A012. Critical Thinking Final Report

***Authors**: Neil R. Thomason[a], Tom Adajian[b], Ashley E Barnett[c], Sandy Boucher[d], Eva van der Brugge[e], John Campbell[f], William Knorpp[g], Larry Lengbeyer[h], David R. Mandel[i], Yanna Rider[j], Tim van Gelder[k], and John Wilkins[l]*

[a] The University of Melbourne, neilt@unimelb.edu.au
[b] James Madison University, adajiatrv@gmail.com
[c] The University of Melbourne, aeba@unimelb.edu.au
[d] The University of Melbourne, sandycboucher@hotmail.com
[e] The University of Melbourne, evavd@student.unimelb.edu.au
[f] Latrobe University, jcampbe1@bigpond.net.au
[g] James Madison University, knorppwm@jmu.edu
[h] US Naval Academy, Annapolis, lengbeye@usna.edu
[i] Defence Research and Development Canada, David.Mandel@drdc-rddc.gc.ca
[j] Yanna Rider Consulting, Melbourne, yanna@yannarider.com
[k] The University of Melbourne, Austhink Consulting, tvg@austhinkconsulting.com
[l] The University of Melbourne, jwilkins@unimelb.edu.au

Table of Contents

Executive Summary: 5

1 Background 8

1.1 The Intelligence Community analysts need improved critical thinking skills. 8

1.2 College Only Slightly Improves Critical Thinking Skills 9

1.3 Critical Thinking Skills Significantly Improve When Students Learn to Argument Map 12

2 Description of project and the hypothesis being tested 17

2.1 Berkeley Workshop 18

2.2 Alexandria VA Workshop 20

3 Materials and Methods 21

3.1 Institutions 21

3.2 Instructors 21

3.3 Course structure and development 23

3.4 Textbook development 26

3.5 Lots of Argument Mapping Practice (LAMP) 26

3.6 Mastery Learning Milestones (MLMs) 27

3.7 Testing procedure 28

3.7.1 Law School Aptitude Test (LSAT) Logical Reasoning subsection 28

3.7.2 California Critical Thinking Skills Test, paper version (CCTST) 29

3.7.3 Halpern Critical Thinking Assessment (HCTA; short form) 30

4 Experimental Results 31

4.1 Anonymized reporting of data for subjects and institutions 31

4.2 Table of Experimental Results 32

4.2.1 Analysis of data by experiment 33

4.2.2 Untrimmed and trimmed data 34

4.2.3 Difference between pre-course subject ability, across individuals and across institutional settings 34

4.3 Gender Differences 35

4.3.1 Comparison of experiment results by gender 35

5 Discussion 36
5.1 Overall discussion 36
5.2 Extraordinary Scrutiny, Critical Thinking Classes and Standard Critical Thinking Tests 40
5.3 Using the LSAT to measure changes in critical thinking ability 50
5.3.1 CT Tests -- Reading Ease, Grade Levels, and Test Statistics 53
5.4 CCTST and HCTA 56
5.5 The Textbook 58
5.5.1 Strengths 58
5.5.2 Improvements on previous Argument Mapping courses/textbooks 61
5.5.3 Weaknesses of textbook 62
5.6 Lots of Argument Mapping Practice (LAMP) 66
5.7 Mastery Learning Milestones (MLM) 67
5.7.1 Possibility of 100% automated MLM assessment and MOOCs 69
6 Acknowledgements: 75
7 Appendix: van Gelder: Meta-analysis "Impact of argument mapping on critical thinking skills" 77
7.1 Abstract 77
7.2 Introduction 77
7.3 Argument Mapping (AM) 78
7.4 Previous Results 79
7.5 Method 84
7.5.1 Search for Studies 84
7.5.2 Data extraction 84
7.5.3 Classifying Studies into Intensity Groups 85
7.5.4 Computing Results 87
7.6 Results 87
7.7 Discussion 88
8 Appendix: Anonymized Reports 92
8.1 Instructor Report A 92
8.2 Instructor Report B 107
8.2.1 Overall assessment. 107

8.2.2 Strengths 107
8.2.3 Weaknesses 108
8.2.4 Specific Aspects of the Course 109

9 Appendix: Other ways to use argument maps to teach critical thinking 112

10 Appendix: Statistical Data, with additional analyses 115
10.1 Data 115
10.2 Analysis of Untrimmed Data 125
10.2.1 Experiment Data 125
10.2.2 Level of Scrutiny 126
10.2.3 Gender 126
10.3 Trimmed Data 127
10.3.1 The Rationale for examining trimmed data and the trimming technique 128
10.3.2 Trimming Methodology 129
10.3.3 Trimmed Data and Analyses 132
10.3.4 Level of Scrutiny (trimmed data) 135
10.3.5 Comparison of experiment results by gender, trimmed data 136

11 References 137

Executive Summary:

There has been considerable evidence that argument-mapping-based critical thinking classes are much more effective (an average student improvement of roughly 0.75 standard deviation (SD) per semester) than regular critical thinking classes (roughly 0.35 SD per semester) which, in turn, are much more effective than undergraduate studies (roughly 0.15 SD per semester). There is also considerable, robust evidence that peer instruction and mastery learning markedly improve educational outcomes. This project was to see whether it is possible to get a 1+ SD improvement measured on standard critical thinking tests, by teaching argument mapping using peer instruction and mastery learning.

Seven experiments, taught by a total of seven philosophically-sophisticated instructors, were run at five institutions: Canadian Border Services Agency, RAF Molesworth, US Naval Academy, and two universities. Pre and post testing was done on the California Critical Thinking Skills Test (CCTST), the Halpern Critical Thinking Assessment (HCTA) and the Logical Reasoning subsection of the Law School Aptitude Test (LSAT). Results on individual experiments ranged from 1.25 SD to –0.05 SD; the project did not achieve a reliable 1+ SD improvement.

In accord with previous critical thinking research, subjects in Experiments 1 through 6 were taught to scrutinize co-premises for *prima facie* plausibility. In Experiment 7, they learned how to scrutinize such co-premises in far more detail and to unpack many of them into background causal and conceptual presuppositions, which in turn were to be scrutinized. Although time-consuming to teach and to apply, such extraordinary scrutiny is often crucial in important espionage, legal, and scientific investigations. Still, subjects in this experiment did much worse on the time-limited standardized tests than subjects in the other six experiments.

	Normal Scrutiny (Experiments 1 to 6)		Extraordinary Scrutiny (Experiment 7)		All Expts	
	Stand'ised ES	95% CI	Stand'ised ES	95% CI	Stand'ised ES	95% CI
CCTST (5 exp'ts)	0.847	[0.57, 1.12]	-	-		[0.57, 1.12]
HCTA (2 exp'ts)	0.721	[0.46, 0.98]	0.008	[-0.45, 0.47]	0.539	[0.32, 0.76]
LSAT (All exp'ts)	0.370	[0.24, 0.50]	-0.054	[-0.38, 0.27]	0.307	[0.18, 0.43]
All	0.505	[0.40, 0.61]	-0.033	[-0.29, 0.22]	0.424	[0.32, 0.52]

As can be seen from the Table, improvements on the first two tests that are explicitly designed to measure critical thinking are considerably greater than improvements on the LSAT, with its high literacy loading. The LSAT was designed to predict first year law school grades and may be closer to a general intelligence test than a critical reasoning test.

Experiment 7's different pedagogical emphasis and quite different outcomes suggest its results should be analyzed separately. Given that the critical thinking tests used in this study do not have items that measure the extraordinary scrutiny skills the course intended to teach, the non-existent average improvement may be because of the techniques taught in this course. Extraordinary scrutiny of test items takes considerable working memory and slows down performance scores in timed tests. Yet, in some circumstances, extraordinary scrutiny is crucial to thinking clearly on important issues. Thus, there is a danger that, when present critical thinking tests are used, excellent courses that emphasize it may wrongly be seen as ineffective. We suggest that the intelligence community look further into those factors that differentiate normal scrutiny critical thinking from extraordinary scrutiny. If the latter seems important to the effective accomplishment of analytic tasks, that it develop an objective extraordinary-scrutiny-oriented critical thinking test and better ways to teach these valuable skills.

The Normal Scrutiny results reinforce the existing evidence on the effectiveness of argument mapping courses for improving the critical thinking skills that standard tests do measure. In two of the three tests, the standardized effect size is above 0.7, as one would expect from the previous research. For the LSATs, the effect size is about half that.

But these results do not show the expected gains over previous research from adding peer instruction and mastery learning. It may be that we are already near ceiling of possible gains arising from any way of teaching critical thinking and so it may be too much to hope for reliable 1+ SD gains. We are disappointed that we did not achieve our expected goal of improving measured performance on critical thinking tests by 1 SD or more. We cannot specify the reasons for falling short, but we do not think that our failure to meet this goal suggests that it is beyond reach. Rather we think that, using the materials developed in this project and applying the lessons learned, 1+ SD might be achieved by:

- Covering more topics important to critical thinking, such as inference-to-the-best explanation and application of basic informal logic forms.
- Moving the textbook fully on-line, with many short videos.
- Refining the mastery-learning questions and dividing them into smaller, more focused lessons, to make it easier to learn the material from the mastery-learning quizzes while making it harder to pass them with brute-force repetition that bypasses learning.
- Integrating on-line mastery-learning quizzes and classroom exercises with the on-line textbook.
- Developing better techniques for peer assessment of each other's maps, by adapting already available online techniques for peer assessment of essays.
- Developing a flipped-classroom MOOC (Massive Open Online Course) could provide considerable reliable evidence of student (mis)understandings for guiding the improvement of the textbook, videos, mastery-learning quizzes, classroom exercises, and peer-assessment of maps.

The improvements we have seen on the two standard CT tests we used suggests that the IC may want to integrate argument mapping into its critical thinking courses and to further improve learning in connection with the introduction of argument mapping. Further, it may want to sponsor further research on ways to enhance the effective teaching of critical thinking by running experiments to assess the effectiveness of the above suggestions.

1 Background

1.1 The Intelligence Community analysts need improved critical thinking skills.

First-rate analysts must be first-rate critical thinkers; for at the core of analysis is critical thinking, however important other analytic skills and knowledge are.[1]

The Intelligence Community (IC) has repeatedly acknowledged the importance of critical thinking. *Intelligence Community Directive Number 203 – Analytic Standards* D.4.e (6) states that a good analytic presentation "Uses logical argumentation – Analytic presentations should facilitate clear understanding of the information and reasoning underlying analytic judgments." Further, *Directive Number 610 – Competency Directories for the Intelligence Community Workforce* lists critical thinking as a "Core Competency" for "Intelligence and Analysis Production Intelligence Community", as well as for many other IC jobs.

However, there is reason to believe that the critical-thinking skills of many analysts need significant improvement. For example, the WMD Commission Report (Whitney 2005) found many serious flaws in analysts' reasoning:

- "Analysts did not question the hypotheses underlying their conclusions, and tended to discount evidence that cut against those hypotheses." (p. 169)
- "Perhaps most troubling, we found an Intelligence Community in which analysts have a difficult time stating their assumptions up front, explicitly explaining their logic, and, in the end, identifying unambiguously for policymakers what they do not know." (p. 389)

---

[1] We used the American Philosophical Association's definition from its "Expert Consensus Statement Regarding Critical Thinking" (Facione 1990):

"We understand critical thinking to be purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological or contextual considerations upon which that judgment is based."

The rest of the statement deals with such issues as the role of critical thinking in personal and public life and the characteristics of an ideal critical thinker.

- "Analytic reasoning must be more rigorous and be explained in clearer terms in order to improve both the quality and credibility of intelligence." (p. 409)

The Commission's findings about the quality of analysts' reasoning can be summarized: analysts' critical thinking must be substantially improved.

1.2 College Only Slightly Improves Critical Thinking Skills

Why do analysts' critical thinking skills need to be substantially improved, given that almost all analysts are college graduates and some have advanced degrees? The answer is straightforward: the evidence shows that critical-thinking skills improve remarkably little during undergraduate years.

For example, The Wabash National Study of Liberal Arts Education (2009) looked at many colleges and found that "although students' improvement on the CAAP (the ACT's College Assessment of Academic Proficiency] Critical Thinking test was statistically significant, the change was so small (less than 1% increase) that it was practically meaningless" – only 0.11 standard deviation in two semesters. Hitchcock's review (2004) concluded that, over the four years of American university education, the average student only gains somewhere between 0.5 and 0.65 standard deviations (SD), or about 0.08 SD per semester, some of which might simply be due to maturation. Pascarella and Terenzini's comprehensive review of tertiary education tells us this study is indicative of what generally been found:

> "A sample of 147 students took the CCTDI as beginning freshman and then again as seniors. From the first to the second testing, the sample increased a statistically significant 7.43 points in critical thinking disposition. *This translated into a gain from freshman to senior of .28 of a standard deviation* (11 percentile points)." (Pascarella and Terenzini 2005, 160, our italics)

Arum and Roksa's recent study over many colleges and universities tested the change in undergraduate synthesizing skills and found:

> "The end result is that many students are only minimally improving their skills in critical thinking, complex reasoning, and writing during their years through higher education. From their freshman entrance to the end of their sophomore year, students in our sample on average have improved

> these skills, as measured by the CLA (College Learning Assessment], by only 0.18 standard deviations. … we observe no statistically significant gains in critical thinking, complex reasoning and writing skills for at least 45 per cent of the students in our study." (Arum and Roksa 2011, 35–36. Our italics)

Whatever other benefits of undergraduate education there are, considerably improved critical thinking is not one of them. The IC must look elsewhere to satisfy its need for analysts with high-level critical-thinking.

The literature on standard dedicated critical-thinking courses promises no breakthroughs. As a leading theorist of critical thinking wrote in 2000 about introductory critical-thinking courses: "I wish I could say that I had a method or technique that has proved successful. But I do not, and from what I can see, especially by looking at the abundance of textbooks on critical thinking, I don't think anyone else has solved this problem either" (Walton 2000).[2] The early literature reviews found little evidence that dedicated critical-thinking instruction produced gains beyond those due to maturation and university education (McMillan 1987). Alvarez's more thorough recent meta-analysis of critical thinking courses found that critical thinking can be taught, but that standardly-taught critical-thinking courses only produce an average increase of 0.34 SD over controls (Alvarez 2007). While this is more than standard university education by itself, the IC needs much more.

Why is critical thinking so hard to teach? In large part because critical thinking itself is a complicated blend of different subskills. An intelligence analyst, for example, must assess the reliability of bits of evidence, assess how each bit fits into relevant sub-hypotheses, and then fit the sub-hypotheses into a single well-reasoned whole. To do this, the analyst must determine which claims, inferences and objections are well supported and which are problematic and in the end should arrive at a reasoned and defensible estimate of the probability that the ultimate hypothesis of interest is true. Further complicating the picture, the analyst will often know that there is a good chance

[2] Such negative evaluations have been around for a while. Consider this from Lehman, Lempert, and Nisbett (1988, 441):

"The truth is that we know very little about reasoning and how to teach it. … just how much we can improve reasoning by instruction, is now a completely open question."

that some of the "evidence" might be come from less-than-fully-credible sources or even be deliberately misleading; this probability and its implications must condition the analysis.

Good analysts (as the Weapons of Mass Destruction committee emphasized) will question the hypotheses underlying their conclusions and will not unduly discount evidence that cuts against those hypotheses. They must be able to state their assumptions up front, be able to explicitly explain their logic and, in the end, identify unambiguously for policymakers what they know and do not know. To enhance quality and credibility, their reasoning must be rigorously explained and suitably transparent.

To achieve this, analysts must reflexively understand the underlying structure of their own reasoning as well as the reasoning of others whose arguments figure in their analyses. Usually, they store those structures in their head – a cognitively demanding task, since working memory is limited. Fortunately, there is a well-tested alternative to keeping an argument's structure in the analysts' head.

The underlying structures, complex though they may be, can be graphically represented in argument maps. Figure 1 shows an argument map of the structure underlying a relatively simple analytic product. For most analytic products, the corresponding map would be considerably more complex. The complexity is an unavoidable characteristic of real world problems that cannot be escaped, other than by greatly over-simplifying the issues.

An argument map, as we see in Figure 1, can substantially reduce the associated cognitive load without over-simplifying the issues. A well-thought-out argument map makes it easier to see which evidence is reliable and why, which evidence is dubious and why, which evidence supports which hypotheses, which evidence cuts against which hypotheses, how well supported the various hypotheses and sub-hypotheses are, and ultimately how the entire argument fits together. Argument mapping has the potential to considerably help analysts as they work systematically through the entirety of the available evidence and their reasoning.

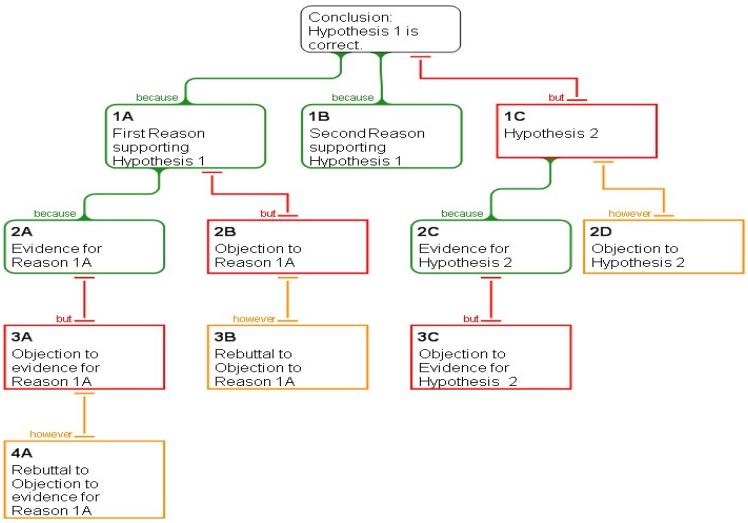


Figure 1. Argument map showing the structure of a relatively simple analytic product.

However, producing a good argument map of a complex analytic problem can be very time consuming, and analysts often work under great time-pressure. Hence the idea behind this research proposal. Our funded seedling proposal was designed to take a step toward answering the question: Can argument-mapping-based critical-thinking courses significantly improve critical thinking done when an argument map is not used? That is, do the skills learned in an argument mapping class transfer to analytic work done without argument maps?

1.3 Critical Thinking Skills Significantly Improve When Students Learn to Argument Map

There is evidence that the skills developed in argument-mapping-based courses substantially transfer to critical thinking done without argument maps. Alvarez's meta-analysis found that such critical thinking courses produced gains of around 0.70 SD, about twice as much as standard critical-thinking courses (Alvarez 2007, 69-70 et seq.).

For this project, van Gelder (Appendix) expanded the Alvarez meta-analysis to include recent research and distinguishing among the intensity of the students' argument mapping study. The High Intensity group did considerably better 0.84 SD [0.7, 1.0] than the Medium Intensity group 0.44 which in turn did better than the Low Intensity group 0.30. The effect size for all studies was 0.57, compared to that of standard (non-

argument mapping) critical thinking subjects 0.34 SD and 0.12 per semester for college without a dedicated critical thinking course.

Almost all the reviewed studies used standard critical-thinking tests. In taking these tests, the students did not make argument maps, did not have access to argument-mapping software, and were under considerable time-pressure. Thus, these results directly measure the transfer to critical-thinking tasks done without argument maps, albeit in non-IC contexts.

The research does not, however, tell us why these gains occurred and the issue has not been experimentally studied. It is possible that the reported effects are not due to learning how to map. For example, the teachers who taught the Argument Mapping based courses are simply more dedicated and effective than teachers who taught traditional critical thinking courses. Or, that argument mapping is more intimidating to many students, so the less dedicated ones quit the classes. Or, that the argument mapping courses require students to spend more time-on-task and that explains the superiority of the argument mapping results. For whatever it is worth, it does not seem to us that teachers who run experiments on their teaching of critical thinking are not dedicated, particularly as their reputation might somewhat depend on how successful their classes are. Further, at least in this project's classes there was not much student attrition. The third explanation, that subjects in argument mapping classes had to spend more time-on-task seems plausible to us and future experiments should control for this possible confounding variable.

Assuming that the gains are real and due to learning how to argument map, several plausible non-exclusive hypotheses may explain these gains. However, a few of them require that, in a vague but psychologically real sense, people have something like maps in their minds. This possibility needs further investigation.

> **Logical Structure:** Argument maps display an argument's logical structure more clearly than does the standard linear way of presenting arguments.
>
> **Critical Thinking Concepts:** In learning to argument map, students master such key critical thinking concepts as "equivocation", "rebuttal", "unstated assumption", "co-premise", "preponderance of evidence", "logical structure", "circular argument", "independent evidence", etc. Mastering such concepts is

not just a matter of memorizing their definitions or even being able to apply them correctly; it also requires understanding why the distinctions these words mark are important and using that understanding to guide one's reasoning.

Visualization: Humans are highly visual and learning argument mapping may provide them with a basic set of visual schemas with which to understand argument structures.

More Careful Reading and Listening: Learning to argument map teaches people to read and listen more carefully, and highlights for them the key questions "What is the logical structure of this argument?" and "How does this sentence fit into the larger structure?" In-depth cognitive processing is thus more likely.

More Careful Writing and Speaking: Argument mapping helps people to state their reasoning and evidence more precisely, because the reasoning and evidence must fit into the map's logical structure. This additional clarity become habituated, transferring to situations where there is not an explicit map.

Literal and Intended Meaning: Often, many statements in an argument do not precisely assert what the author meant. Learning to argument map enhances the complex skill of distinguishing literal from intended meaning.

Externalization: Writing something down and reviewing what one has written often helps reveal gaps and clarify one's thinking. Because the logical structure of argument maps is clearer than that of linear prose, the benefits of mapping will exceed those or ordinary writing. Externalizing also unburdens working memory and allows more pieces of an argument to be examined at once than would be possible using memory-based argument representation. This hypothesis would only apply to actual argument maps, and would not explain how learning how to argument map would improve ones reasoning when one does not actually make a map.

Anticipating Replies: Important to critical thinking is anticipating objections and considering the plausibility of different rebuttals. Mapping develops this anticipation skill, and so improves analysis.

It would be surprising if any single hypothesis were the complete story for all people.

The increases of 0.7 SD for argument mapping subjects that Alvarez found and the 0.84 SD van Gelder found for intensive argument mapping subjects are remarkably large for the tertiary research literature. These results are comparable to this project's CCTST

and HCTA results. As far as we know, no previous tests had used the LSAT. Still, further gains might be possible.

Although Alvarez's was a careful, comprehensive reviews of the available data, and a roughly 0.7 SD increase is substantial, there was good reason to believe that taking a mastery-learning approach to teaching an argument-mapping-based course would be more effective. In mastery-learning courses, students master the material in one section before progressing to the next, where they build on their newly acquired understanding. Thus, students are not working at "catching up", having to learn new material while continuing to learn the partially-mastered previous material. This project used two types of mastery learning: Keller's Personalized System of Instruction (PSI) which is aimed at individual students' mastery of the material, and Mazur's peer instruction (PI) which is aimed at class-wide mastery.

A Personalized System of Instruction (PSI) course is divided into units; a student must pass each unit at mastery level before moving on to the next. There is no penalty for failing a quiz, and to achieve mastery the student may have to take many different quizzes on each unit's material. Thus, PSI requires many quiz questions for each unit. Learning is promoted because there is immediate feedback on each quiz as well as explanations for why right answers are right and the wrong answers wrong. In addition, students set their own pace and, guided by the detailed feedback PSI provides, can focus on the material they are weakest at.

The evidence for PSI's effectiveness is long-standing. Here is the evaluation in the chapter on "Research on Teaching in Higher Education" from the 3rd edition of American Educational Research Association*'s Handbook of Research on Teaching:*

> "The single most significant conclusion to be reached from research on innovatory teaching methods in higher education is that the Keller Plan (PSI] is clearly superior to other methods with which it has been compared. *Indeed, the Keller Plan has been so consistently found superior that it must rank as the method with the greatest research support in the history of research on teaching*." – (Dunkin and Barnes 1986, 759. Our italics)

Subsequent meta-analyses have shown similar results, covering a range of subjects (physics, mathematics, English, etc.) and measuring techniques (final grades, pre-post

testing, etc.). Kulik, Kulik, and Bangert-Drowns (1990a) found that PSI produces an average one-semester gain of about 0.5 SD more than standardly-taught controls. Spencer's meta-analysis, which used slightly different criteria for study inclusion, found a gain of 0.7 SD against standardly-taught controls (Spencer 1991). Possibly, PSI courses can be made even more effective by collecting data to determine precisely which topics students are finding particularly difficult and developing on-line mini-lectures focused on those topics.

In recent years, much of the research on Massive Open Online Courses (MOOCs) and other forms of on-line instruction has centered on mastery learning. For example, in the Khan Academy, students are not supposed to progress to the next lesson until they have answered ten consecutive questions correctly. In classroom-based Kahn Academy teaching, the teacher can see whether each student is following this advice.

The second mastery-learning technique, peer instruction (PI), devotes part of each lecture to short small-group discussions of conceptual questions, with students explaining their own answers to each other.[3] To ensure that students are prepared for their discussions, well before each lecture they are asked two content questions about the assigned reading and are also asked how well they think they understand the material.[4] Answers are emailed to the lecturer at least an hour before the class. Among other benefits, this helps the teacher focus the lectures on material the students find difficult.

During the discussions with their peers, many students change their minds. Remarkably, they are roughly five-to-seven times more likely to shift from an incorrect to a correct answer than from a correct to an incorrect one (Crouch and Mazur 2001, Giuliodori, Lujan, and DiCarlo 2009, Smith et al. 2009). There is evidence that this increase in understanding transfers to similar questions. In one study, when questions were "difficult", discussions produced a remarkable increase from about 25% correct

[3] Mazur (1997) remains the best introduction to PI. For more recent refinements, see Crouch et al. (2007).

[4] Interestingly, Mazur found a strong *inverse* correlation between students correctly answering the two content questions and their confidence about their level of understanding. (Mazur, personal communication)

answers on the initial questions to about 60% correct answers on isomorphic questions – without any feedback from the teacher or the rest of the class (Smith et al. 2009)! Meta-analysis has shown that PI courses produce an average one-semester gain of about 0.39 (Hake's average standardized gain <*g*>) more than standardly-taught controls (Crouch et al. 2007).[5]

2 Description of project and the hypothesis being tested

This project tested the hypothesis that a critical-thinking course can robustly improve critical thinking by at least one Standard Deviation on standard critical thinking tests. To do this, the project developed and tested an argument-mapping-based critical-thinking course (Critical Thinking by Argument Mapping, CTAM) taught using mastery learning and peer instruction techniques. The teachers were experienced, philosophically-sophisticated teachers of critical thinking, although most had not taught critical thinking with such an intensive focus on argument mapping and some had not used argument mapping-based instruction at all. Moreover, the mastery learning approach and the use of peer-instruction were new to most instructors.

During the pre and post testing, students did not have access to argument mapping software. Thus, the tests measured the transfer from argument mapping-based learning to critical thinking questions without argument mapping. The intervention had approximately 140 subjects, ranging from USNA midshipmen through Canadian Border Services Agency (CBSA) and NATO analysts to bright undergraduates in a couple of institutions.

General course development for co-ordinated experimental courses and the conceptualization and production of courses materials, was undertaken from April 2012 to January 2013. In this period, two workshops were conducted, which guided the course development: One in Berkeley CA in June, and one in Alexandria, in December.

[5] All other reported or projected effect sizes are measured in Standard Deviations (SD), using Cohen's *d*. For technical reasons, it is not possible to directly compare Cohen's *d* to Hake's <*g*>, but since this project used both PSI and PI, a comparison of their relative effectiveness when used in isolation is not needed.

During this period, Lengbeyer ran a 10-week trial at USNA, providing valuable feedback. The experiments were launched in January 2013 in five locations with a total of seven teachers. Their feedback guided ensuing revisions.

2.1 Berkeley Workshop

The Berkeley discussions centered on the role of CTAM and a literature review of prior work (the full report was submitted as A002: Critical Thinking Workshop Presentations and Papers"). The Workshop intensively discussed philosophical (What is an unstated premise?), practical (How can this course best improve analysts' performance?) and pedagogic (How, if at all, should the fallacies be taught?) questions. The Review covered the scholarly literature (published and gray) regarding argument mapping, critical thinking, mastery learning, learning through discussion, and transfer from classroom to work.

The issue of whether teaching the fallacies was useful in teaching critical thinking was the first to be discussed. The literature on this was considered, and the general consensus was that there are several problems with teaching fallacies in critical thinking. One is that the skill taught is the skill of identifying and using technical names such as *ignorantio elenchi* rather than critically thinking about the arguments as such. Another is that nearly all proposed "fallacies" are in fact not fallacious in some contexts (for example, scientific studies often rightly affirm the consequent, and legal arguments often make reasonable ad hominem arguments). Hence the difficulty in identifying the fallacy *as* a fallacy adds cognitive overhead to learning how to reason. Finally, most bad arguments involve more than one fallacy, or can be identified as different fallacies with equal support, making assessing the students' grasp of the subject matter difficult or even merely the subjective opinion of the assessor.

The question of how to read a text passage, particularly in the context of the IC, was raised. Developing a standardized test for an IC argument-map-based critical thinking training programs was suggested, but was generally agreed to be a subsequent task, dependent on the successful completion of this project.

The role of Bayesian reasoning, a widely discussed and accepted account of reasoning, was also considered. Some consideration for using CTAM with Bayesian probabilities

was given, but it was thought to be outside the scope of the present project. It was agreed that a Bayesian approach would illuminate many issues that arise with argument maps, but that this was a complex task, and the challenges that would arise in developing materials and testing the core hypotheses was all that could reasonably be expected to be accomplished in the time and with the available funds.

The educational aspects of teaching critical thinking by argument mapping were discussed, in particular the questions of peer instruction and personalized system of instruction techniques. It was argued that some improvement on CT ability is just a fact of maturation in early adulthood. The nature of the textbook used in the project was discussed, focusing upon the overall purposes of the course. Issues like the role of technical terminology (such as *modus ponens*), epistemic relevance, and validity in teaching were considered. It was agreed that this was a course in informal rather than formal logic, and as such using letter-based arguments should be avoided. The role of scoring rubrics to make evaluating text answers was also presented and discussed. This involved considering logical correctness, implicit premises (unstated premises) and completeness of maps. The experience of briefly teaching CTAM in the US military, at Annapolis and in the US Army in Africa was presented.

There was a panel discussion on "Arguments as components as debates: "Can Argument-Mapping of debates enhance decision-making? Inculcate critical thinking?" The difficulties and advantages of using argument mapping to map debates rather than single arguments were discussed from a philosophical and an IC perspective, noting that these debates were typically dense and complex.

Biases in reason and the difficulty of identifying and individuating reasons and parts of arguments were discussed, including the Lehman, Lempert, and Nisbett (1988) hypothesis that statistical training improved reasoning skill in ordinary life and in science. The use of heuristics and the existence of biases pose a well-known threat to good reasoning.

Under the rubric of practical and theoretical issues, the question of whether bias can be attenuated by instruction in CT was raised, in the context of a recent *Behavioral and Brain Sciences* essay by Mercier and Sperber (2010). This argues that reasoning did

not evolve in order to attain truth or the best-attested theory, but to convince others to adopt a view through argumentation. That is, the point of reasoning matches much closer to what historically has been called "rhetoric" than "informal logic."

We also discussed the practical issues of teaching CT as part of a standard content course (history, sociology, etc.) where the subject matter would often swamp the CT elements; many students tend to show little interest in anything that is not seen as furthering vocational goals. Further, CT was seen by those within various disciplines as having a low status and value.

Finally, there were presentations illustrating the use of the *Rationale* software, and practice sessions to allow those who had previously been unacquainted with the software to learn it. The presentations and hand-on experiences illustrated the legitimate disagreements that often arise when mapping all but the simplest arguments.

2.2 Alexandria VA Workshop

The Alexandria discussions focused on Teaching, particularly on integrating the developed materials with in-class activities (the full report was submitted as "A005: Teaching Workshop Presentations and Paper"). It considered the assessment and structure of the course used in the experiments.

It was agreed that the course should be practice-oriented, with almost no lecturing. Problems with achieving a 1+ SD gain were discussed with respect to measurement in evaluation tests.

The modules of the course plan were considered. Module 1 was designed to let students become familiar with the software and core argument mapping techniques. The utility of Critical Thinking using Argument Maps (CTAM) to intelligence analysts was discussed as well. The complexities of understanding co-premises, particularly unstated premises, were discussed. The fact unstated assumptions (co-premises) are highly contextual, leading to some concern particularly with regard to marking maps. Each of the Mastery Learning Questions (renamed in the project Mastery Learning Milestones, or MLMs) was discussed in turn. The method of teaching by Lots of Argument Mapping Practice (LAMP) was covered and the process of integrating

mastery learning (the "Keller Plan") explored. The schedule for the early weeks of the course was organized and teaching tasks were worked out. The draft textbook was reviewed and some suggested changes were discussed. Many of them were adopted.

Tom Adajian presented the history and philosophy of logic diagrams and notation, and the ways in which they were used to teach. Bill Knorpp presented on the efficacy of checklists in reducing cognitive load on pilots, and how AM can be seen as implementing a checklist.

Ashley Barnett took instructors through the course's practice activities, to ensure they had a clear understanding of the way to teach and use AM. A draft test, and questions to measure the subjects' background and attitudes, were presented and debated. Rick Lempert presented a fictional case study for IC analysts, with a quiz to test analysts' critical thinking ability.

3 Materials and Methods

3.1 Institutions

The lead institution in the project was the University of Melbourne, Melbourne, Victoria, Australia (History and Philosophy of Science, in the Faculty of Arts). The collaborating institutions were:

- James Madison University, Harrisburg, Virginia, USA
- US Naval Academy, Annapolis, Virginia, USA
- Defence Research and Development Canada, Toronto Research Centre (Sensemaking and Decision Group, Socio-Cognitive Systems Section)
- Strategic Risk Assessment Division, Canadian Border Services Agency
- Royal Air Force, Molesworth Air Force Base, East Anglia, England

3.2 Instructors

All the instructors were philosophically sophisticated educators with experience in teaching critical thinking. They were:

Ashley Barnett, Research Assistant, University of Melbourne. He is a PhD student working on the evaluation of arguments and has taught critical thinking at Swinburne, Monash and Melbourne universities. He has worked closely with Tim van Gelder, who pioneered argument mapping in education and workplace. For

this project, he was tasked with integrating mastery learning into an argument mapping course. He was the lead writer and designer of the learning materials, managing the production of the mastery learning quizzes and the argument mapping practice questions, and wrote over eighty per cent of the questions used in the final course.

Robert Blair RAF Molesworth, Cambridgeshire, UK. Blair has 15 years experience in the U.S. Intelligence Community as an analyst, instructor and manager, and is a certified U.S. Defense Intelligence Agency (DIA) Joint Military Intelligence College Master Instructor. In his current role, Bob has led intelligence engagements with allies in partners in over 30 countries across Europe, Asia, and Africa, and he writes and speaks regularly in IC-related forums. Bob expects to earn his Masters of Science in Strategic Intelligence for the U.S. National Intelligence University in Spring 2014. He has a Certificate in Terrorism Studies from the University of St. Andrews, and graduated from Davidson College.

Dr Sandy Boucher, Research Assistant, University of Melbourne, did his Masters at Monash University and his PhD at the University of Melbourne, under Greg Restall. He has taught at the Universities of Melbourne and Connecticut in a variety of areas, including philosophy of science, philosophy of biology, epistemology, logic, ethics, metaphysics, and philosophy of mind. His research interests are mainly in the philosophy of science and the philosophy of biology.

Eva van der Brugge is a doctoral candidate in History and Philosophy of Science at the University of Melbourne, from which she was awarded a full Melbourne International Student Scholarship. Her provisional thesis title is "External Validity of Critical Thinking Assessment". She has an MA in "Rhetoric, Argumentation Theory and Philosophy" from the University of Amsterdam and an MSc in "Methodology and Statistics of Psychology" from Leiden University. She has a BSc in Psychology and a BA in Comparative literature, both from Leiden University. Until recently she was a Research Fellow (psychometrics) at the Australian Centre for the Educational Research. Eva is currently a visiting student researcher at Princeton University.

Dr William Knorpp, faculty, James Madison University. Dr Knorpp has taught logic, philosophy and critical-thinking courses at North Carolina, William and Mary, and James Madison. He is the author of "The Relevance of Logic to Reasoning and Belief Revision" (*Pacific Philosophical Quarterly* 1997), and co-author of papers evaluating critical-thinking exams, including "Examining the Exam: A Critical Look at the Watson-Glaser Critical Thinking Appraisal Exam (*Inquiry* 2001). His current work on critical thinking includes Beardsley-type argument diagrams, and pedagogical issues in the theory of fallacies. He is co-authoring an on-line critical-thinking book *Digital Critical Thinking and Logic*.

Dr Lawrence Lengbeyer (AB Applied Mathematics, Harvard; JD Law & Hermeneutics, Yale; PhD Philosophy, Stanford) is Associate Professor of Philosophy in the Department of Leadership, Ethics, and Law at the United States Naval Academy. His research focus is in moral psychology, more specifically belief and related cognitive processes, emotion, interpretation, and ethics. He has taught courses on intellectual virtue, scientific reasoning, legal reasoning, and critical thinking, and was an instigator of the Critical Thinking Working Group at the Naval Academy.

Dr Neil Thomason, Principal Investigator, University of Melbourne. Dr Thomason has long focused on understanding and teaching informal reasoning. Many of his publications examine the structure of actual scientific and philosophical arguments. He taught critical thinking at Berkeley (as a TA), Reed, Vassar, and in Australia and, before retiring in 2008, he had integrated argument maps into many of his classes. He is an expert in social science experimental design. He has long been interested in improving the critical thinking in political science and intelligence; Cf. Lieberman and Thomason (1986). Rieber and Thomason's (2005) argues for the necessity of careful scientific testing of proposed analytic techniques, a concern behind his National Research Council report on the Analysis of Competing Hypotheses (Thomason 2009).

3.3 Course structure and development

The course was designed to introduce students directly to the skills needed to analyze text into arguments and produce AMs based on that analysis. Since it was not intended to be a formal logic course, all material about logical operators, fallacies, and even the nature of conditional arguments was left implicit. There was no attempt to train students to recognize these under a technical name or method such as truth tables. To enable the clear and systematic mapping of arguments, we decided to use the *Rationale* argument mapping software, Advanced Reasoning format.

The course was divided into eleven modules, each dealing with a number of sub-topics:

1: Introduction: The basic parts of arguments:

- The parts of arguments and argument maps in a nutshell
- Arguments
- Conclusions, reasons and objections
- Claims
- Complex arguments
- Argument units
- Indicators

- Standard terminology used for argument mapping
- The software used

2: Reasons, Objections and Conclusions

- We're starting with fully-stated arguments
- Reasons
- Reasons and evidence
- Objections
- Reasons and objections – strengthening and weakening
- Reasons and agreements; objections and disagreements
- Combinations of reasons and objections
- Analyzing argument units one at a time

3: Premises

- Co-premises
- Basic premises and intermediate conclusions (= intermediate premises)

4: Claims

- Claims, truth and confidence
- How many claims?

5: Mapping Completely Stated Arguments

- Non-argumentative material
- Repetition
- The parts of arguments and their locations
- Arguments versus explanations

6: Evidential Relevance and Unstated Premises

- Evidential relevance
- Danglers
- Danglers in conclusions: The Rabbit Rule
- Rabbit Rule helps, but it's not enough
- Danglers in premises
- Unstated premises – Where the work often is done
- Trade-offs between relevance and plausibility
- When unstated premises can stay unstated – Don't belabor the obvious
- Cheap co-premises – they may look good, but they don't help

7: More on Incompletely Stated Arguments

- Reasons for and objections to unstated premises
- Reasons for unstated premises

- Objections to unstated premises
- Unstated final conclusions
- Unstated premises and faulty reasoning
- Different acceptable interpretations of arguments: often there is no such thing as *the* argument

8: Refining Stated Claims

- Claims vs. non-claims
- Metaphorical language
- Ambiguity
- Vagueness
- Emotional language and euphemism
- Technical jargon and buzzwords

9: Basic Premises and Basis Boxes

- Expert opinion
- Common belief
- Shared belief
- Personal experience
- Publication
- Treating bases as arguments

10: Evaluation

- Evaluating individual reasons and objections
- Justifying and persuading
- What makes a good reason?
- Evidential relevance and belief
- Evaluating Objections
- 'Bad reason' doesn't mean 'false conclusion'; 'Bad objection' doesn't mean 'true conclusion'
- Lines of Reasoning
- Combinations of reasons and objections
- Counting reasons and objections
- How reasons combine
- Evaluating complex arguments
- Short Cuts
- Fallacies

11: How to convert maps to prose arguments

- Step 1: Literal conversion
- Step 2: Placing the conclusion

- Step 3: Signposting the argument
- Step 4: Removing repetition
- Step 5: Dealing with objections
- Step 6: Adding in basis text
- Step 7: Polishing

The textbook followed this plan, so the course structure *is* the textbook structure and vice versa, and course and textbook coevolved through the project. Earlier versions were tested in practice and in discussion.

The terminology used was in general ordinary language rather than the usual Latin and technical terms. This was a deliberate decision, made to reduce the cognitive load upon students, to avoid teaching technicalities rather than skills, and with the recognition that if IC analysts were trained in critical thinking using argument mapping, the teaching would be to develop sound, well-structured analyses and not to put an analysis into technical philosophic terminology.

3.4 Textbook development

Initially the plan was to develop several online versions of the text, including web-based pages, electronic publications, and PDFs, so that each aspect of the course could be put online in a format that suited the needs of different students. We started in this direction using Adobe Creative Suite software, but encountered difficulties of shared editing and coauthoring using that software. So we reverted to Microsoft Word. As a result, we could only prepare a PDF version in time for the courses. However, with additional time, online versions could be produced with relatively little effort, if wanted to use in, for example, a MOOC. For a critique of strengths and weaknesses of the textbook, see *5.5 The Textbook* below.

3.5 Lots of Argument Mapping Practice (LAMP)

Our goal was to teach argument mapping intensely. To this end, we provided more items posing argument mapping challenges than any other such course; over 300 were written, of which about 120 were used in the final course materials. They were made available as *Rationale* files, to be downloaded from improvingreasoning.com. Students were not expected to do all of them; they varied in difficulty and teachers were to pick

the ones that would challenge the students at the right level. Students were to work with them in class, working closely with their peers sharing a computer and constructing the map together. This way, individuals would get rapid feedback from their colleagues and from the teacher moving around the room. In classes where the necessary technology was available, an answer would be projected onto a screen in the front of the room and there would be general discussion, with the map modified as suggestions came in.

3.6 Mastery Learning Milestones (MLMs)

To implement mastery learning (Kulik, Kulik, and Bangert-Drowns 1990b), we developed a series of tests for each module. Students were to complete and pass each module before they were able to progress to the next. Initially we referred to these as Mastery Learning Questions, but decided that was less descriptive and might deter students from doing them, and so we renamed them as Mastery Learning Milestones to emphasize the achievement aspect.

Questions were developed with the IC in mind, and where possible using realistic cases. We produced these from a range of sources:

- Examples widely used in the critical thinking literature.
- Examples from mass media.
- Examples developed in discussion by project staff, notably Ashley Barnett and John Campbell.

The MLMs were developed with a standard map solution, so that the instructors could correct them easily, although alternative maps were possible in many cases, posing a difficulty that needs to be addressed. They were accessible via the QuestionWriter site (see CDRL A008) and results were sent to the instructors automatically. Students were able to retake each milestone level on randomly selected questions as many times as they needed, without penalty, but could not progress until they had passed with at least 80% correct answers for each milestone. For a report on how MLMs worked, see *5.7* Mastery Learning Milestones (MLM) below.

3.7 Testing procedure

We used three well-established critical thinking tests: the Logical Reasoning subsection of the Law School Aptitude Test (LSAT), California Critical Thinking Skills Test (CCTST) and the Halpern Critical Thinking Assessment (HCTA). The last two are standard critical thinking tests; the first has a clear prima facie relationship to critical thinking and is designed to measure a number of the skills that critical thinking courses attempt to teach, but it has been validated on its ability to predict first-year law school grades. Almost every subject in each experiment took two pre-tests from them as well as the appropriate post-tests. To measure differences across groups, the LSAT was given in every experiment. To minimize test-familiarity effects, subjects were tested in AB/BA design – that is, given one form of a test (e.g., A) for the pre-test and the alternative form (B) for the post-test.

We had hoped to use those who applied but could not fit into the courses as controls, but we did not have enough excess course applicants to allow this. Thus our results do not control for maturation effects or the results of simply taking another course. But, given the research cited above and in van Gelder's appendix, we have no reason to believe that such effects would be large.

3.7.1 Law School Aptitude Test (LSAT) Logical Reasoning subsection

According to the Law School Admission Council
(http://www.lsac.org/jd/lsat/prep/logical-reasoning):

> Logical Reasoning questions evaluate the ability to analyze, critically evaluate, and complete arguments as they occur in ordinary language. The questions are based on short arguments drawn from a wide variety of sources, including newspapers, general interest magazines, scholarly publications, advertisements, and informal discourse. These arguments mirror legal reasoning in the types of arguments presented and in their complexity, though few of the arguments actually have law as a subject matter.
>
> Each Logical Reasoning question requires the examinee to read and comprehend a short passage, then answer one question (or, rarely, two questions) about it. The questions are designed to assess a wide range of skills involved in thinking critically, with an emphasis on skills that are central to legal reasoning.

These skills include:

- Recognizing the parts of an argument and their relationships
- Recognizing similarities and differences between patterns of reasoning
- Drawing well-supported conclusions
- Reasoning by analogy
- Recognizing misunderstandings or points of disagreement
- Determining how additional evidence affects an argument
- Detecting assumptions made by particular arguments
- Identifying and applying principles or rules
- Identifying flaws in arguments

Form A was taken from Form 9LSS44 of the October 1999 test. Form B was taken from Form 2LSS53 of the October 2002 test. Both were retrieved (with permission) from the Law School Admission Council (2007).

We desired and anticipated that the intellectual sophistication of the experimental groups would vary, and so we needed a measure common to all experiments for their pre-tests. We chose the Logical Reasoning component of the LSAT, with the CCTST also being seriously considered. It has several substantial advantages: unlike most other critical thinking tests, it is widely known outside of psychometric circles and there was universal agreement on the quality of the LSAT questions and the correctness of its answers. Despite these virtues, as argued below, the heavy literacy-loading of the LSAT Logical Reasoning probably lowered the measured increase in critical thinking ability.

Because we were using the LSAT questions to measure critical thinking ability in general, and not critical thinking ability under considerable time pressure, we extended the time to 50 minutes.

3.7.2 California Critical Thinking Skills Test, paper version (CCTST)

The CCTST is a product of Insight Assessment, a division of the University of California Press. It is based upon the 1990 Delphi Report (Facione 1990). It has been published in paper form and an updated version is now available as a computer based test. We used the 1990 version as Form A of the pre- and post-tests, and the 1992 version as Form B. We used the 1990s CCTST because most of the argument mapping

experiments used it and we wanted to directly compare the results to them. The items in the on-line CCTST differed only slightly from the earlier paper versions.

3.7.3 Halpern Critical Thinking Assessment (HCTA; short form)

The HCTA was developed by Professor Dianne Halpern of the Claremont McKenna College. The test offers 25 everyday scenarios and asks open-ended and forced choice questions. It has been validated in many languages as a reliable indicator of the real life choices and inferences people make. There is a long and a short form version of this test.

The long form consists of questions that subjects must first write a few sentences in responses to questions. For example, they are to assess the suggestion about increasing students' GPAs:

> A recent study in a magazine article written for college undergraduate professors showed that the more time students spend chatting online, the lower their GPA. That is, as the number of hours spent chatting online increases, a student's GPA decreases. One suggestion made in this article is that we could increase students' GPAs by restricting their access to online chat programs.

Then, after submitting their sentences, subjects revisit the same target only this time assessing the plausibility of various answers provided:

> A recent study in a magazine article written for college undergraduate professors showed that the more time students spend chatting online, the lower their GPA. That is, as the number of hours spent chatting online increases, a student's GPA decreases. One suggestion made in this article is that we could increase students' GPAs by restricting their access to online chat programs.
>
> **Based on this information, which is the best answer? (Choose one.)**
>
> ☐ Students' GPAs will increase if we restrict internet chat access because the researchers found that as hours spent chatting online increased, the students' GPA decreased.
>
> ☐ Students' GPAs probably will increase if we restrict internet chatting access, but we cannot be certain because we only know that GPA decreases when hours spent chatting online increases, not what happens when hours spent chatting online decreases.

- ☐ There is no way to know if students' GPAs will increase if we restrict internet chatting access because we only know that chatting online and students' GPAs are related, not whether chatting online causes the students' GPAs to change.
- ☐ There will probably be no effect on students' GPAs if we restrict internet chatting access because the magazine is written for college professors, so it is probably biased against students chatting online.

The short form has the same items, but only with the multiple-choice questions.

4 Experimental Results

The detailed experimental results are found in *Section 10*. Appendix: Statistical Data, with additional analyses.

4.1 Anonymized reporting of data for subjects and institutions

We have made the data anonymous not only in regard to the subjects (as required by the IRBs) but also the institutions. Since the latter is more unusual, perhaps a brief explanation may be called for. First, some of the experimental groups were small enough that a suitably ingenious investigator possibly could figure out which student got which result. It is unlikely that this would happen, but anonymizing the data by institutions as well makes it nigh impossible. In this way, the IRB for student anonymity requires the institutions be kept anonymous. Second, some of the experimental results are poor. Whatever the explanation, it is clear that the teachers and students were intelligent, conscientious, and knowledgeable. We were concerned with the possibility that, at some future time, a misguided administrator or academic superior might penalize some teacher for doing worse than the rest of the experimenters and for candidly reporting the results. This would be unfair; anonymizing the institutional results should make that unfairness less likely. Finally, although this was not a formal requirement of the IRBs, we did not want invidious comparisons made across institutions. The pre-test LSAT scores for some institutions were distinctly higher than for others. Informally, we told the experimenters that their institutions would not be shown in a bad light. Reporting the pre-test results by institutions would have violated that commitment.

So in the publically available version of this Report, we have reported the data without reporting the institution where each particular experiment happened.

4.2 Table of Experimental Results

The data for every subject, divided by experiments and by test, along with additional analyses, are found in *Section 10. Appendix: Statistical Data, with additional analyses*.

Since there is some controversy over the correct standard deviation to use calculating effect size, the effect sizes below are calculated on the pre-test SD only and also on the average of pre and post test SDs.

In these analyses, the effect size has been calculated on both the widely used (but biased) Cohen's *d* and the unbiased Cohen's *d*, d_{unb}. For larger samples, the two figures are very similar, but for smaller samples, there is sometimes a reasonably large difference. For example, on the Experiment 4, with its sample of 7, has an effect size 0.764 on (biased) Cohen's *d* and one of 0.664 on unbiased Cohen's *d*. Cumming (2012, 294) explains the difference this way:

> We have been using *d* as our estimate for the population effect size δ, but unfortunately *d* overestimates δ, especially for small samples. Thus *d* is a *biased* estimator of δ. Fortunately, the bias can be removed by multiplying *d* by an adjustment factor, to give an unbiased estimate of δ. An unbiased estimate has a sampling distribution whose mean equals the population parameter being estimated. In other words, if an estimate is unbiased it will on average neither underestimate nor overestimate the parameter.

Although d_{unb} is the appropriate unbiased estimate of effect size, we provide both because the biased Cohen's *d* is the more widely used measure and presenting this statistic will facilitate comparing our results to those elsewhere in the experimental critical thinking literature.

The difference between the standardly used Cohen's *d* and the unbiased Cohen's *d*, d_{unb}, pretty much disappears when sample sizes are large, as they are in the meta-analyses below. Therefore, in our cross-course analysis we only present Cohen's d_{unb} calculated on the average of pre-test SD and post-test SD as standardizer.

The tables below present the results of the study by experimental course. We see that with the exception of Experiment 7, each of these courses produced a meaningful increase in the effect size.

4.2.1 Analysis of data by experiment

Descriptive Statistics

				SD		Mean			
Experiment	**Test**	**N**	**% of improved cases**	**Pre-Test**	**Average Pre/post**	**Pre-Test**	**Post-Test**	**Difference**	**95%CI**
1	HCTA	14	71.4%	12.0	8.9	60.1	68.1	8.1	[0.7, 15.5]
	LSAT	14	92.9%	3.1	3.5	11.9	14.3	2.4	[1.1, 3.8]
2	CCTST	19	84.2%	3.1	3.2	17.2	21.4	4.2	[2.6, 5.8]
	LSAT	17	70.6%	3.4	3.8	12.6	14.4	1.8	[0.4, 3.1]
3	CCTST	10	70.0%	4.3	3.9	21.0	23.9	2.9	[0.4, 5.4]
	LSAT	12	66.7%	4.0	4.4	13.7	15.2	1.5	[-0.5, 3.5]
4	CCTST	7	85.7%	3.0	4.1	17.3	20.4	3.1	[-1.7, 8.0]
	LSAT	7	57.1%	3.8	3.3	14.3	14.7	0.4	[-1.6, 2.4]
5	CCTST	13	84.6%	3.3	2.8	25.2	28.2	3.0	[1.2, 4.8]
	LSAT	15	53.3%	3.2	3.4	16.6	18.1	1.5	[0.0, 3.0]
6	HCTA	38	81.6%	7.4	6.6	67.2	72.0	4.8	[-5.1, 13.8]
	LSAT	39	61.5%	4.1	4.0	14.3	15.7	1.4	[0.2, 2.5]
7	HCTA	24	50.0%	4.0	5.0	71.4	71.5	0.0	[-9.0, 9.1]
	LSAT	23	39.1%	3.7	3.9	16.3	16.1	-0.2	[-1.6, 1.1]

Standardized Effect Sizes

		Using Pre-Test SD			Using Pre/Post Av SD		
Experiment	**Test**	***d***	d_{unb}	**95% CI d_{unb}**	***d***	d_{unb}	**95% CI d_{unb}**
1	HCTA	0.674	0.634	[0.064, 1.723]	0.908	0.854	[0.064, 1.723]
	LSAT	0.775	0.729	[0.245, 1.143]	0.703	0.661	[0.245, 1.143]
2	CCTST	1.350	1.293	[0.663, 1.923]	1.303	1.248	[0.663, 1.923]
	LSAT	0.515	0.490	[0.094, 0.815]	0.460	0.439	[0.094, 0.815]
3	CCTST	0.671	0.614	[0.071, 1.373]	0.737	0.674	[0.071, 1.373]
	LSAT	0.372	0.346	[-0.092, 0.766]	0.344	0.320	[-0.092, 0.766]
4	CCTST	1.034	0.899	[0.048, 1.442]	0.764	0.664	[0.048, 1.442]
	LSAT	0.112	0.098	[-0.368, 0.622]	0.132	0.115	[-0.368, 0.622]
5	CCTST	0.906	0.848	[0.341, 1.745]	1.057	0.989	[0.341, 1.745]
	LSAT	0.452	0.427	[-0.007, 0.863]	0.435	0.411	[-0.007, 0.863]
6	HCTA	0.646	0.633	[0.417, 1.021]	0.722	0.707	[0.417, 1.021]
	LSAT	0.334	0.327	[0.094, 0.573]	0.336	0.329	[0.094, 0.573]
7	HCTA	0.010	0.010	[-0.453, 0.470]	0.008	0.008	[-0.453, 0.470]
	LSAT	-0.058	-0.056	[-0.382, 0.272]	-0.055	-0.054	[-0.382, 0.272]

4.2.2 Untrimmed and trimmed data

For individual experimental results, there is sometimes a substantial effect size difference between the trimmed and untrimmed results. In every case, this is an artifact of the trimming technique we adopted, and different trimming techniques with different cutoffs would give different results. However, we found that when the tests were combined in meta-analyses, the difference between untrimmed and trimmed data was relatively small. The largest was for the CCTST, which had an effect size of 0.847 [0.571, 1.123] untrimmed and 0.966 [0.744, 1.188] trimmed. But, even the trimmed data is not above 1.0 SD and for the tests are roughly similar, so the large effect sizes on the tests are not due to a subset of subjects being ill or particularly disengaged when taking their pre-tests. For details of the trimming method and the results, see Appendix 10.3.

4.2.3 Difference between pre-course subject ability, across individuals and across institutional settings

The pre-test results showed significant differences in critical thinking ability within each group and between several of the groups.

		Pre-Test		
Experiment	**Test**	**N**	**Mean**	**SD**
1	HCTA	14	60.1	12.0
	LSAT	14	11.9	3.1
2	CCTST	19	17.2	3.1
	LSAT	17	12.6	3.4
3	CCTST	10	21.0	4.3
	LSAT	12	13.7	4.0
4	CCTST	7	17.3	3.0
	LSAT	7	14.3	3.8
5	CCTST	13	25.2	3.3
	LSAT	15	16.6	3.2
6	HCTA	38	67.2	7.4
	LSAT	39	14.3	4.1
7	HCTA	24	71.4	4.0
	LSAT	23	16.3	3.7

4.3 Gender Differences

As one would expect with such small samples, sometimes there were very few women or men in an experiment, producing dramatic but misleading differences in percentages gender difference percentages. For example, in Experiment 3, 0% per cent of women improved on the CCTST, whereas 78% of men did. In the same experiments, 100% of the women's LSAT scores improved. However, as there was only one woman in the group, this is not strong evidence of anything at all. Analysing gender differences across experiments is more informative.

4.3.1 Comparison of experiment results by gender

For CCTST, HCTA and LSAT, the difference between females and males in pre-post scores is negligible. For every test, the 95% CIs on female and male mean scores overlap considerably, and *p* values on differences (female-male) do not even remotely approach statistically significant thresholds.

Test	Gender	N	Pre-Post Mean Difference	95% CI	Female-Male Mean Difference (Standardized ES)	*p*-value
CCTST	F	17	3.59	[1.61, 5.57]	0.18 (0.06)	0.85
	M	32	3.41	[2.54, 5.85]		
HCTA	F	31	4.68	[2.42, 6.93]	1.66 (.21)	0.37
	M	46	3.02	[0.36, 5.68]		
LSAT	F	48	1.08	[0.34, 1.43]	-0.22 (-0.08)	0.67
	M	79	1.30	[0.62, 1.99]		

While pre-post improvement is equivalent for males and females, there were non-trivial gender differences on pre-score measures. Female mean scores were lower in all three tests, approximately one third of a SD (HCTA) to half a SD (CCTST and LSAT) lower than male scores. Small sample sizes mean these differences did not always reach statistical significance (CCTST and HCTA, see table directly below), but regardless of their statistical non-significance the differences seem potentially important. The difference in p values may well simply reflect the differences in N for each test.

Test	Gender	N	Pre-score Mean	95% CI	Female-Male Mean Difference (Standardized ES)	*p*-value
CCTST	F	17	18.5	[16.0, 21.0]	-2.9 (-0.54)	0.15
	M	32	21.0	[19.4, 22.5]		

HCTA	F	31	65.6	[63.1, 68.1]		
	M	46	68.5	[65.8, 71.2]	-2.5 (-0.34)	0.079
LSAT	F	48	13.2	[12.0, 14.3]		
	M	79	15.1	[14.3, 16.0]	-1.97 (-0.51)	0.006

5 Discussion

5.1 Overall discussion

The project did not achieve a reliable 1+ SD improvement across tests and experiments. This is true regardless of whether one uses untrimmed[6] or trimmed data, uses Pre-test SD or Pre/Post test average SD, or whether one includes or excludes the results from Experiment 7 (Extraordinary Scrutiny), or the results from the LSAT Reasoning test. If the HCTA and the CCTST are regarded as better measures of critical thinking than the LSAT, the degree of average improvement were a large number of courses. If, however, the LSAT is taken as the better measure of critical thinking, then our goal was seldom approached.

Here is the untrimmed data again. The standardized, Cohen's d_{unb} , results on individual tests using the average Pre/Post SD average, ranged from 1.29 SD to -0.05 SD:

		Using Pre-Test SD		**Using Pre/Post Av SD**	
Experiment	**Test**	d_{unb}	**95% CI**	d_{unb}	**95% CI**
1	HCTA	0.634	[0.064, 1.723]	0.854	[0.064, 1.723]
	LSAT	0.729	[0.245, 1.143]	0.661	[0.245, 1.143]
2	CCTST	1.293	[0.663, 1.923]	1.248	[0.663, 1.923]
	LSAT	0.490	[0.094, 0.815]	0.439	[0.094, 0.815]
3	CCTST	0.614	[0.071, 1.373]	0.674	[0.071, 1.373]
	LSAT	0.346	[-0.092, 0.766]	0.320	[-0.092, 0.766]
4	CCTST	0.899	[0.048, 1.442]	0.664	[0.048, 1.442]
	LSAT	0.098	[-0.368, 0.622]	0.115	[-0.368, 0.622]
5	CCTST	0.848	[0.341, 1.745]	0.989	[0.341, 1.745]
	LSAT	0.427	[-0.007, 0.863]	0.411	[-0.007, 0.863]
6	HCTA	0.633	[0.417, 1.021]	0.707	[0.417, 1.021]
	LSAT	0.327	[0.094, 0.573]	0.329	[0.094, 0.573]
7	HCTA	0.010	[-0.453, 0.470]	0.008	[-0.453, 0.470]
	LSAT	-0.056	[-0.382, 0.272]	-0.054	[-0.382, 0.272]

[6] For details of the trimming data and method, see *Section 10.3 Trimmed Data*.

As the following table shows, the data on a course-by-course basis is not much different from the combined scores from students taking the same test regardless of the course they were in. Since the meta-analysis results are roughly the same whether the pre-test SD or the average pre/post test SD is used, for ease of presentation below we will simply discuss the latter.

The LSAT results are generally considerably lower than the CCTST and HCTA results, and the Extraordinary Scrutiny (Experiment 7) results much lower than the results from the Normal Scrutiny (Experiments 1 to 6) results:

Using Pre/post Test SD

Effect Size Measure	Test	Experiment 7		Experiments 1 to 6		All Expts	
		Effect Size	95% CI	Wgtd Avg	95% CI	Wgtd Avg	95% CI
d_{unb}	CCTST	-	-	0.847	[0.571, 1.123]	0.847	[0.571, 1.123]
	HCTA	0.008	[-0.453, 0.470]	0.721	[0.464, 0.977]	0.539	[0.317, 0.760]
	LSAT	-0.054	[-0.382, 0.272]	0.370	[0.238, 0.502]	0.307	[0.185, 0.428]
	All	-0.033	[-0.288, 0.223]	0.505	[0.397, 0.613]	0.424	[0.324, 0.523]

Trimming the data set by 5% at both the high and low ends to exclude outliers does little to change the picture. For details see Appendix 10.3.

There were non-trivial gender differences in mean pre-scores for all three tests, with mean female scores between one third and one half a SD lower than mean male scores. However, the mean improvement (pre-post difference) afforded by the intervention was equivalent in male and female participants. Effect sizes for gender differences were extremely small (0.2 SD or less), and did not approach statistical significance.

We propose an explanation for the Experiment 7 and LSAT anomalies below.

There are several factors which no doubt played some role in the project's not achieving a robust 1+ SD improvement. But there is no reason to believe that this project was unusual in facing them; instead, they should be viewed as impediments any critical thinking education research should take into account. That is, these are factors that *mutatis mutandis* would impede the achievement of substantial improvements in critical

thinking ability in future experiments, or impede the transfer of proven techniques to the classroom, or both:

- *Some subjects did not work very hard*. Some students were members of the military or intelligence community, and were unable to commit sufficient time to the AM practice required to show a result. Others appeared to not see the activity as very relevant to their work and thus not worth much effort. In other experiments undergraduates, being undergraduates, had more exciting things to think about than unstated co-premises. Some students just didn't care enough to put in the requisite time and thought.
- *Some instructors might not have fully absorbed the underlying purpose and philosophy of CTAM* Some instructors were philosophically trained in the traditional styles of teaching critical thinking, and may have somewhat continued to teach in the old manner despite using the techniques of CTAM. This could have reduced the effects of CTAM.
- *There was selection bias in the subjects at some institutions* Some students were students of disciplines like philosophy that require high levels of critical reasoning, and so they may have been already primed to learn the techniques and to be interested in the material. Some students and analysts took the workshop semi-involuntarily, and at least a few were not particularly interested in mastering the requisite skills, seeing them as irrelevant to their education, lives, or jobs.
- *In educational research generally, student incentive to do well on the pre and post tests can considerably determine the effect size*. A 2012 study by Liu et al., "Measuring Learning Outcomes in Higher Education" (Liu, Bridgeman, and Adler 2012), has some useful data, although the basic effect should not be surprising:

> With the pressing need for accountability in higher education, standardized outcomes assessments have been widely used to evaluate learning and inform policy. However, the critical question on how scores are influenced by students' motivation has been insufficiently addressed. Using random assignment, we administered a multiple-choice test and an essay across three motivational conditions. Students' self-report motivation was also collected. Motivation significantly predicted test scores. *A substantial performance gap emerged between students in different motivational conditions (effect size as large as .68). Depending*

> *on the test format and condition, conclusions about college learning gain (i.e., value added) varied dramatically from substantial gain (d = 0.72) to negative gain (d = −0.23).* The findings have significant implications for higher education stakeholders at many levels. [Our italics]

While Liu's study did not include students taking a class for a grade, its lesson is important for setting up future CT educational experiments and for evaluating experiments in the CT literature.

While these factors may well have played some role in this project's not reaching a robust 1+ SD improvement on critical thinking, they are difficulties that go with the territory, *not* reasons to believe that the project really did reach 1+ SD. They will play a role in almost any critical thinking educational research and, indeed, almost all educational research. Further, some of them interfere with almost all critical thinking classes, often substantially.

Turning now to two anomalies that we did not expect:

- Experiment 7, the Extraordinary Scrutiny one, did far worse than the other six experiments and indeed worse not only than almost all other previous argument mapping experiments but even than standard critical thinking courses.
- The improvements on the LSAT Reasoning were distinctly less than that on the CCTST and HCTA, regardless of whether the results of Experiment 7 are included in the analysis.

As statistical tests reported below demonstrate, these anomalies are not due to statistical noise. There is no reason to think that they are primarily due to different testing situations, or to differences in instructors' or subjects' abilities. Below, we will look at the tests themselves to explain, in large part, these anomalies.

We are *not* saying in any way that this projected failed to robustly reach 1+ SD deviation problem because of inadequacies with standard CT tests. The LSAT is a superb test and the HCTA and CCTST are more than adequate for our purposes. Rather, we are proposing that they have unexpected aspects which are of important for future critical thinking research.

5.2 Extraordinary Scrutiny, Critical Thinking Classes and Standard Critical Thinking Tests

It was obvious early on that the Extraordinary Scrutiny approach of Experiment 7 differed substantially from the scrutiny applied to co-premises in the other six experiments as well as in previous argument mapping courses. But, it was not obvious what the effect would be. As befits its experimental nature, all three possibilities were open: the Extraordinary Scrutiny might improve the scores on standard critical thinking tests, might make little or no difference, or might substantially decrease the scores. The data are unambiguous:

	Using Pre/Post Av SD					
	Normal Scrutiny (Experiments 1 to 6)		Extraordinary Scrutiny (Experiment 7)		All Expts	
	Stand'ised ES	95% CI	Stand'ised ES	95% CI	Stand'ised ES	95% CI
CCTST	0.847	[0.57, 1.12]	-	-		[0.57, 1.12]
HCTA	0.721	[0.46, 0.98]	0.008	[-0.45, 0.47]	0.539	[0.32, 0.76]
LSAT	0.370	[0.24, 0.50]	-0.054	[-0.38, 0.27]	0.307	[0.18, 0.43]
All	0.505	[0.40, 0.61]	-0.033	[-0.29, 0.22]	0.424	[0.32, 0.52]

Clearly, the scores were substantially lower, compared with the other six experiments in the project. Could this difference plausibly be due to random noise? The following table strongly indicates that this is probably not the explanation, although it might be part of it. It reports the means and 95% CIs for Studies 1 to 6 combined by meta-analysis, for Experiment 7 alone, and for the difference between Studies 1 to 6 and Experiment 7. Results are provided for LSAT scores, for either CCTST or HCTA scores – just one of these was used in any study – and for the average of LSAT and either CCTST or HCTA. The measure is unbiased Cohen's d, d_{unb}, calculated using the average of pre-test SD and post-test SD as standardiser.

	Normal Scrutiny (Experiments 1 to 6)			Extraordinary Scrutiny (Experiment 7)			Difference between Expts 1 to 6, and Expt 7			
	Wtd Avg *d*unb	95% CI		*d*unb	95% CI		Wtd Avg *d*unb	95% CI		p[1]
LSAT	0.37	[0.23,	0.51]	-0.05	[-0.38,	0.27]	0.42	[0.07,	0.7 0.78]	.02
CCTST or HCTA	0.80	[0.60,	1.00]	0.01	[-0.45,	0.47]	0.79	[0.29,	1.29]	.002

Av. Of (LSAT, and CCTST or HCTA)	0.58	[0.41,	0.75]	-0.02	[-0.42,	0.37]	0.60	[0.17,	1.03]	.006

[1]Two-tailed *p* value for the difference between Experiments 1 to 6, and Experiment 7.

The low p-value of 0.006 follows from the 95% interval of [0.17, 1.03]. It is very unlikely that the difference between Experiment 7's scores and those of Experiments 1to 6 is due to chance.

Further, the results of Experiment 7 are also dramatically different from the results of previous argument-mapping educational experiments. From van Gelder's meta-analysis below, we find that High Intensity Argument Mapping courses have an average effect size of 0.84 [069, 0.99]. All experiments in this project were definitely High Intensity on van Gelder's criterion. None of the experiments studied by van Gelder used the LSAT. Even if one looks at all argument mapping courses, including low intensity ones, the effect size was 0.57 [0.42, 0.72]. The results of Experiment 7 are not at all close to falling within the 95% CI of the previous argument mapping literature either.

Thus, Experiment 7's results were so markedly and consistently different from those in Experiments 1–6 and previous research, that it is necessary to consider its results separately. Why are they so different? Here is our best explanation, given that the instructor is an extremely bright philosopher and a very dedicated teacher, and the subjects are highly talented and most of them worked hard. The poor results of Experiment 7 are not due to superficial features of instructors or subjects.

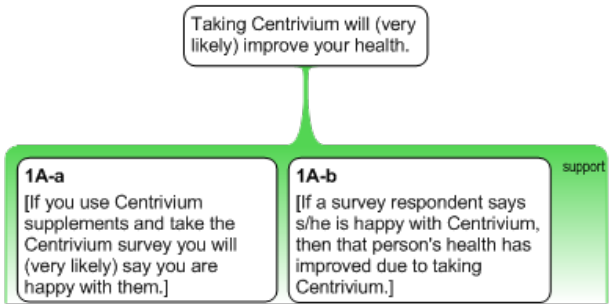
As mentioned above, subjects in Experiments 1 through 6, in accord with previous critical thinking research, learned to scrutinize co-premises for *prima facie* plausibility. Let's call this Normal Scrutiny. It contrasts with Extraordinary Scrutiny, which involves unpacking even plausible co-premises into multiple plausible but potentially false evidential and logical presuppositions and scrutinizing these. From almost the beginning of Experiment 7, there were many discussions about the advantages and disadvantages of what we are now calling Extraordinary Scrutiny. It was clear that Experiment 7 was testing an importantly different hypothesis: *A critical thinking course based on argument mapping + peer instruction + mastery learning could be further*

improved by practicing and habituating the intensified analysis suitable to situations where arguments merit extraordinary scrutiny.

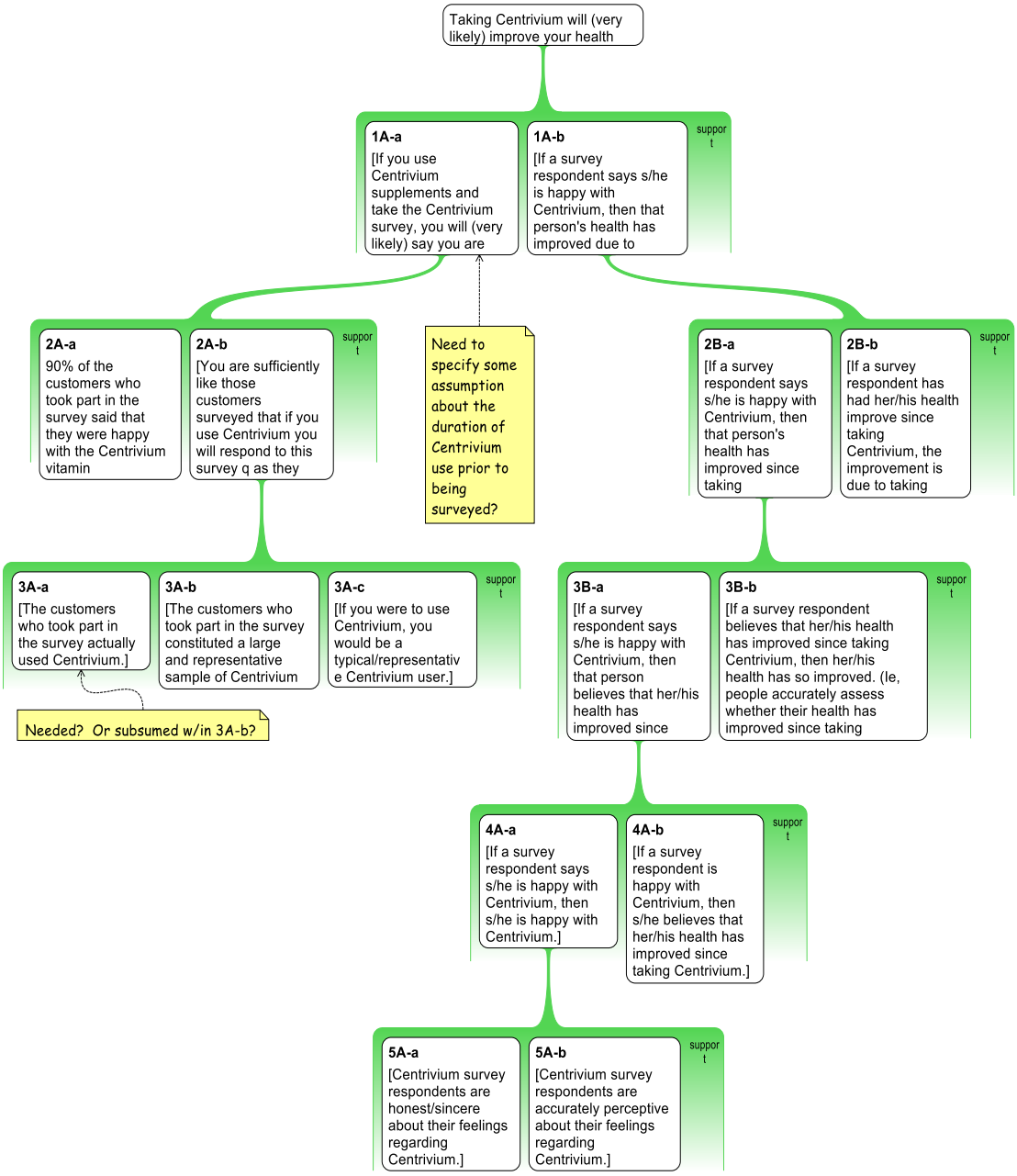
Up to now, the distinction between Normal and Extraordinary Scrutiny has been described abstractly. Here are two examples of Normal and Extraordinary Scrutiny, as found in classroom in the project. Here is Argument 5 from Module 8:

> 90% of the customers who took part in the survey said that they were happy with the Centrivium vitamin supplements. So you can buy Centrivium with confidence that it will improve your health.

Here is how it was mapped in the Normal Scrutiny classes. While the exact wording differed from class to class (and indeed student to student), the maps in these six experimental situations were similar:



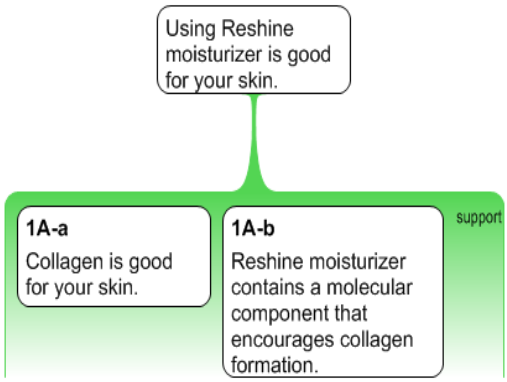
Here is how it was mapped in the Extraordinary Scrutiny Experiment 7:



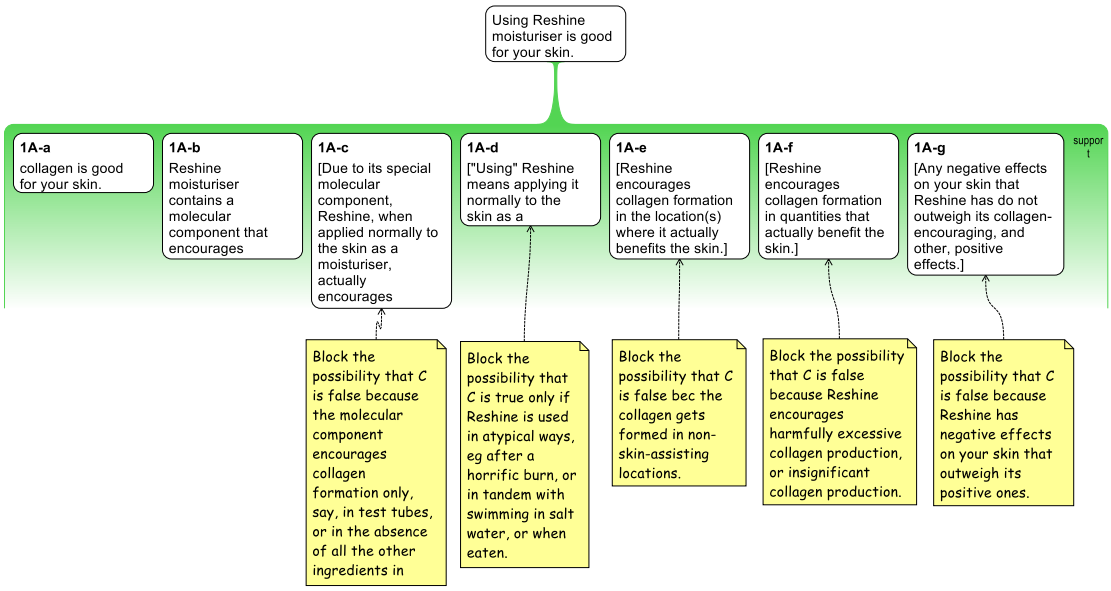
There is no doubt that this is a much more thorough analysis of this argument and there can be little doubt that it took much more time. Now consider argument 5 from Module 7, the Reshine Moisturiser:

> Using Reshine moisturiser is good for your skin. Reshine moisturiser contains a molecular component that encourages collagen formation. As we all know, collagen is good for your skin.

Here is a typical Normal Scrutiny map:



Here is the Extraordinary Scrutiny map produced in Experiment 7:



In both of these cases, and many others, Extraordinary Scrutiny took far longer and went into far more detail than Normal Scrutiny. Excellent though it is for some situations, it is not clear that a subject taking a critical thinking test is one of those situations.

The Extraordinary Scrutiny orientation is not entirely alien to ordinary CT education. To the contrary, it is often briefly touched upon in CT textbooks and a case can be made that Alec Fisher's renowned *The Logic of Real Arguments* (Fisher 1988) is an example of Extraordinary Scrutiny. But such nods toward heightened scrutiny are aimed largely

at arousing and maintaining student interest and motivation, as when texts make the case that CT is generally important and gratifying via specific illustrations showing the need for (in effect) Extraordinary Scrutiny when dealing with advertising and political rhetoric.

In these standard CT textbooks, the vague and situation-dependent boundary between Extraordinary Scrutiny and Normal Scrutiny is not discussed, nor is the distinction even given its due acknowledgement. More seriously, the hard issue of when Extraordinary Scrutiny is appropriate is ignored, probably in part because it depends on practical factors not relevant to most CT courses: the seriousness of the consequences if the conclusion is false; the opportunity costs in pursuing laborious Extraordinary Scrutiny; the need to acquire further data; the grounds for suspecting that superficially competent reasoning might be corrupt; and the determination of the border between reasonable Extraordinary Scrutiny and wasteful hyper-skepticism.

At this point, then, Extraordinary Scrutiny is not far developed as a distinctive approach in CT education, or even well recognized as such. Yet, undoubtedly, such a stance is properly called for in many contexts, as when an intelligence analyst assesses the chances that someone is a double-agent or that a foreign government is doing something importantly different than what it says it is doing, or when a lawyer scrutinizes the opponent's carefully-worded deposition, or when a careful layperson senses the possibility of a fraudulent seduction or con job, or when a consumer advocate monitors for deceptive sales practices, or when a scientist is evaluating the evidence for and against a theory, or when … .

Of course, in the ordinary course of life it is usually not worth the resources to carefully investigate each argument that is encountered, because the consequences of accepting an unsound argument are minor enough or the likelihood of error in the reasoning is low enough. And even where it would be worth the resources, Extraordinary Scrutiny may be practically impossible or the resources needed for it may be needed more elsewhere. Still sometimes one does need to judge such factors, and improving such judgment calls is not easy to teach; demonstrating that one has improved them would be even harder.

Some of the situations dictating Extraordinary Scrutiny are cases where, although the final conclusion and the reasoning seem plausible, the consequences of accepting the conclusion if it should turn out to be wrong are particularly serious, and thus the reasoning deserves close double-checking. More common are cases where the consequences of accepting a false conclusion are serious but not so serious as to call for automatic Extraordinary Scrutiny, yet where there is also good reason to doubt the soundness of the reasoning despite its superficial credibility – for instance, if the final conclusion is implausible, or if circumstances give cause for suspicion regarding the trustworthiness (honesty or competence) of the arguer. Again, determining whether a suspicion is reasonable or excessive is a complex judgment call that is not generally open to mechanical decision.

The complexities involved at the extreme end of the Extraordinary Scrutiny spectrum can certainly be immense. Consider this intelligence example below, from the work of Schaum *et al* (2009) on chains of evidentially relevant custody. Note the many points at which doubts can be raised about the fidelity of the eventual report to the initial evidence. And this example covers only one alleged fact out of many in typical intelligence cases. Similar complexities arise in scientific arguments whose every step may involve inferences about competence, credibility, veracity, and candor.

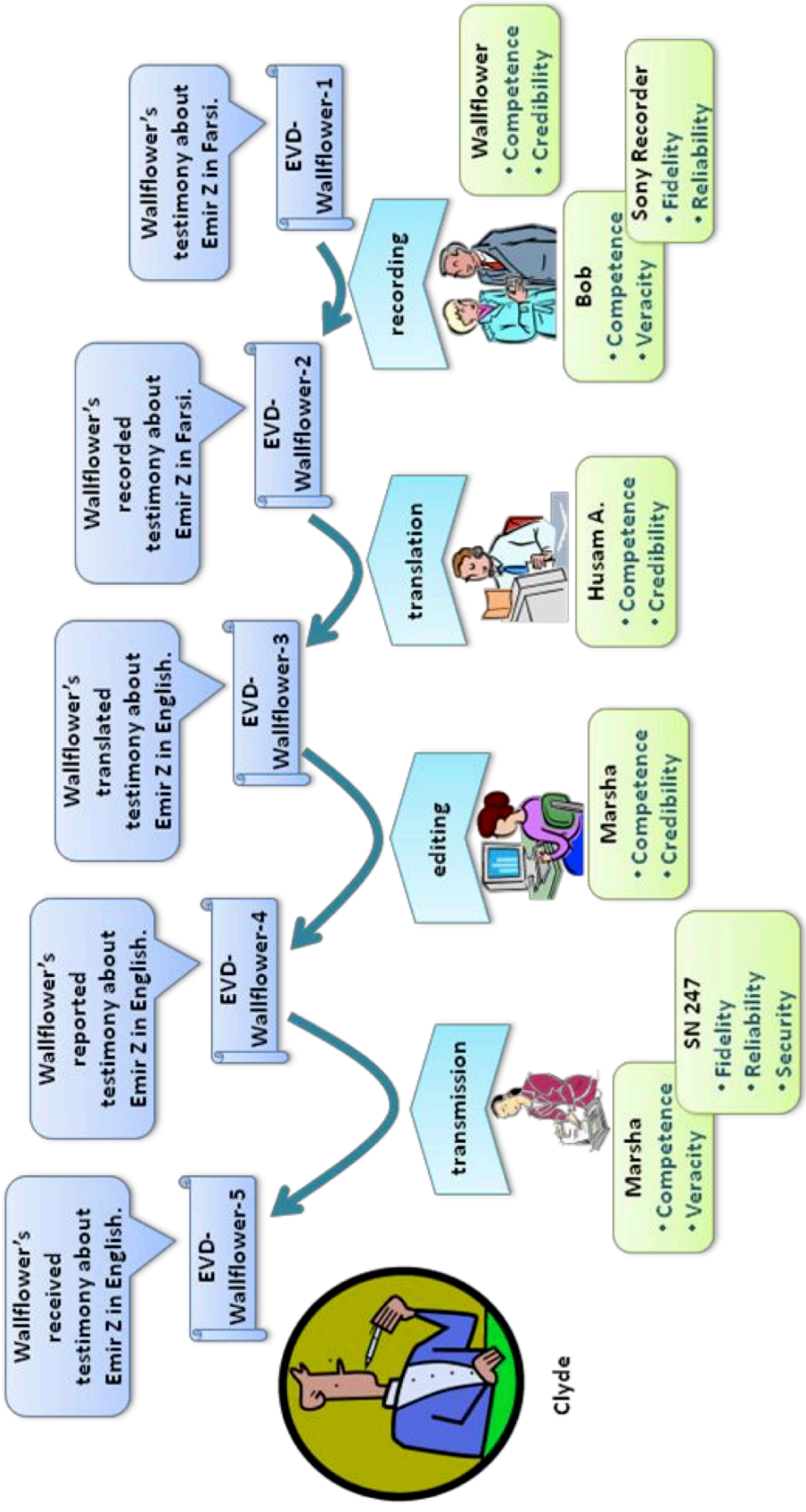


Figure 1: Chain of custody of Wallflower's testimony and the processes in this chain.

It is easy to see how, if the Extraordinary Scrutiny approach were to spread uncontrolled throughout everyday life, it would gravely undermine our functioning. At the grocery store, we normally and reasonably accept the cash register's reported total, although we could double check that all-and-only the right items are being charged for and that the addition is done correctly. Automobile associations don't recommend double-checking our car engines each morning to ensure that bombs are not present. More extreme versions of Extraordinary Scrutiny would be too onerous as default positions even in legal, scientific, or intelligence matters. Beyond a certain point, Extraordinary Scrutiny can become a debilitating hyper-skepticism. Determining that point (or, more accurately, the range) is a judgment call that, except in extreme cases, may not be easily taught.

Yet, despite caveats about the difficulties in teaching how to make good judgment calls, the Extraordinary Scrutiny orientation, and the development of associated CT skills, might be of particular value for the Intelligence Community. It certainly is needed periodically, the Weapons of Mass Destruction episode being only one of many.

Thus it is fortuitous the way that Experiment 7 was unexpectedly conducted by an instructor who has the background preparation, personal inclination, and habitual practice that are especially suited for teaching an argument-mapping-based CT course with a considerable emphasis on Extraordinary Scrutiny. (Unfortunately, presenting the details of his/her training would violate the Project's commitment to institutional anonymity.)

Combining this instructor with the Project's course materials produced a CT course whose argument-mapping activities and class discussions were distinctively detailed and rigorous, certainly more so than the other six experiments and the other critical thinking experiments we know of, argument-mapping-based or not. Plausible and sometimes even implausible but interesting objections to co-premises were raised, and in turn evaluated. This took considerable class time, and fascinated some while it bored the more "practically-minded" subjects. Experiment 7 test outcomes, as noted above, were undoubtedly poor. The data, however, are compatible with at least two explanations:

1. The tests didn't detect CT improvement because there was little if any to detect, maybe due to the instructor's distinctive approach and/or incompetence and/or …
2. There was a substantial CT improvement that these tests could not detect, in part because they do not measure Extraordinary Scrutiny ability and in part because so carefully scrutinizing test items slowed the subjects down in part by absorbing considerable working memory.

The former explanation is belied by general considerations – the strong evidence that most critical thinking courses, particularly argument-mapping-based ones, have substantial effects. And by specific considerations – the instructor's obvious competence, intelligence, and classroom ability, as well as the obvious commitment of most Experiment 7 subjects and their informally apparent CT improvements across the semester-long course. Given the training that they had received, these subjects could reasonably have assumed that Extraordinary Scrutiny would be expected, and helpful, on the post-test. They were unaware that these kinds of CT tests do not call for, or reward, Extraordinary Scrutiny. As a result, many of them may well have invested too much time and working memory into needlessly detailed scrutiny of the test items. Further, sometimes the overly close scrutiny might have produced non-standard readings of question stems, questions, or answer options. But however plausible, this explanation is only conjectural, and calls for further experimentation.

Whichever explanation is correct, Experiment 7 has special relevance to IARPA and the purpose for which this Project was undertaken. Because Extraordinary Scrutiny is so central to key parts of the IC's tasks, its lessons should be further explored:

- Present critical thinking tests do not test for Extraordinary Scrutiny skills. Such a test is clearly needed, ideally as a part of measuring the other critical thinking skills, and ought to be developed. But developing one might not be easy, not the least because it is unclear how to operationalize some key questions, such as How should a test-taker be expected or required to choose which *prima facie* plausible co-premises to scrutinize so carefully? and How careful should that scrutiny be in the test situation, where the only opportunity cost is time spent?

- Unpacking a *prima facie* plausible co-premise's causal and conceptual presuppositions is not a mechanical task, since it depends on eliciting the most plausible of the generally implausible objections while neglecting hyper-skeptical objections. The important distinctions between Normal Scrutiny and Extraordinary Scrutiny and between reasonable Extraordinary Scrutiny and Excessive Scrutiny are not sharp. No doubt their applications depend on the situation, including the probability that someone is deliberately attempting to deceive and the importance of knowing the truth of the matter.
- While Extraordinary Scrutiny is taught in law schools and elsewhere, we do not know how well the lessons are learned or whether there are better ways of teaching it. We suspect that argument maps might play a key role here, but this use of them has not been explored.
- Serious consideration should be given to developing an IC course in CT that contains substantial elements of the Extraordinary Scrutiny approach, with an emphasis on when to use it and when not.

5.3 Using the LSAT to measure changes in critical thinking ability

As mentioned above, we anticipated that the intellectual sophistication of the experimental groups might vary considerably and so decided to use one test in all experiments, so we could measure the relative pre-test abilities of the various groups of subjects. We chose the Logical Reasoning component of the LSAT for these measurements, and in that regard our choice was a good one.

However, we failed to adequately appreciate that its very excellence for law school admissions purposes meant that it was less good for ours in measuring changes in critical thinking ability. To be a good law student and lawyer, one must not only be a good critical thinker, one must also be able to read large amounts of complex material rapidly, retaining major points and many details. Obviously, analysts need analogous skills. Even more obviously, these are not skills that a critical thinking course can do that much about; we never expected to substantially improve reading speed, general literacy, and working memory. So, insofar as the LSAT measures them, any

measurement of improved critical thinking *per se* would be diminished by the high literacy factors which remained more or less constant.

Across all three tests (CCTST, HCTA and LSAT) the combined standardized ES for all 7 experiments was 0.424 [0.32, 0.52]. Whilst this didn't reach the 1+SD expected, it is non-trivial. As mentioned elsewhere, removing the results from Extraordinary Scrutiny Experiment 7 raises the ES to 0.505 [0.40, 0.61]. Further removing LSAT leaves ESs in the range expected from previous literature (0.721 HCTA and 0.847 CCTST). For example, the HCTA and CCTSTS ESs are comparable to Van Gelder's meta-analysis ES of 0.84, all of which were Normal Scrutiny and none of which used the LSAT to measure the changes in critical thinking ability.

Here is a typical LSAT question, from the project's Form A:

> Columnist: It is impossible for there to be real evidence that lax radiation standards that were once in effect at nuclear reactors actually contributed to the increase in cancer rates near such sites. The point is a familiar one: who can say if a particular case of cancer is due to radiation, exposure to environmental toxins, smoking, poor diet, or genetic factors.
>
> **The argument's reasoning is most vulnerable to criticism on which one of the following grounds?**
> (A) The argument fails to recognize that there may be convincing statistical evidence even if individual causes cannot be known.
> (B) The argument inappropriately presupposes that what follows a certain phenomenon was caused by that phenomenon.
> (C) The argument inappropriately draws a conclusion about causes of cancer in general from evidence drawn from a particular case of cancer.
> (D) The argument ignores other possible causes of the increase in cancer rates near the nuclear reactor complexes.
> [E) The argument concludes that a claim about a causal connection is false on the basis of a lack of evidence for the claim.

Compare that to a typical multiple choice question from the HCTA:

> In order to try to reduce violence in middle school for delinquent children, one parents' group suggested that some students be included on the board that writes the school violence reform proposal. The principal of the middle school

> was totally opposed to the idea saying that it was like having the mentally ill write the rules for their mental institution.
>
> **In his analogy, what is the principal assuming? (Choose as many as apply.)**
> ☐ Students are not able to work well with others so they could never agree on rules.
> ☐ Students belong in a mental institution.
> ☐ Students are like mental patients and cannot be trusted to write rules they would have to follow.
> ☐ Students don't respect the rules of the school now, so they won't respect new rules either.
> ☐ Students are incapable of making good decisions when they work with the adults on the committee.

Again, compare both of those to a typical CCTST question, italics in original:

> Consider these statements true: "Stylish dressers are neither flashy nor dull. If someone is not flashy, then such a person is tasteful." Which of the following *must* be true, if both of the above are true?
> A = Stylish dressers are neither tasteful nor dull.
> B = If someone is a stylish dresser, that person is dull but tasteful.
> C = Every stylish dresser is tasteful and not dull.
> D = No stylish dressers are dull.
> E = None of the above.

The key relevant difference here is the considerably higher level of background literacy required for many LSAT questions. In the LSAT target and possible answers above we have: "lax radiation standards", "cancer rates", "environmental toxins", "genetic factors", "statistical evidence", "nuclear reactor complexes", and "causal connection". Many other LSAT questions also require analogous levels of sophistication. For example, another question from the same test discussed "indigenous people", "Tasmania", "land bridge" and "disappearance of the land bridge", "spear thrower", "boomerang", "technological innovations" and "polished stone tools".

In addition to such vocabulary differences, standard readability tests show substantial differences among the critical thinking tests. These test results are calculated on the basis of such items as words per sentence, but not on the difficulty of the words themselves. The LSAT reading ease measure is around 50% of the HCTA and CCTST, and the grade level

required to understand the LSAT questions is around second year college level, where the HCTA is around Grade 10, and the CCTST is Grade 6, for the questions above:

5.3.1 CT Tests -- Reading Ease, Grade Levels, and Test Statistics

A higher Flesch-Kincaid Reading Ease score indicates easier readability; scores usually range between 0 and 100, but very complex texts can have a negative score and very easy ones can go beyond 100.

Readability Formula	**LSAT**	**HCTA**	**CCTST**
Flesch-Kincaid Reading Ease	38.1	63.1	75

A grade level (based on the USA education system) is equivalent to the number of years of education a person has had. Scores over 22 should generally be taken to mean graduate level text.

Readability Formula	**LSAT**	**HCTA**	**CCTST**
Flesch-Kincaid Grade Level	13.4	8.4	5.4
Gunning-Fog Score	17.4	11.1	7
Coleman-Liau Index	13.7	12.2	8.1
SMOG Index	12.7	8.1	6
Automated Readability Index	13.2	9	2.9
Average Grade Level	**14.1**	**9.8**	**5.9**

	LSAT	**HCTA**	**CCTST**
Character Count	887	685	337
Syllable Count	306	217	119
Word Count	177	144	83
Sentence Count	8	9	8
Characters per Word	5.0	4.8	4.1
Syllables per Word	1.7	1.5	1.4
Words per Sentence	22.1	16.0	10.4

Although it would undoubtedly be a good thing if all students and analysts had the level of sophistication the LSAT questions presuppose, many clearly do not. Or, more precisely, the LSAT regularly uses terms such as “statistical evidence” and “land bridge” that are not part of many subjects’ every day working vocabularies and its sentences are longer. For such subjects, answering LSAT-type questions will take more time, and further tax their working memory.

With occasional exceptions, questions on the HCTA and the CCTST do not require the sophisticated background that the LSAT often does, and so those tests should measure any changes in critical thinking itself more accurately. In effect, they are not nearly as strongly constrained by literacy, working memory, and background knowledge.

To repeat, this is *not* a weakness in the LSAT; the more we worked with it, the more we appreciated its considerable virtues for its set purpose of assisting in law school admissions. We think it is, overall, an excellent measure of critical thinking ability as well as reading speed and vocabulary needed for law school admission requirements. Its weakness for us was in using it to measure changes in critical thinking ability. Since there is little or nothing in argument mapping (or indeed in most critical thinking courses) to improve literacy, working memory, etc., the test functionally caps how much the LSAT can measure any courses improvements in subjects' critical thinking.

In addition, there are some other substantial differences between the LSAT test questions and what the course did. Some LSAT questions use aspects of critical thinking that aren't addressed in the course materials, at least not directly. These now seem to be weaknesses in the course, one that should be addressed in future iterations.

- While the LSAT questions don't use phrases like "deductively implies" or "deduced from", they do use lots of euphemisms for deduction; they talk about what conclusion 'must' be true or false if certain claims are true, what might be 'logically inferred' from certain claims if true, etc. Many LSAT examples involve deductive validity. Since the project textbook studiously avoids reference to any distinction between deductive and non-deductive arguments, even such euphemisms were avoided. Future development of the course should address this issue.
- In a related vein, some LSAT questions ask what assumption an argument in a passage depends on. The idea seems to be to get people to pick out what is (logically) necessary to get to the conclusion as distinct from something that would, e.g., suffice or help to get to the conclusion. Both would involve picking out unstated material but the LSAT is making the distinction between what is strictly needed and what would be sufficient and, different again, what would help. In the textbook, there was no emphasis on these distinctions. There are no quiz questions of this type.

- LSAT questions sometimes raise 'inference to the best explanation' issues. For example, they sometimes ask which answer, if true, would most strongly support the claim in a text. While there is no mention of explanation, the correct answer is the one that would best explain the conclusion. The course had little material on 'inference to the best explanation' arguments (see *5.5.3 'Weaknesses of the Textbook'*). Pretty clearly there should be some.
- LSAT questions sometimes raises issues about causality. For example, they will ask which claim if true would most undermine some view. The view turns out to involve some causal element. And the claim that if true would most undermine the view turns out to be, e.g., a claim that would imply that the effect occurred before the cause, or that a factor similar to the alleged cause did not have a similar effect elsewhere, etc. In the course, there was little if any discussion of causality and how claims about it might be supported or undermined. This may well have been a mistake.
- LSAT questions sometimes ask what the 'main point' of a passage is, where the main point isn't the final conclusion or indeed any particular claim found in the passage. If you mapped the passage, the main point would be a "meta-map" point, some point about the map rather than something found in the map. For example, the main point may be that you don't need to deny some premise in order to deny the conclusion. So construed, there may be many possible 'main points' that someone could theoretically come up with but only one of the multiple choice answers could possibly be right. In the textbook, there was little if any discussion of how you might, e.g., 'look at the map and sum up its key points in a sentence or two'. There aren't any quiz questions of this type.
- In a related 'meta-map' vein, some LSAT questions present a text with some reasoning and then ask which of the possible answers exhibits or most closely parallels the 'pattern' of reasoning in the text. The patterns are not restricted to deductive schemas like *modus ponens*, etc. The pattern might be something like "p because q. But r. So not-p." Perhaps the patterns sometimes could be seen by looking at maps constructed both for the text in the passage and for the possible answers. This possibility should be explored far more than it has been.

In sum, the LSAT is an excellent test with a large literacy and culture weighting. For law school admissions purposes, this is as it should be. But, this weighting is measured in subjects' pre-test scores, and the student's literacy and cultural background will change very little during this period. Therefore, Getting a robust 1+ SD on the LSAT would require substantial change across a number of factors (literacy, cultural background, etc.), not just a change in critical thinking. Critical thinking courses do not deal with those other factors and should not, because of the time it would take. We would recommend seriously considering the LSAT for measuring the critical thinking skills of analysts, insofar as a good analyst should be highly literate. Not in the sense of reading lots of novels, but in the sense of wide read and knowledgeable. But, we would not recommend the LSAT for measuring *changes* in critical thinking ability.

5.4 CCTST and HCTA

There difference between the CCTST results and the HCTA lies well within measurement uncertainty. This is not surprising since both tests aim at the same relatively low level (roughly high school graduates) of literacy and background knowledge.

Still, there is a difference worth noting both for its theoretical and practical value: "Critical thinking" and "critical thinker" are polysemous, not only with the general public but among the experts in teaching critical thinking. These differences in understanding are reflected in standardized critical thinking tests. For example, the CCTST has test items that use imaginary terms to examine the subjects' capacity to determine logically necessary conclusions, along these lines:

> "All blonks are tonks, and all blonks eat squanks. Further, no non-squank is eaten by a blonk tonk." Assuming that those two sentences are true, which of the following must be true? Possibly more than one answer is correct, so answer as many as you think are correct:
>
> A) All blonk tonks eat squanks.
> B) Blonk tonks only eat squanks.
> C) No squanks are eaten by tonks that are not blonks.
> D) Some squanks are eaten by non-blonk non-tonks.
> E) Some non-blonk tonks eat non-squanks.
> F) None of the above.
> G) Who gives a damn?

Since the project did not cover Venn Diagrams, there is little reason to believe that the subjects would do better on such questions. Does this show a weakness in the course? Or is it a weakness in this aspect of the CCTST for the purposes? The answer to those questions depends both on the account of critical thinking one accepts and psychological questions about the capacity and advisability of people to think in terms of logical form (formal or informal).

Unlike the CCTST, the HCTA focuses on practical questions, the sort that might come up in subjects' day-to-day life, as opposed to many the more abstract questions in the CCTST. This approach may provide a lesson for future critical thinking research, whether or not argument-mapping based. Possibly, the project focused too much on more abstract questions and thoughts, and not enough on ones that students are likely to think about outside of a critical thinking class. Perhaps, there should have been more emphasis on more practical, less academic, questions; this might enhance transfer to outside the classroom.

There is another aspect to the HCTA that should be mentioned. Several instructors expressed concern with it, because some students' performance on the test appeared to be uncorrelated with the instructors' impressions of their thinking skills. Several disagreed with some of the "correct" answers; this did not happen with the LSAT. Others have looked closely at the difference between the A and B versions and have found that they more different than at first appears. However, since we used AB/BA design, any such differences should not have affected the average effect size on the HCTA, although it might have increased the SD and thus decreased the standardized effect size.

1. Motivational differences in taking pre and post tests

The above proposed explanations for the considerable variation in results assumes that the pre- and post-tests measure, at least roughly, the cognitive ability of the subjects before and after the course. This would be true even if, as proposed, the LSAT also has a heavy loading of factors such as reading speed. But, it was not obvious at actual testing that all subjects always tried their hardest on both pre and post tests. In fact, in some test conditions, a fair number left the test room

remarkably early, to the disquiet of those running the experiment. Obviously, this problem is a widespread psychometric problem in education.

A recent study (Liu, Bridgeman, and Adler 2012) studied the learning gain of students, measuring motivational conditions. They found that depending upon the test format and conditions, learning gain varied from substantial gain (d = 0.72) to negative gain (d = –0.23). One explanation may lie in the differences in motivation of the subjects. Those who are students may have had more motivation than those who are professionals, and motivation might vary based on the perceived benefits of completing a course in CTAM for their career or course credit.

5.5 The Textbook

5.5.1 Strengths

5.5.1.1 Style

Writing

The textbook is written in a clear, engaging style. It discusses the issues in as much detail as is necessary, and doesn't talk down to or lecture the student. There is some occasional humor (Harris cartoons, absurdly obscure passages as examples of jargon and buzzwords, etc.).

Most who have read the textbook have commented on its clarity, and the way that it gets at the central issues without any superfluous discussion, or unnecessary asides. It focuses on the key skills students need to learn, and is accessible and approachable for beginning students in critical thinking.

Formatting

The formatting is elegant and attractive, and makes the text easy to read. All quoted text is in green, which makes it stand out more from surrounding text. Problems with maps or arguments are indicated with a clear yellow sticky note with a big red 'X'. Thus there is no danger of students reading a 'mistaken' map or argument and believing it to be correct.

5.5.1.2 Structure

Organization of material

The textbook is well organized. Each section builds on the previous section in a logical way. The textbook begins by covering the key elements of arguments (reasons, conclusions, objections, premises and claims), with all examples being 'fully-stated' (i.e., arguments that have no unstated premises). After this the textbook covers fully-stated arguments with non-argumentative material, then introduces incompletely stated arguments (arguments with unstated premises or conclusions) along with the concepts of evidential relevance, danglers, and the 'Rabbit Rule'. It then looks at refining claims, basis boxes and basic premises, and finally covers evaluation (bringing together several important topics covered earlier), and how to convert maps to prose. This structure matches the structure of the course the textbook was designed for, and represents, in our view, the best way of organizing the material for beginning students.

Length

The textbook, at 129 pages, is clearly considerably shorter than standard critical thinking and informal logic textbooks, partly because it doesn't cover many of the standard topics in critical thinking and partly because it doesn't include the many exercises. Its brevity, in focusing only on the essentials of critical thinking, is a strength: it makes it more accessible for the beginning student, and it increases the likelihood that the student will read all of it, as well as making it relatively cheap for students to print.

5.5.1.3 Content

Examples

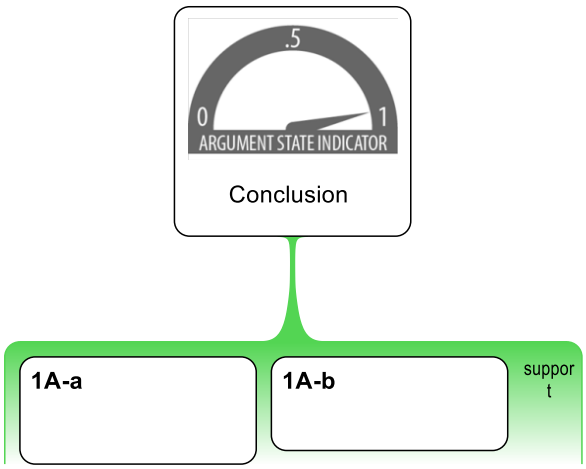
The examples in the textbook are for the most part clear and easy to understand. They convey the relevant points successfully.

Justification for mapping as an approach to critical thinking

In the first chapter the general reason for argument mapping is spelled out clearly, with a helpful example of the superiority of maps to information presented in prose.

Argument state indicator graphic

The 'argument state indicator' graphic, which vividly shows the way in which reasons and objections of varying strengths ought to influence one's confidence in the conclusion of an argument, is a helpful visual device, and is used in several chapters. For example:



*No such thing as '**The** Argument'*

The textbook makes the important observation that generally there are several conflicting yet acceptable interpretations of the argument in question, so there is often no such thing as '*the* argument' or '*the* correct map of the argument'. This point, so often neglected in standard critical thinking textbooks, with their focus on baby arguments, is crucial once one starts dealing with real argumentation.

5.5.1.4 Improvements on typical critical thinking courses

Topics

The textbook covers only the fundamental topics of critical thinking, centered around argument mapping, that we think can help students become better reasoners. Thus we do not cover (or only very briefly cover) many topics typically covered in standard critical thinking courses, such as deduction vs induction, fallacies, statistical reasoning, scientific reasoning, validity and soundness, and other aspects of formal logic. There is

evidence that students do not retain much of this material when they take critical thinking courses, as shown by the small effect size such courses have. But the same sort of evidence strongly indicates that argument-mapping centered critical thinking courses transfer far more successfully to interpreting and evaluating real life arguments.

Terminology

The aim of the textbook was to keep logical terminology to a minimum, thus we avoided terms like 'valid', 'sound', 'proposition', 'conditional', 'deductive' and 'inductive'. For the most part this was an improvement, although see Weaknesses section below ('terminology' subsection) for some reservations about this strategy.

5.5.2 Improvements on previous Argument Mapping courses/textbooks

The textbook is an improvement on previous argument-mapping-oriented textbooks in several respects:

- All the examples use the *Advanced Reasoning* mode in *Rationale*. This makes for consistency and clarity, and makes sense given this is the mode the students will be using in the course.
- The textbook uses less logical and critical thinking terminology than previous argument mapping oriented textbooks. Unnecessary terms such as 'lemma', 'multi-branch argument', 'multi-layer argument', 'well-formed reasons', 'cases' and 'proposition' are not used.
- The textbook focuses on 'cheap *co-premises*' rather than 'cheap *conditionals*', since cheap co-premises need not be conditionals. For example '[*Premise*] *shows that* [*conclusion*]', or '[*Premise*] *means that* [*conclusion*]' – these are cheap co-premises, but are not conditionals.
- Previous argument mapping oriented textbooks had included complex lists, flow charts and procedures for constructing, interpreting and evaluating arguments, which we felt were unhelpful and confusing. The textbook avoided these.
- It was felt that the 'holding hands' rule was not helpful to students, so we left it out, and introduced only one 'rule' for constructing arguments, the 'rabbit rule',

in the context of a more general discussion of 'danglers' in arguments (concepts that only appear once in an argument unit).

5.5.3 Weaknesses of textbook

5.5.3.1 Conceptual issues

Rebuttals

The category of 'rebuttal' is an unnecessary and confusing category in the context of argument mapping, and is not consistent in the way in operates in *Rationale*, in that an objection to a rebuttal is not called a 'rebuttal', but an 'objection', despite the fact that it plays the same role as a rebuttal. Removing rebuttals from both the software and the textbook would be a good idea.

Evidential relevance

There is a tension in the treatment of evidential relevance, reflecting a tension in the concept itself. On the one hand a reason is said to be relevant to a conclusion when that reason makes the conclusion more likely to be true than not. On the other hand a reason is said to be relevant to a conclusion when it raises the probability of the conclusion, which it may do even though it doesn't make the conclusion more likely to be true than not. Whether this ambiguity is something that can be avoided, and whether it would be a problem for students, is not clear.

5.5.3.2 Terminology

Logical terminology

The textbook in its current form eschews much traditional terminology used in logic and critical thinking textbooks and courses. The reasoning behind this was that such terminology, and the concepts and distinctions associated with it, while helpful for philosophers, do not help students to become better reasoners. However it is possible that the textbook goes too far in this direction. The experience of those teaching the argument mapping course was that including some traditional logical terminology ('modus ponens', 'deductive', etc.) could be helpful for both students and teachers.

Problems with 'relevance'

We chose to use the term 'evidential relevance' for the version of (probabilistic) validity, given that we wanted to avoid traditional logical vocabulary like 'valid' in its truth-preservation sense. The textbook makes clear that this sense of *evidential* relevance is different from the standard meaning of *topical* relevance, but it seems likely that students will still confuse the two senses of 'relevance', and thus an alternative term may be needed.

5.5.3.3 Improvements to existing chapters

Evaluation chapter

The chapter on evaluating arguments is weak in several respects.

The point about both evidential relevance and acceptability of premises being matters of degree is an important one, but is discussed too briefly. The same goes for the point that the amount of support a reason provides for a conclusion is determined by the degree to which the premises are relevant to the conclusion, combined with the degree to which the premises are acceptable. These points should be fleshed out and explained in more detail, and some examples provided to illustrate them.

The 'evaluating complex arguments' section is too brief and sketchy. This section should bring together many of the points made earlier in the chapter, and in other chapters, to sum up the fundamental skill of evaluating real life arguments with a number of levels and lines of reasoning. At present this section doesn't do this.

Refining claims chapter

The Refining Claims chapter had some weaknesses. One weakness was the discussion of vagueness. Vagueness is clearly a problem for many claims and arguments, but much of the time a degree of vagueness is unavoidable and unproblematic, and saying exactly when vagueness is a problem, and when it isn't, can be difficult. The section on vagueness does not solve this problem very satisfactorily ('The answer to the question, 'when does vagueness matter?' is 'when it matters for something else'') and needs a

rethink. Some more examples of when vagueness is genuinely a problem, and when and how vague claims need to be refined, would be helpful.

5.5.3.4 Possible Additions

Fallacies

At present there is minimal discussion in the textbook about fallacies of reasoning (the fallacy of equivocation is discussed briefly), which form a major part of most critical thinking courses and textbooks. The question of whether and how fallacies should be taught as part of critical thinking courses is a controversial issue, but having at least a section of the textbook dedicated to common fallacies would be a good idea. To fit in with the approach of the project the treatment of fallacies would need to be based around argument mapping. The natural way to do this would be to indicate the ways in which mapping fallacious arguments can bring out precisely what is wrong with them, typically that they have false unstated premises. This can then be used to further support the value of the argument mapping approach to critical thinking.

Argument forms

It was felt by some who taught the course that the textbook would profit from having a section on common argument forms, possibly as an appendix. Presently argument forms do not appear in the textbook, as it was felt that their value was limited, and encouraged an overly formal approach to critical thinking. But teachers have found that students, when mapping, find it useful to have a set of argument forms to draw on. These would of course need to be presented in the form of argument maps, and would include both deductive and inductive, and possibly explanatory, forms, although to be consistent with the approach of the project such distinctions (between for example inductive and deductive arguments) would not be emphasized.

Critical thinking course vs. argument mapping course

The guiding assumption of the project is that argument mapping is the best way of teaching critical thinking in all, or most, of its aspects. Yet argument mapping is sometimes treated as one component of critical thinking among many. This leads to the

mistaken view that the course that we have designed focuses on just one aspect of critical thinking – argument mapping – at the expense of all the others. In fact we see argument mapping not as one 'topic' within critical thinking, but as a tool with which to approach many topics in critical thinking. Making clear this point – that this is not so much an 'argument mapping course' as it is a critical thinking course which utilizes argument mapping – at the beginning of the textbook would be helpful.

Rhetorical questions

Because they feature prominently in real life arguments, it would be useful to have a section on rhetorical questions, explaining their functions and how to deal with them when interpreting and mapping arguments.

Inference to the Best Explanation

There is very little in the textbook about inference-to-the-best-explanation arguments, yet this is a widespread form of reasoning. A section dealing with explanatory arguments would be a useful addition. It could go over the best ways to map such arguments, including identifying their unstated premises.

5.5.3.5 Integrating textbook with the course

How to use the software

The *Rationale* software is central to the argument mapping course the textbook is designed for, yet there are currently minimal instructions on using the software in the textbook. Since mastering the software is a prerequisite for undertaking the course, it is likely that including more detailed instructions, across several chapters, in the textbook, would be helpful to students, and would more closely integrate the textbook with the software and the course. An alternative to this may be to produce short instructional videos covering the basics of using the software, which could be put on the course website. Such instructional videos are commonly used and their form is familiar to students. They may be better for introducing the software than written instructions in the textbook.

Textbook should be printed

There was evidence that students taking the argument mapping course were not making use of the textbook to the extent that we had hoped, or planned for. Probably, this was partly because the textbook was only available online through the course website - hard copies were not made available to students. In future hard copies as well as online versions should be available, and students encouraged to regularly refer back to the textbook throughout the course.

5.6 Lots of Argument Mapping Practice (LAMP)

It is safe to say that we used far more LAMP activities than other mapping courses. JMU, USNA, and Melbourne used the LAMP material for homework activities. Despite our best intentions, subjects did not get as much quality practice and feedback as we had hoped. Often the small-group and the class-wide discussions were animated and illuminating; all too often, they were not.

LAMP improvements needed:

1) Model answers would have helped teachers considerably. Mapping is a challenge, and it is not easy to figure out an answer while discussing a student's work.
2) The questions should be more focused on the day's particular topic, and the take-home lesson for each day's activity made clearer. We found it more difficult than anticipated to design questions that are seen by students as focusing on the topic of the day, despite the detailed structure of the textbook and the entire course.
3) We tried to make the arguments interesting; there is some disagreement about how far we succeeded. Perhaps the truth is that the questions are more interesting to more students than the questions found in most critical thinking courses, and that with additional ingenuity and time we could have done better. Certainly a lot of effort was spent on this.
4) In theory, a student was getting feedback from her peers at her computer terminal as they collectively constructed their map, from the teacher going from terminal to terminal, and from the group as a whole where the map was shown on the screen for group discussion. Also, seeing the strengths and errors of other groups' maps could illuminate. That is the theory, and it still sounds good. In fact, though, the feedback

was not as regular or as useful as we had intended. We do not understand why the feedback technique did not work better.

5.7 Mastery Learning Milestones (MLM)

Writing, revising, and revising again the Mastery Learning Milestones (MLMs) was a massive task. Over 300 questions with answers were written, of which about 200 were used in the final course, with quizzes covering the first seven topics (modules). They were on the improvingreasoning.com website, and students worked from the website. However, at some venues, for security reasons internet access was not an option, so PDFs were used. We found that online access is clearly preferable.

In early trials, we found it impossible to write questions that directly measured mapping mastery. So, we added questions that required you to map an argument in order to answer the question, by filling in the on-line answer sheet.

MLM improvements needed

1) As befits classical mastery learning theory and practice, the students could redo each MLM quiz as often as required to pass at the set mastery learning level. Depending on the test, we set the pass mark at 80 or 90%. This approach is standard in the mastery learning research literature which developed when there were written tests and, as a practical matter, a student could only do one test per day. But the technology, and the students, have changed.

   To enhance learning, we gave immediate feedback on each question, as well as telling the student how they were doing overall on the quiz. Students soon realized that, after answering one or two questions incorrectly, it was not possible for them to pass the module. So, many of them simply stopped that test and immediately started retaking the test. They did not do as we had intended – take the MLM quiz through to the end to discover what they needed to know. Nor did they go back to the textbook or the LAMP material to think about the topic. Essentially, these students used trial-and-error to rush through the quizzes. For these students, there was no incentive to think about the questions and so

they did not get the required mastery. Which, if true, would help explain why the results were not as strong across the board as we had hoped.

Here are two ways to deal with this problem:

- The Khan Academy approach, where the student is required to answer 10 consecutive questions correctly. This way, the student has no incentive to stop the particular test just because incorrect answers were given.
- The test could not give the answers until the entire quiz has been taken. This may have the advantage of getting students to actually think about the results when they are handed out at the end of the test. Instant feedback is useful only if it is used.

We do not know which approach is educationally superior, but suspect that contemporary students are so used to instant feedback that the second approach might irritate them. So, we suspect that the Khan Academy approach is preferable.

2) More questions are needed, or so at least some students said. These usually were students who took the MLMs many times, perhaps by not completing the entire MLM once they found they could not pass it. Possibly going to the Khan Academy's approach of 10 consecutive correct answers will solve this problem, since such students see the first few questions often, even though the questions were shuffled after each test taken. Or, it could aggravate it, since it might be harder for some of them to get ten correct in a row. In any case, at least some of us do not have a lot of sympathy for this complaint, since we believe it emerged from these students' not looking at the textbook to figure out what they hadn't understood.
3) The MLM interface needs improving. It could be made more attractive and, more importantly, in some ways it was annoying to use. Questions with longer targets and/or lengthy answers sometimes could not properly display on the page. Many images could not be displayed easily. These difficulties emerge from the amount of material that needs to be on the screen simultaneously – map, target and choice of answers. Once one has gone beyond baby examples, the maps have several reasons and objections, the target has a fair number of sentences,

and often the alternative answers are rather lengthy. Having the map expand when the cursor slid over it was a considerable improvement, but more is needed. There is not an easy solution to this problem.

4) More focused questions are needed for each topic within a module. Because we did not have a large enough question bank, we could not use the current MLMs to diagnose problems. There are no theoretical or other obstacles here; it is simply that making good MLMs takes considerable time, particularly since what may seem *prima facie* as an excellent question all too often proves to be open to misinterpretation or have some unsuspected property that we had not detected.
5) Although the quality of the questions and answers was remarkably good given the constraints on their construction, there is room for improvement. Vague questions and incorrect answers annoyed the students, and reasonably enough.

5.7.1 Possibility of 100% automated MLM assessment and MOOCs

Marking students' maps is a time consuming, and often depressing, activity. It would be far better both for the teachers' morale and for widespread uptake of such a course if automated assessment were an option. However, it is unlikely that 100% automated assessment in a course like this is feasible in the foreseeable future. By 'a course like this' we mean a course that involves assessment of the maps that students construct to represent some text. This may be a good thing, because automated assessment might not be educationally preferable to the alternative of having students intelligently mark each other's work.

Certainly a good deal of assessment, particularly of the more boring bits, can be done in an automated way. You can even do quite sophisticated assessment involving maps, e.g., presenting students with a complex map and asking them multiple choice questions. Or, if desirable, you can present a text with several non-trivially different correct multiple-choice answers, to reflect different legitimate interpretations of the text.

The first difficulty with automatic assessment of students' maps is that computers do not understand natural language. There are many different ways that a student can reasonably represent the same thought, however narrowly "the same thought" is taken. For example, "All red-haired men are bad tempered" should be taken as the same as "All

men who have red hair are bad tempered" and as "If a man has red hair, he is bad tempered." This is a particularly simple case. It will be a long time before computers can determine whether a student's answer is close enough to the instructor's answer.

Further, as the textbook says, generally more than one map can reasonably represent a text. This is true even when you ignore trivial variants, such as different orderings of a reason's premises. The rule of thumb "There is no one right map" comes from a deep fact about informal logic: excluding baby arguments, there is no one right reconstruction of an argument. Some instructors think the textbook should emphasize this even more than it does. In any case, it is important theme in the classroom.

The non-existence of "the right answer" is true even for very simple arguments. Consider Scriven's example, an argument about as simple as one can find: "I know John has a bad temper – he has bright red hair." There are many plausible unstated candidate premises here:

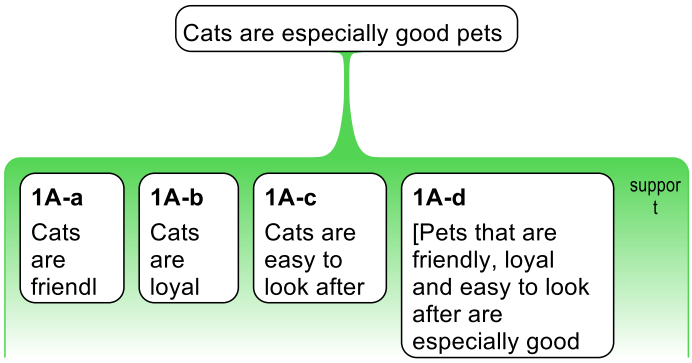
- All red-headed people are bad tempered.
- All red-headed men are bad tempered.
- Almost all red-headed people are bad tempered.
- Almost all red-headed men are bad tempered.
- All red-headed Caucasians are bad tempered.
- …
- All red-headed people around me are bad-tempered.
- …

In a third person case (when reading a text, for example) there generally is no knowable answer to "What is the correct unstated premise?" First-person reflection indicates that often there is no truth to the matter at all.

The multiplicity of reasonable maps for a given text isn't only due to the problem of many plausible unstated premises. Even where there is *agreement* on all the premises, stated and unstated, some may reasonably think one is so obvious that it does not need to be put into the map, others may reasonably slot it into the map.

In addition, real-life arguments often have so many gaps and unclarities that there may often be more than one respectable way to represent them. Turning now to fully-stated arguments where the premises themselves are unproblematic, consider "Cats are friendly and loyal and, moreover, they are easy to look after, so they make especially good pets." Even something as simple as that has a couple of acceptable maps, because there are a couple of reasonable interpretations of it. These are maps with different structures, not just maps with different content for the premises and conclusions in a map that has a given structure.

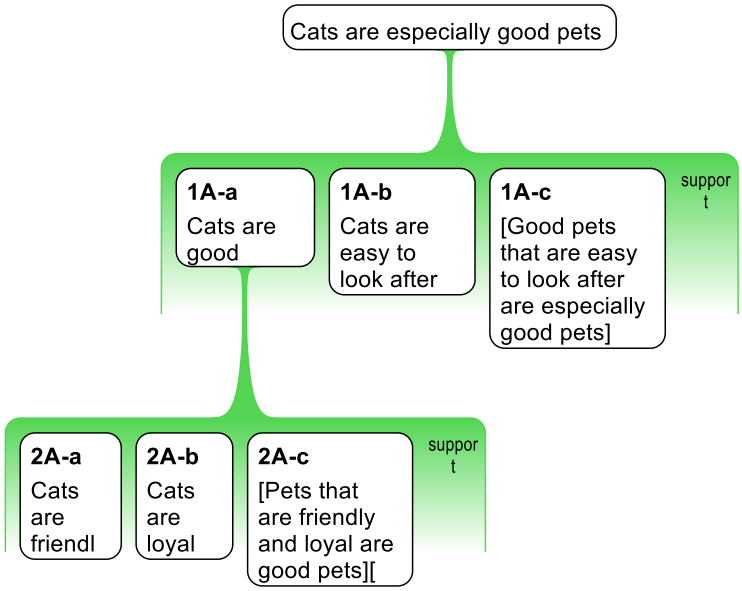
Here is one legitimate way to represent that text:



There are other possible maps even for such a simple argument. For example, the person who made this map regards "Cats are pets" as an unstated premise that is too obvious to include in the map but other people may include it in the map. That is a fairly minor disagreement. It isn't even a disagreement about what the unstated premise is, it is only a disagreement about whether it needs to be in the map.

However there can be more significant disagreements about even this simple argument, involving different ways of interpreting just what the argument amounts to. For example, some people would be attentive to those bits of the text "moreover" and "especially" and think that it is significant that those bits are there. They might interpret the argument as suggesting, first, that cats being friendly and loyal makes them good

pets and, second (i.e., moreover), that cats being easy to look after makes them not just good pets but especially good pets. Here is a suitable map:

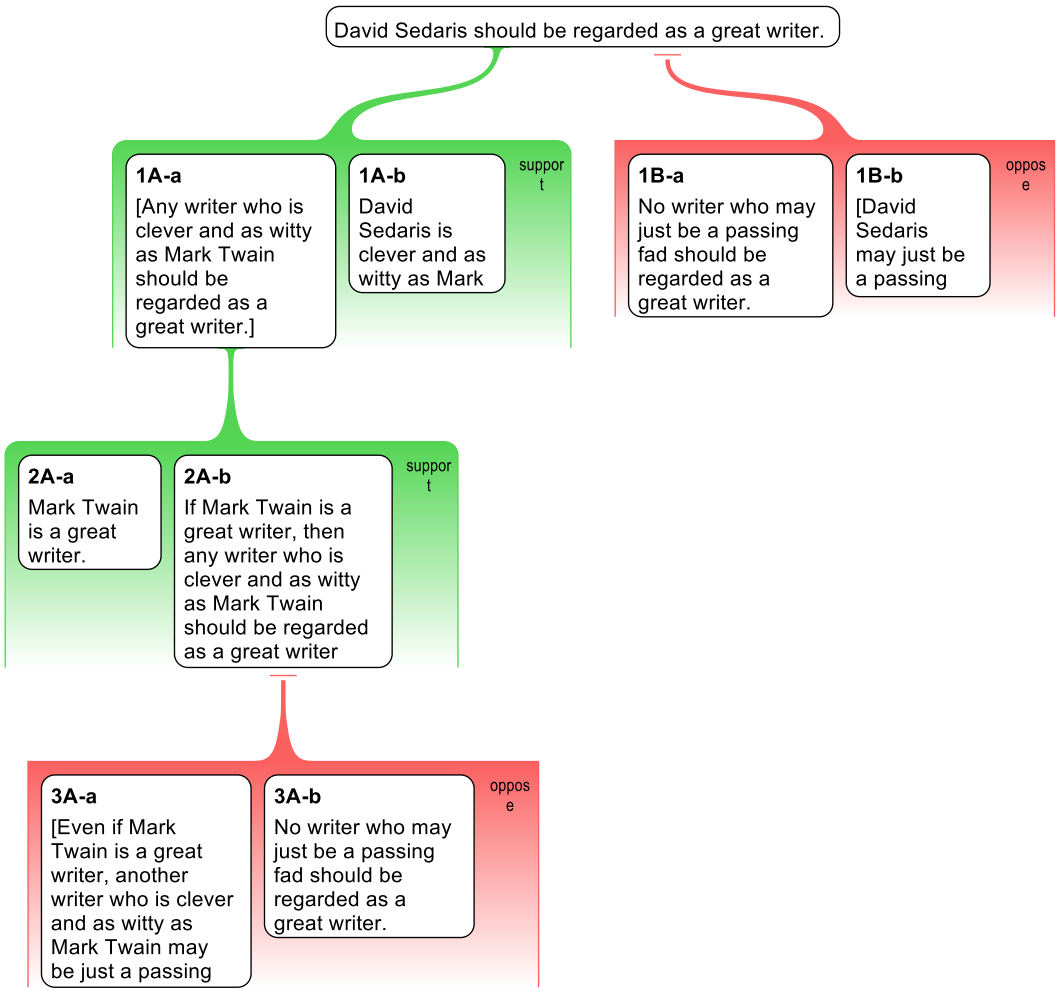


As simple as the text is, perhaps you can even imagine even other possibilities for the map representing it. But the main point has been already made, viz., that there can be alternative reasonable maps for even a very simple text. If this issue can arise with very simple arguments, then imagine the range of possibilities with arguments of any complexity. Of course this does not mean that anything goes. There are lots of defective maps, most of them obviously defective. But the fact that there are lots of wrong maps does not mean that there is a single right or best map.

Often, even when people agree an objection needs to be in a map, they reasonably disagree about what it is an objection to. Here is an actual disagreement amongst the people working on the textbook about where an objection should be placed:

> "There's no doubt that Mark Twain is a great writer and surely if he is a great writer, then any writer who is clever and as witty as he is should be regarded as a great writer too. Anyone who reads David Sedaris can see that he is clever and as witty as Mark Twain, so he should be regarded as a great writer too. Of course, someone will object that no writer who may just be a passing fad should be regarded as a great writer."

Here is the map preferred by mapper #1:



In this map, 1B-a and 3A-b are the same. Mapper #1 thinks that this is fine, interpreting the text as implicitly involving both objections 1B and 3A. However mappers #2 & #3 think that reads too much into the text and they each have a different view on how to map the text.

Mapper #2 would delete objection 3A from the map and retain 1B only. However, if you only have objection 1B and then someone successfully undermines the claim 1B-b that David Sedaris may just be a passing fad, then the chain of reasoning on the left hand side would stand unchallenged since you've deleted 3A from the map. But clearly the objection 3A to the general claim 2A-b could be made and could stand even if someone can undermine the claim that David Sedaris in particular may just be a passing fad.

Mappers #1 & #3 interpret the text as implying some objection to the general claim 2A-b and think that 3A needs to be there.

Mapper #3 would delete objection 1B from the map and leave 3A only. However, that leaves it open whether David Sedaris in particular may just be a passing fad. Mappers #1 & #2 think that the text strongly suggests that David Sedaris in particular may just be a passing fad and that objection 1B needs to be there. And Mapper #4 thinks it should go under 1A-a.

So which map is the 'right' or the 'best' map – the one with both 1B & 3A, the one with only 1B, or the one with only 3A? While the three mappers each have a preference for one over the others, it is hard to see that any one of those maps is clearly right or best and that the other two are clearly wrong or inferior. They all seem to be respectable interpretations of the text.

That there are often many possible reasonable maps representing any given text of any complexity seems both true and important, and agreed by everyone working on the project. How could you construct an automated system that will have in its bank all the reasonable maps that students might possibly construct from a given text? Maybe someday somehow it will be possible to do that but it is clearly beyond the capacity of a project like this. If possible at all, it would be a massive and complex programming task, perhaps akin to programming a computer to mark essays. Or perhaps even worse, if you had to anticipate all the different permutations; truly a combinatorial explosion.

If true, this has implications for developing the course as a MOOC (Massive Open Online Course). Some MOOCs now leave room for some human assessment by the students themselves. In fact, if rumors are correct, considerable progress has been made in enabling students to reliably mark each other's short papers in some humanities-oriented MOOCs. We recommend a systematic review of this literature, with a focus on whether student marking of one another's maps can be made demonstrably reliable and relatively rapid.

6 Acknowledgements:

We would like to acknowledge and thank all the participants in the courses/experiments for their participation and forbearance.

When Larry Lengbeyer taught his non-credit trial run at USNA, the textbook, MLMs, etc., were in early drafts; Lengbeyer's detailed feedback from Annapolis considerably improved all of them.

Ashley Barnett's was tasked with integrating mastery learning into an argument mapping course. He did this and much more, becoming the lead writer and designer of the learning materials, developed the mastery learning quizzes and the argument mapping practice questions, and wrote over eighty per cent of the questions used in the final course. His efforts and achievement were heroic, far beyond the call of duty.

Richard Lempert, although not a participant in the implementation of the experimental courses, was deeply involved in the planning process for the experimental courses and elements of the research design. He also made extensive comments on the first draft of this final report for which we are very grateful.

Charles Twardy's participation in the Berkeley Workshop and comments on earlier drafts of this report substantially improved both.

The University of Melbourne's School of Historical and Philosophical Studies gave major support and advance funding. In particular we thank Professor Trevor Burnard, Head of School, and Josie Winther, School Manager, for their support and encouragement. Without Merc Fox's diligence and understanding of the intricacies IRBs, we would still be mired in getting Ethics Approval. Geoff Cumming's and Steve Kambouris' understanding of statistical and administrative matters has made our life much easier.

Without the unstinting efforts of Melinda Heron, this project would have died an early, painful death.

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via SPAWAR under Contract Number N66001-12-C-2004. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors, and do not necessarily reflect the views, either expressed or implied, of IARPA, SPAWAR, or the U.S. Government.

7 Appendix: van Gelder: Meta-analysis "Impact of argument mapping on critical thinking skills"

7.1 Abstract

How can we improve critical thinking skills? One pedagogical technique receiving increasing attention is argument mapping. This paper presents a small meta-analytic review of studies of the impact of argument mapping on critical thinking skills. The analysis suggests that argument mapping-based instruction typically is much more effective than standard undergraduate education, and substantially more effective than other forms of critical thinking instruction. However it is no magic wand. The extent to which it improves critical thinking depends strongly on the extent and quality of the argument mapping instruction; the more intensive the instruction, the greater the gains.

7.2 Introduction

The centrality of critical thinking (CT) as a goal of higher education is uncontroversial. In a recent high-profile book, *Academically Adrift*, Arum and Roska report that "99 percent of college faculty say that developing students' ability to think critically is a "very important" or "essential" goal of undergraduate education" (Arum and Roksa 2011, 35), citing (HERI 2009).

However a major message of their work is that college education generally makes little progress towards this goal: "many students are only minimally improving their skills in critical thinking, complex reasoning, and writing during their journeys through higher education." Indeed for many students college education appears to be failing completely in this regard: "With a large sample of more than 2,300 students, we observe no statistically significant gains in critical thinking, complex reasoning, and writing skills for at least 45 percent of the students in our study …".

Their message is barely more positive than H.L. Mencken's acerbic comment, over a century ago: "Certainly everyday observation shows that the average college course produces no visible augmentation in the intellectual equipment and capacity of the student. Not long ago, in fact, an actual demonstration in Pennsylvania demonstrated

that students often regress so much during their four years that the average senior is less intelligent, by all known tests, than the average freshman." (Mencken 1997, 98).

Yet we also know that college education *can* positively impact CT; simply put, critical thinking (CT) can be taught. In a meta-analysis of 117 studies of college-level efforts to teach critical thinking, Abrami *et al*. found "a generally positive effect of instruction on students' CT skills." (Abrami 2008, 1119). However the amount of gain found in these studies varied widely, and Abrami et al concluded that it makes quite a difference *how* CT is taught. They say "both the type of CT intervention and the pedagogical grounding of the CT intervention contributed significantly and substantially to explaining variability in CT outcomes." (p.1120).

Thus, in the long term achieving one of higher education's most central goals will involve meeting two challenges:

1. Identifying and refining those instructional practices that are most conducive to CT skill gains; and
2. Overcoming the numerous practical barriers to the widespread adoption of those practices.

This paper is a modest step with regard to the former challenge. It focuses on one promising instructional practice, argument mapping (AM), and attempts to quantify the extent to which it in fact improves CT skills, drawing on all known studies conducted to date.

7.3 Argument Mapping (AM)

Argument Mapping (AM), also known as argument diagramming or argument visualisation, is visually depicting the structure of reasoning or argumentation (Davies 2011, Macagno, Reed, and Walton 2007, van Gelder 2013). Typically an argument map is a graph-type or "box and arrow" diagram, with nodes corresponding to propositions and links to inferential relationships.

AM's roots reach back into the nineteenth century, but it has only become popular in the last decade or two, primarily as a tool to help students build reasoning and CT skills.

Indeed its immediate precursor was the kind of argument diagramming found in many introductory logic textbooks (see, e.g., Fisher 1988, Govier 1988). The most common type of exercise involves providing an argumentative text and requiring the student to map out the argument it contains, i.e., to produce an argument diagram faithfully representing the reasoning. This can be surprisingly difficult. Other AM exercise types are (1) creating a map to represent an argument of one's own creation; and (2) translating an argument map into fluid argumentative prose.

A key factor in the recent growth in popularity of AM has been the development of software tools designed specifically to support this activity. Previously argument diagrams would have to be created "by hand" (whether on paper, or on a computer using generic drawing software), which made producing maps of any complexity tedious and time-consuming. New software packages eliminate much of the "futzing around" with boxes and arrows, as well as providing varying amounts of guidance, scaffolding and inbuilt exercises.

7.4 Previous Results

AM's adoption as an instructional practice has generally been as part of an attempt to make instruction in reasoning or CT more effective. On many such occasions there has naturally been some interest in whether the new approach is working as intended. Thus a number of studies have attempted to quantify the extent to which students' skills are developing, making the plausible assumption that any above-normal gains could be at least largely attributed to the AM-based instruction.

Some of these studies were included by Alvarez in an important review of studies of CT gains over one semester in undergraduate subjects (Alvarez 2007). Her main goal was to empirically evaluate the claim, made so often by philosophers and philosophy departments, that philosophy instruction is particularly effective at enhancing CT skills. Her meta-analytic review aimed to pool all available pre- and post-test studies of CT gains in philosophy subjects, and to compare the results with the expected gains in other relevant contexts, such as other CT subjects, or other university subjects (i.e., simply being at college).

Alvarez' review drew upon prior work by The Reason Project at the University of Melbourne. Starting in the late 1990s, that project had been developing a new approach to teaching CT in a one-semester, dedicated CT subject. The new approach largely ditched the traditional textbook- and lecture-heavy pedagogy in favor of extensive deliberate practice of CT skills (van Gelder, Bissett, and Cumming 2004, Ericsson, Krampe, and Tesche-Römer 1993). The form of practice used was software-supported AM; hence the approach was nick-named "LAMP" for "lots of AM practice" (Rider and Thomason 2008).

A series of studies using pre- and post-testing with objective tests indicated that students undergoing this instruction reliably showed gains in CT skills. However the significance of these gains could only be assessed by comparing with the gains which would have occurred anyway, whether due to simply growing up (maturation), being at university, or due to the mere fact of undergoing some kind of CT instruction. As it happened, at the time, good information on those "would-have-happened-anyway" gains was not available – hence the need for a broad meta-analytic review.

Alvarez' results are summarized in Figure 1:

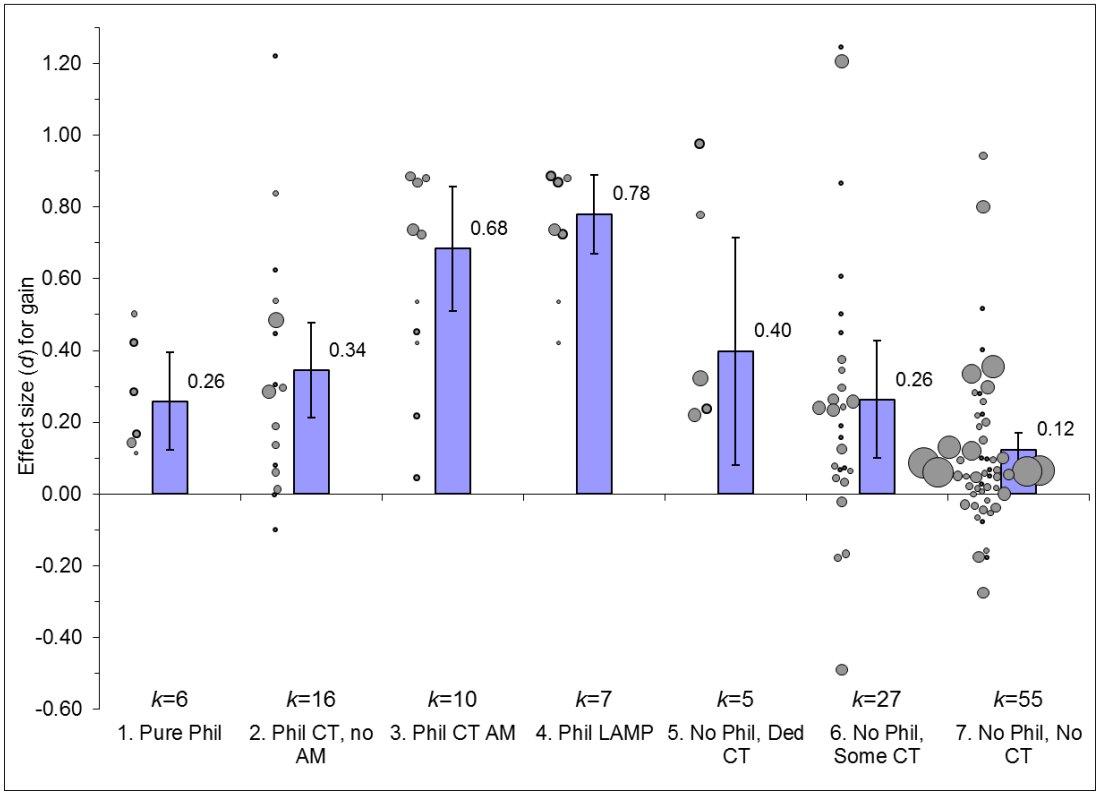


Figure 1: Critical thinking skill gains over one semester, as reported in (Alvarez 2007).

Each circle in Figure 1 corresponds to a study of CT gains over a semester of undergraduate education, with the size of the circle representing the number of participants, and its vertical position representing the effect size found in that study. The unit for effect sizes is Cohen's *d*, a widely-used measure of the strength of an effect found in a study. Its purpose is to serve as a kind of common yardstick across studies which might vary in important respects. For example, studies of CT use a variety of different tests. Some tests typically give much larger mean gains than others, as measured in raw points. Thus, comparing mean gains using just the scores on those tests can be very misleading. Cohen's *d* tries to overcome this problem by relating mean gains to the typical variability in test scores. In Alvarez analysis, and in the one below, Cohen's *d* is calculated as mean post-test score minus mean pre-test score, divided by the standard deviation for the test used in the study. Cohen's *d* is thus a number of standard deviations. The standard deviation is, wherever possible, the one given by the developers of the test as the norm for the relevant population. It is common, although

arbitrary, to regard *d* values of 0.2, 0.5, and 0.8 as, respectively, small, medium, and large, a rule of thumb suggested by Cohen himself (Cohen 1969).

Each column in Figure 1 corresponds to a category of studies, based on the type of undergraduate subject in which the study was carried out. The categories are:

1. Pure Phil: Standard philosophy subjects
2. Phil CT, no AM: CT subjects taught in a philosophy program, but without AM
3. Phil CT AM: CT subjects taught in a philosophy program, using AM to some extent
4. Phil LAMP: CT subjects taught in a philosophy program, using the LAMP approach. Note that studies in this column also appeared in the previous column.
5. No Phil, Ded CT: Dedicated CT subjects taught outside philosophy programs.
6. No Phil, Some CT: Subjects outside philosophy programs which included some explicit attempt to improve CT
7. No Phil, No CT: A standard university subject, not including philosophy subjects, and where no explicit attempt is made to improve CT.

The height of the column is the weighted average effect size for studies in that category, with 95% confidence intervals.

For current purposes, the third and fourth columns, and the seventh column, are most important. The third column indicates that on average, students in subjects using AM substantially improved their CT skills. In somewhat more technical terms, the weighted average effect size for studies in this category was 0.68, with 95% confidence that the "true" effect size was between 0.51 to 0.85. The fourth column indicates that students experiencing a particular kind of AM instruction – the LAMP method – improved their CT skills by even more, with an effect size of around 0.78. Note that the studies in column four are a subset of those in column three.

The seventh column shows the results of pre- and post-test studies of undergraduate students enrolled in standard undergraduate subjects or programs. It indicates that undergraduates' CT skills typically improve by a modest amount over one semester, around a tenth of a standard deviation. It is noteworthy that there are so many studies in

this category, and that they are relatively tightly clustered. Thus this seems like a fairly robust result, as suggested by the tight confidence intervals. Note also that the modest gain should not be necessarily be credited to wholly undergraduate education. Some gain would be expected in those students anyway, due to maturation and other experiences. The *net* per-semester gain due to university education specifically is likely to be much smaller than one tenth of a standard deviation.

In assessing the effectiveness of AM-based instruction, we should also compare it with other ways of teaching CT. The Alvarez review did include results from studies of other forms of CT instruction, with weighted mean effect sizes ranging from around 0.26 to around 0.4, depending on the category of instruction (though with wide confidence intervals – see Figure 1). Thus her review suggested that AM was substantially more effective, on average, than other kinds of instruction.

This suggestion was further supported by the results of a subsequent meta-analysis, which presented a considerably more comprehensive review of studies of attempts to teach CT, viz., the Abrami *et al*. paper mentioned earlier (Abrami 2008). This review found an overall effect size of around 0.34, very consistent with Alvarez' findings for non-AM CT instruction (columns 2, 5 and 6). Thus the Abrami paper establishes a strong baseline for assessment of the impact of AM-based instruction.

Technical note: Abrami *et al*. included the refinement that d values were corrected for small-sample bias using the formula $g = (1-(3/(4n-9))d$. This correction makes only tiny differences to the overall meta-analytic results.

Overall, the strong suggestion when we combine these reviews is that CT instruction using AM really works: An undergraduate experiencing such instruction will typically improve their CT skills many times faster than they would have with no CT instruction, and substantially faster than when receiving other forms of CT instruction.

These results have helped inspire wider adoption of AM-based instruction, in contexts ranging from primary schools through to intelligence agencies. This activity has resulting in more studies of the impact of such instruction, shedding further light on the

kind of gains one might expect to see. Hence the need for this paper, an updating and refinement of the core result on gains associated with AM-based instruction.

7.5 Method

This study conducted a simple version of a classic meta-analysis, which involves taking relevantly similar studies and "pooling" them to obtain an overall results, more or less as if they were one large study. Our approach is guided by what has come to be called the "new statistics" – replacing traditional null-hypothesis significance testing with effect sizes, confidence intervals and meta-analysis (Cumming 2012). Further, also in the spirit of the new statistics, findings are communicated visually (Cumming and Finch 2005). Since some readers may not be familiar with the process of meta-analysis, an overview of the procedure used here is provided.

7.5.1 Search for Studies

The review first conducted a search for relevant studies as reported in publications or manuscripts. This consisted of searching relevant databases, broadcasting calls on email lists, and querying authors. To be included in the meta-analysis, a study would have to be conducted in the context of an undergraduate subject which used AM; use a pre- and post-testing design to assess skill gains; use an objective or near-objective (adequate inter-rater reliability) test of CT; and report effect size or data sufficient to calculate it. To keep the pool as large as possible, the studies did not need to be published; indeed, unpublished studies were sought in order to avoid the "file drawer" problem. Further, they did not need to show a positive or "statistically significant" result. Finally, note that a "study," for current purposes, is the pre- and post-testing of a cohort of students. A single report (journal article, manuscript etc.) might describe a number of studies. Conversely, a study might be described in more than one report, in which case it was of course included only once in the meta-analysis.

7.5.2 Data extraction

Having identified a set of studies meeting the criteria, the next challenge was to extract the key data needed to compute the key results.

The main objective was to identify the overall effect size associated with AM-based instruction's impact on CT skill, as described above. For this, for each study, the mean pre-test and post-test scores, the test used, and the standard deviation for performance on that test had to be identified. Since the reports of studies were a motley crew ranging from refereed psychology publications through to brief descriptions sent via email, the required information was often difficult to identify in the report and often not contained there at all – which meant contacting (if possible) the author(s) and asking them to send further information, perhaps even the original spreadsheet with raw data. This can be a slow and painstaking process.

With the basic data assembled, effect sizes for individual studies were calculated. The studies and effect sizes are listed in Table 1 below. (More detailed data can be obtained from the original reports, or by contacting the lead author of this paper.)

Study	Citation	n (participants)	Effect Size
1	(Carwie 2009)	56	-0.09
2	(Dwyer, Hogan, and Stewart 2011)	15	0.23
3	(Carrington 2011)	25	0.23
4	(McCoy unpublished)	26	0.36
5	(Bessick 2008)	25	0.79
6	(Dwyer, Hogan, and Stewart 2011)	23	0.17
7	(Butchart 2009)	41	0.22
8	(Butchart 2009)	43	0.45
9	(McCoy unpublished)	27	0.46
10	(Dwyer, Hogan, and Stewart 2012)	42	0.61
11	(ter Berg unpublished)	18	0.62
12	(Donohue et al. 2002) *	53	0.39
13	(Twardy 2004) *	126	0.74
14	(van Gelder, Bissett, and Cumming 2004) *	232	0.80
15	(Donohue et al. 2002)*	50	0.89
16	(Donohue et al. 2002)*	114	0.89
17	(Harrell 2011)	68	1.10
18	(Harrell 2011)	35	1.15

Table 1: Studies of AM-based instruction included in this meta-analysis

7.5.3 Classifying Studies into Intensity Groups

Reviewing studies and the effect sizes emerging from them, three things were obvious. First, there were huge differences in overall effect sizes, ranging from –0.09 to 1.15.

Second, there were large differences in the extent to which participants were required to engage in AM, ranging from Carrington *et al*.'s "minimalist" intervention (Carrington 2011) through to the intensive semester-long AM training described in (Twardy 2004). Third, it was obvious that these were related; the bigger effect sizes tended to be found in the subjects requiring more AM.

The idea that amount of gain might be correlated with and indeed caused by amount of AM has natural plausibility. CT is a skill, and it is a truism that, by and large, skills improve with practice. As mentioned earlier, this idea was refined in (van Gelder, Bissett, and Cumming 2004) into the "quality practice hypothesis" – that CT skills should improve as a function of extent of quality practice of those skills. Further, in some of their studies van Gelder and colleagues actually recorded the amount of practice participants were doing, and found a positive correlation with skill gain.

In general, practice alone does not uniformly produce skill gains, whether in tennis or anything else. The *quality* of that practice also makes a big difference. While it is not easy to say exactly what quality of AM instruction consists in, when perusing the studies, it also seemed clear that there were important differences in key dimensions.

Consequently, to further explore these relationship, studies in the present meta-analysis were coarsely classified into three groups based on what can be termed the "intensity" of AM, combining both quality and quantity in a single rating. The following considerations were taken into account, based on whatever information we could obtain either from the reports, or by contacting authors:

1. Overall duration of the subject
2. Quantity of AM practice within that subject
3. Whether the subject was a dedicated CT subject
4. Whether/how much individual feedback was provided
5. Whether/how much homework was required
6. AM and informal logic experience of the instructors

We were careful to classify studies by intensity having no regard to the results they reported. While this kind of intensity rating is far from a perfect science, it did appear to be reflecting important differences between the studies.

7.5.4 Computing Results

Using a random effects model for meta-analysis (Cumming 2012, 209) the effect sizes and confidence intervals for the entire group of 18 studies, and for the three intensity-based subgroups, were calculated.

7.6 Results

The results are summarised and displayed in Figure 2, whose format is closely based on that of Figure 1:

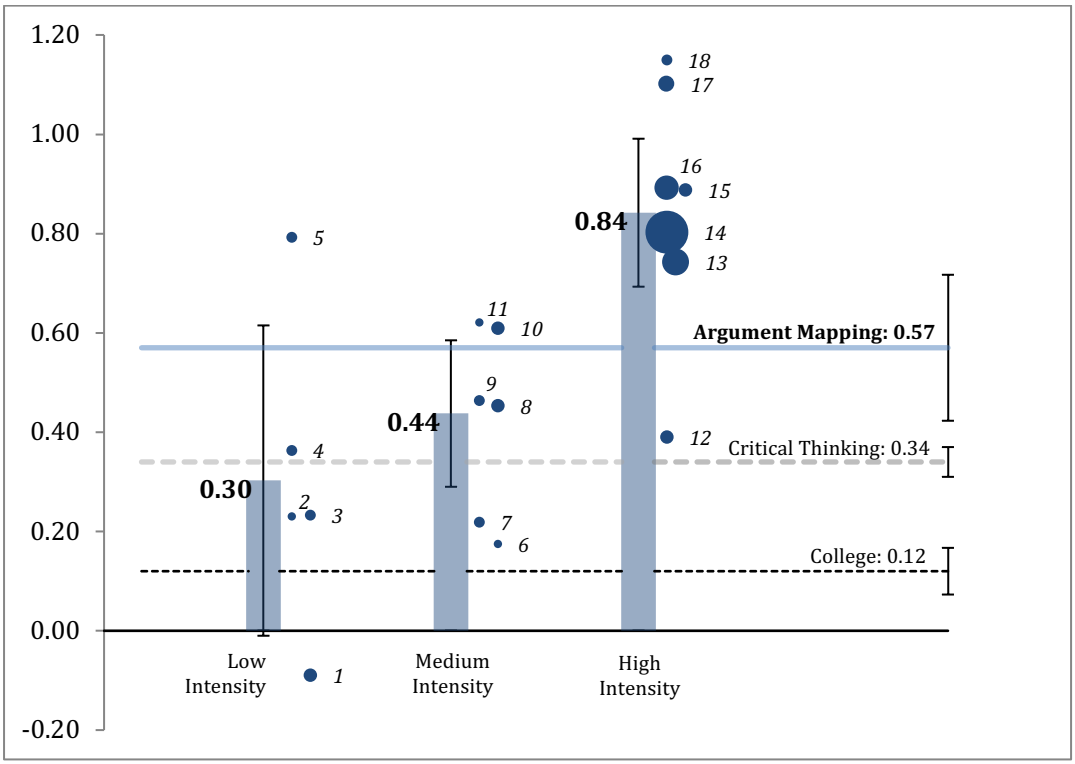


Figure 2: Critical thinking effect sizes for AM studies: results of the current meta-analysis.

In this figure:

1. The number beside each circle indicates which study it is (see Table 1).
2. The "Argument Mapping" horizontal line represents the effect size of argument mapping instruction, based on the 18 studies included in this analysis, with 95% confidence intervals.
3. The "Critical Thinking" horizontal line represents the effect size of CT instruction generally, as found in the Abrami *et al*. meta-analysis.

4. The "College" dotted horizontal line represents the baseline, i.e., effect size found in general undergraduate subjects, as found in (Alvarez 2007) – see Figure 1, column 7.
5. The three columns correspond to low intensity, medium intensity, and high intensity argument mapping studies.

Here is the meta-analysis data, tabulated:

Low Intensity	0.30	[-0.01,	0.61]
Medium Intensity	0.44	[0.29,	0.58]
High Intensity	0.84	[0.69,	0.99]
Argument Mapping -- All	0.57	[0.42,	0.72]
Non-Argument Mapping CT	0.34	[0.31,	0.37]
College per semester	0.12	[0.07,	0.17]

7.7 Discussion

Four main conclusions seem to emerge from the results just displayed.

1. CT instruction based on AM appears to substantially accelerate CT skill gains.

AM-based instruction in general is associated with an effect size of over 0.5. When compared with the baseline gain (i.e., the gain which would have occurred even without any CT instruction) of 0.12, this represents a substantial expected net gain or "value add". Not all of this net gain would be attributable to AM specifically; some of it would have happened anyway because these students were undertaking some kind of CT instruction. However it seems safe to assume that much of it would be. Thus, AM-based CT instruction appears to be substantially more effective than standard undergraduate education at improving CT skills.

2. AM-based instruction appears to be more effective than other kinds of CT instruction.

The effect size associated with CT instruction generally is around 0.34, substantially less than the 0.57 for AM-based instruction. Note that the confidence intervals for CT

instruction, and those for AM-based instruction, do not overlap, meaning we can be confident that the "true" effect of AM-based instruction is larger than that of CT instruction generally. Of course, this does not imply that any particular implementation of AM-based instruction will be more effective than any particular implementation of an alternative form of CT instruction. Indeed, a number of the AM studies included in our analysis showed effect sizes less than that of CT instruction generally.

3. The expected gain from AM instruction depends strongly on the intensity of the AM instruction.

The strong relationship revealed in the difference in the height of the three columns in Figure 2 confirms what we would expect. AM is not some kind of fairy dust which, sprinkled lightly on a university subject, somehow produces extraordinary gains. It is essentially a framework or medium within which CT skills can be practised. The amount of practice done will affect the amount of gain, as will many other aspects of the argument mapping instruction and the instructional context more generally.

4."High intensity" AM-based instruction can produce large gains in CT.

The effect size for high-intensity argument mapping studies is around 0.84, with 95% confidence that the true effect size is between 0.7 and 1.0. By the rule of thumb mentioned earlier, this counts as large effect size. Put another way, it accelerates CT skill gains fivefold or more over the gain typically found over an undergraduate semester. Alternatively, it is triple or more the CT gain over two years of undergraduate education (0.18SD) as identified in Arum and Roska's large study (Arum and Roksa 2011).

Thus, this meta-analysis underscores, with a larger data set, an important conclusion already emerging in Alvarez' analysis. There appears to be at least one way to reliably produce strong gains in CT skills, viz., teach CT using a high-intensity argument mapping approach. However, as usual, the results and conclusions open up a host of new questions, inviting further research.

1. Why does it work?

Insofar as AM does accelerate CT skill gains, why is this? What are the causal mechanisms? Little research has been done on this. The question was partially addressed by van Gelder, who asked why a specific argument mapping software package might facilitate better thinking *performance* (van Gelder 2007). He canvassed three potential causal mechanisms:

- That such software is more "usable" than the standard technologies we use for representing and manipulating reasoning;
- That such software complements the strengths and weaknesses of our inbuilt cognitive machinery; and
- That argument mapping represents a semi-formal "sweet spot" between natural language and formal logic.

It is not hard to imagine how each of these mechanisms may also play a role in facilitating not just performance on a given task, but learning of CT skills. Another potential causal mechanism is that working with argument maps builds, in the learner's minds, mental templates or schemas for argument structures, making it easier for them to critically evaluate argumentation.

2. What dimensions of CT are being enhanced?

CT is multi-dimensional. For example, the HCTA has five "subscales" for different dimensions of CT: verbal reasoning, argument analysis, thinking as hypothesis testing, likelihood and uncertainty, and decision making and problem solving (Halpern 2010). It is plausible that AM-based instruction will be more effective in enhancing some dimensions – say, verbal reasoning, and argument analysis – than others. Closer analysis of data from existing and future studies may shed some light on this.

3. How much CT gain can be generated?

The meta-analysis suggests a strong relationship between intensity of AM and CT gain. Could even greater gains be achieved by even more training? Even the most time-consuming AM regimes in the studies included in this meta-analysis were not

particularly demanding, being only somewhat more challenging than typical undergraduate subjects, and certainly much less intensive than, say, college athletics training. Thus it is plausible that substantially higher gains could be achieved, though of course there must also be practical limits. Given that high-intensity AM-based instruction is already showing gains of around 0.8 standard deviations, it is a reasonable conjecture that this practical limit would be somewhere between one and two standard deviations – which does not of course rule out even larger gains from exceptionally thorough instruction.

What would it take to achieve gains of this order?

- Combining argument mapping (AM) with other general approaches known to enhance learning, such as mastery learning (Kulik, Kulik, and Bangert-Drowns 1990a) and peer instruction (Crouch and Mazur 2001), as suggested by Thomason.
- Developing and deploying automated feedback. One of the enabling conditions for rapid skill acquisition, in general, is timely, good-quality feedback. Having human instructors provide sufficient feedback of adequate quality is a very substantial challenge for AM-based CT instruction under normal resource constraints. Thus we must develop and use rich automated feedback systems of various kinds (Butchart 2009).
- Improved mapping tools. The argument mapping software in use today, while better than nothing, is much less sophisticated than it could be. In particular, improved educational mapping tools will need to integrate automated feedback.

To the extent that conditions such as these can be satisfied, the prospects for very substantial gains in CT being reliably *achievable* via semester-sized instruction using AM are very good. But this would solve only the first of the challenges described in the introduction – that of identifying and refining a suitable instructional practice. The other challenge – achieving widespread adoption – remains formidable. However some reason for optimism can be found in the way technology is driving dramatic transformation in undergraduate education, with the emergence of MOOCs an obvious current example (Davies 2012). The development of argument mapping software was

the trigger for wider adoption of argument mapping as an instructional practice. Continuing that development, particularly in the area of automated feedback, may be able to make high-intensity AM-based instruction widely deployable under practical resource constraints. We would then have a truly effective and affordable way for higher education to achieve its most central ambition.

8 Appendix: Anonymized Reports

8.1 Instructor Report A

Course Description

The course was aimed at improving students' skills of interpreting, reformulating or clarifying, critiquing, and evaluating arguments. Its novel approach was to pursue these aims entirely by way of *argument-mapping* activities, whereby students would transform textual arguments into graphical, tree-structure form, using the Rationale software designed for this purpose, and would also examine, modify, and assess argument maps.

Technique of teaching

The course followed the www.improvingreasoning.com order of modules. Individual pacing was limited, as the entire cadre moved together, with common homework assignments and exams.

The method of instruction was LAMP: students were assigned mapping activities for homework to be brought to class, along with occasional textbook readings, and the class periods were devoted entirely to reviewing homework and working on new mapping activities. There was no lecture, but there was a great deal of discussion – first among small groups of students (two or three) working together on mapping activities on a single laptop computer, then in plenary, with the instructor displaying student maps – and, sometimes, instructor maps – for group analysis on a large screen at the front of the room. (A few students sometimes resisted the directive to work in small groups, insisting upon doing the mapping exercises on their own, though subsequently usually interacting with the assigned group.) Student maps were projected via a long cable that

connected to the instructor's classroom PC, the cable being reconnected to differing student laptops during the session.

During student small-group mapping activities, the instructor often circulated around the room, observing student efforts, asking pointed questions, offering helpful suggestions to push students in more productive directions. Some students met for individual Extra Instruction with the instructor outside of class.

Class morale was generally high. Most students (about 22 of the 24) seemed to be enjoyably engaged in the mapping activities and discussions. There was some small student frustration at times, in reaction generally to having fault found with their maps. The 'indeterminate' or 'pluralistic' nature of mapping (such that there are typically multiple reasonable maps for a given argument text) sometimes seemed to lead some (very few) students to infer that critique of maps is out of place, that the latitude of the mapper is broader than in fact it is.

Course progress: We reached only the beginnings of Module 9, and then hurried to do some additional mapping exercises supplied by Ashley Barnett.

Emerging Concerns about Pacing of Course and the Level of Analytical Detail (Normal vs. Extraordinary Scrutiny)

There were early concerns, after only one month of the course, that my course was lagging significantly behind others. Ashley Barnett repeatedly expressed worry that my students might become bored. These concerns about the pacing of my teaching lasted through the remainder of the course. I was unable to complete all the course modules, despite hurrying toward the end.

As I taught the course, I was only slightly cognizant of any boundary between what have now been called Normal Scrutiny and Extraordinary Scrutiny, other than the need to avoid (1) delving routinely into semantic definitional issues, and (2) "engaging in superfluous speculation (economic or sociological, say) about the grounds underlying some claim made in an argument," even if "a complete map of the issue (or a fully-grounded decision regarding the Contention) might well need to explore these matters"

(from a 13 Apr 2013 email to team members). These two kinds of distracting superfluous efforts aside, I saw no reason to limit the analytical intensity brought to bear in trying to reconstruct and map arguments. My background assumption was, 'The more penetrating analysis and dissection of concepts and issues, the more revealing and enlightening the analysis.' Indeed, I still find myself unable to grasp where to draw a line, in teaching a LAMP course like this, between some appropriate level of analysis and a level that is excessive and to be discouraged.

The Normal/Extraordinary Scrutiny distinction is a crucial but neglected one. For some purposes, such as analyzing whether Iraq had weapons of mass destruction, Extraordinary Scrutiny of the information and argument is clearly needed. For many other purposes, such analysis would be far too time consuming; any person or organization who did this to every claim would be unable to function. Although neglected by critical thinking scholars and teachers, the distinction is crucial to the theory and practice of critical thinking.

Should Critical Thinking courses emphasize Normal or Extraordinary Scrutiny? There is a lot to be said for either approach, and there is no reason to believe that there is a general answer for all students. In this project, I explored the Extraordinary Scrutiny approach, while the other experiments explored the Normal Scrutiny approach. This was not a deliberate choice on anyone's part, but in retrospect the difference is vivid.

As noted below under "Student suggestions for improvement," multiple students reacted to the course by proposing that less time be devoted to the probing analytical discussions that absorbed so much of our classroom time. In the words of one, they suggested "more mapping, less talking."

The question of proper degree of analytical thoroughness had emerged even in a voluntary 'beta' pretrial course run by me during the previous semester. In 22 Oct 2012 email, Ashley Barnett wrote, in response to some of the 'nitpicky' items included by my students in their maps, "if you start going down the path of including definitional or classificatory claims, unless it is really necessary logical analysis can get bogged down. E.g., should the model answer include 'avocados are things' as a co-premise?" He noted that he found this problem "very difficult to deal with when writing mapping

questions"; for instance, some items that were intended to be evidentially irrelevant reasons "actually turned out to be relevant with the inclusion of a stupid co-premise."

The issue surfaced within one month of the start of my experimental course. In the 7 Feb 2013 class, where the task was to improve the provided map of 'Karinna Moskalenko was poisoned by Russian agents' (Module 4, ex. 9), we spent about 60 minutes in an analytical discussion that went a good deal beyond Ashley Barnett's model answers. We raised in class issues such as 'Is deliberate poisoning different from deliberate placement of poison within someone's car?' and 'Is exposure to poison sufficient for being poisoned, or must there be sickness afterwards?' and 'Mustn't it be specified that the sickness was caused by the poison, and was not unrelated in origin?' In my post-class notes, I commented that it was "Helpful to cast our nitpicky analytical discussion in Intel Community terms – analysts needing to ensure they have solid case for Pres to make specific allegations ag'st Russians."

This example nicely shows the advantages and disadvantages of Extraordinary Scrutiny. For a report to the President on a major issue, it would almost be dereliction of duty to not have done a very thorough examination of the issue – an examination that, in many other circumstances, would be reasonably considered "nitpicky."

In a subsequent email exchange on 8-9 Feb, Ashley Barnett expressed concern that I might be going needlessly slowly, pushing the entire class to spend lots of time on advanced mapping problems that had been included for the sake of fast students who might need to keep busy. Maybe more important, the questions I and my students raised about this 'Moskalenko' map raised questions about just what constitutes a completely stated argument.

In the 19 March class, as recorded in my post-class notes, the 'Pakistan' argument ["Pakistan wants to be able to defend itself against the Indian navy. Pakistan needs nuclear submarines if it is to defend itself against the Indian navy. Hence, Pakistan wants nuclear submarines] sparked a commentary by me as to the need to mind a distinction between the all-things-considered variant of certain concepts and the narrower variant. Here, the concept was 'wants,' and my expressed view, which might

well have been pushing the students for more sophisticated analysis than was expected by other IARPA team members in their courses, was as follows:

> … [I discussed] the narrow (pro tanto/prima facie) vs broad (all things considered) notions of 'want.' Yes, if I desire goal G and I believe that H is essential for G, then, *considering that ground/perspective only*, I do desire H. (Eg, I want to do my HW because I want an 'A' in the course and I think doing HW is essential for getting an 'A.') But it might be that this desire for H is outweighed by other factors that make me not want A (I might get horribly nauseated or bored or suicidal whenever I do this HW). So *all things considered*, I do not want H (to do the HW). Thus, I can both want(1) and not want(2) something at the same time. (Better still, I can both want(pro tanto) H and not want(pro tanto) H, with the result that I want (or don't want) H all things considered.)
>
> Thus, if the reasoner is concluding that Pakistan wants nuke subs, all things considered—and this seems the most plausible reading of the argument text (people don't typically offer serious arguments in favor of merely pro tanto claims)—then he's assuming (and we need to supply) a copremise to the effect that 'The counter-considerations against nuke subs from the Pakistanis' point of view are not so serious as to outweigh this reason to want nuke subs.' (Actually, that's not strictly correct; the concern is that the net impact of all the other considerations, both pro and con, is not so con as to outweigh this pro reason—which in my map I states as 'The balance of other perceived nuke sub pros and cons is not so negative as to outweigh, in the Pakistanis' minds, the pro of defending ag'st the Indian Navy.')

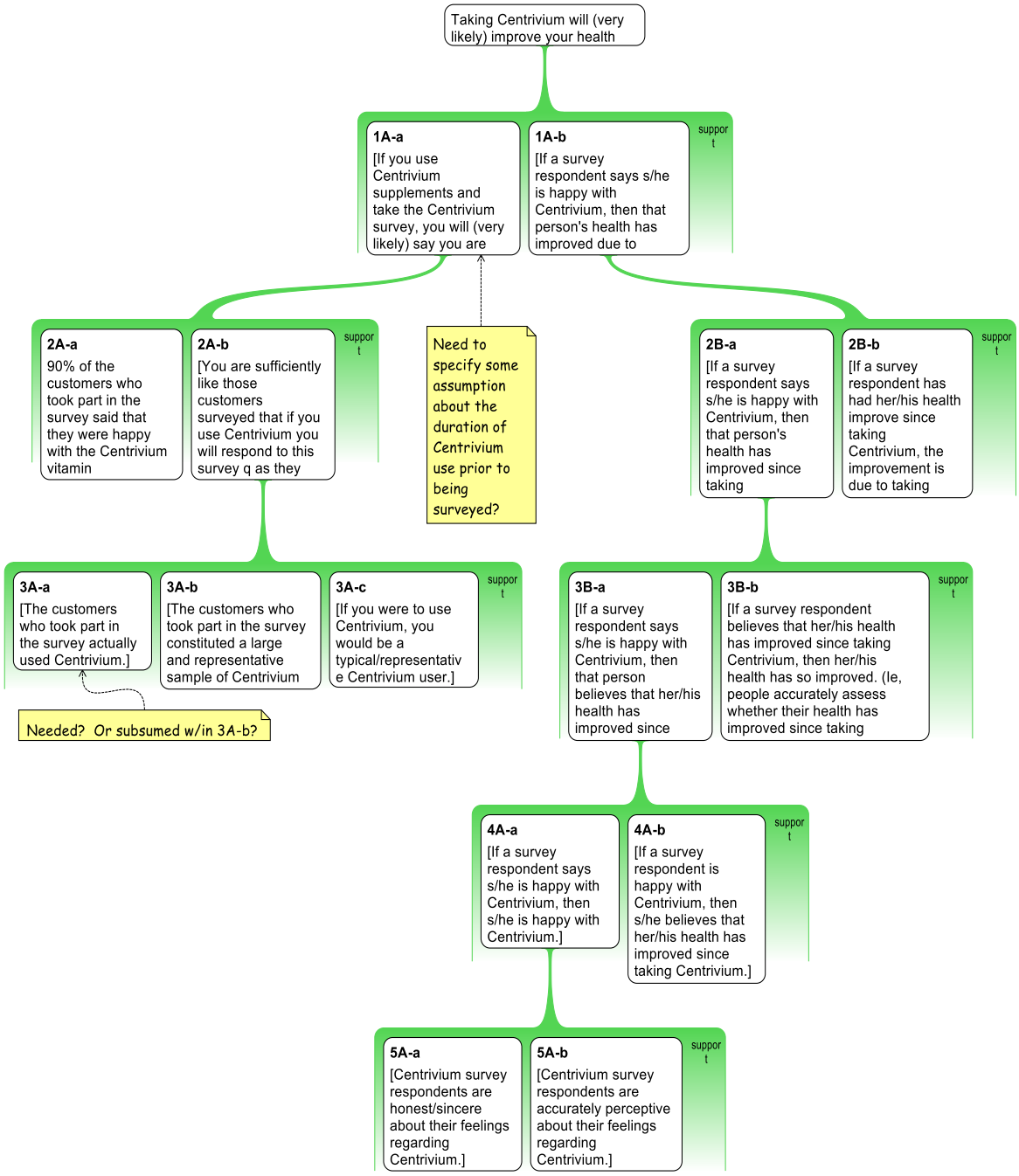
Two days later, in the 21 March class, as recorded in my post-class notes, I "Found myself trying to describe the cognitive method needed for discerning unstated premises":

> Ask yourself: If the existing premises are all true, is the conclusion from them (as reasonably construed) definitely true—or are there possible counter-cases ('what-ifs') whose facts will make the conclusion false (or true only if given some unexpected, deviant construal)? If the latter, then being charitable might demand that we add unstated premises that block those counter-cases. In short: ask yourself how the given premises can be true and yet the conclusion from them still false (or true only in a distorted sense), and block those logical gaps with copremises.
>
> Why all that stuff about "reasonably construed," "unexpected, deviant construal," "true only in a distorted sense"? Because of cases (often intentionally deceptive, as in advertising) where premises support only a weird reading of the conclusion, but the audience is not (supposed to be) aware of this. E.g., 'Reshine moisturizer is good for the skin [if consumed orally in large quantities]'; 'Yes, you've caught me; I am in fact closer to 30 (years old) than I am to 20 [because I'm over 30, not just over 25]'; 'Elect

> me, and you taxes will be lower [though your payments to gov't, when we add in things we won't call "taxes," will be higher].'

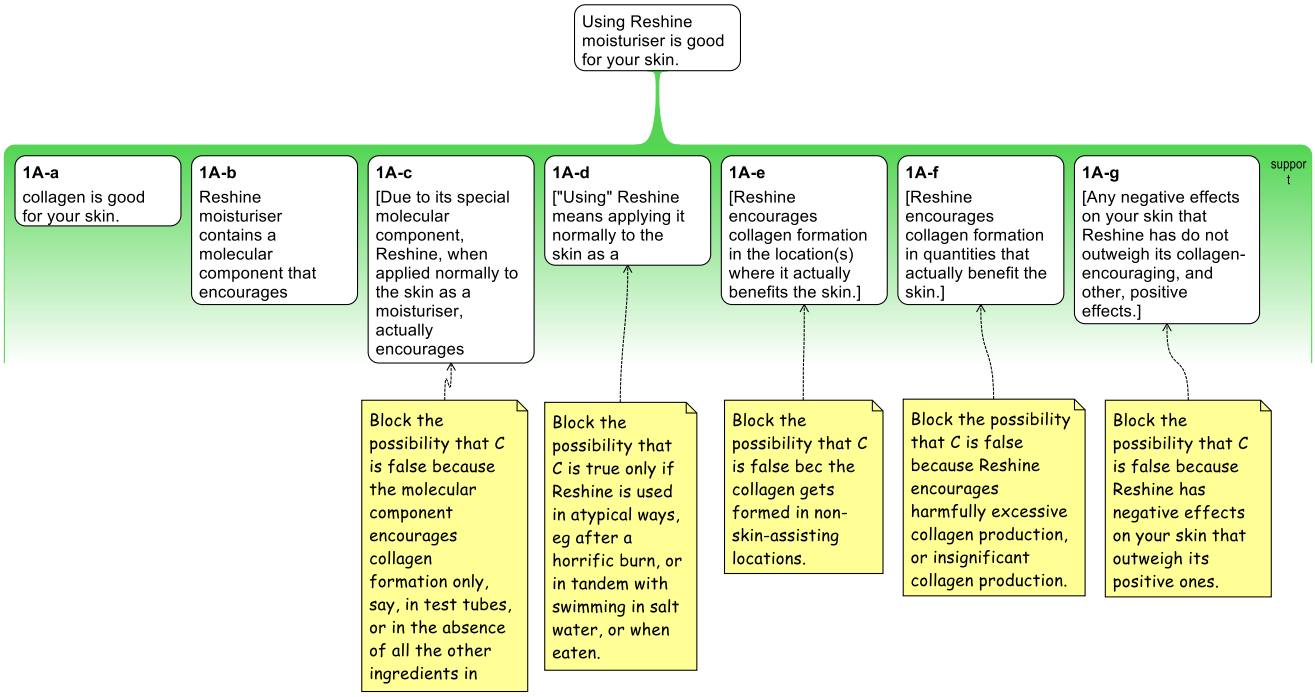
In a 13 Apr 2013 email to team members entitled "Mapping question: Where draw the line on unstated assumptions?", a key element in distinguishing Normal from Extraordinary Scrutiny, I noted my own ongoing struggles with this line-drawing problem and my eagerness to come up with some guideline to share with my students. In the email I asked about 3 different mapping problems. I began as follows:

> First, argument 5 from Module 8:
>
> > 90% of the customers who took part in the survey said that they were happy with the Centrivium vitamin supplements. So you can buy Centrivium with confidence that it will improve your health.
>
> To me, the main point of this exercise is to emphasize the kind of CT that's essential to deploy in dealing with advertising (and which, I think, is prototypical of what CT courses around the world emphasize) – i.e., to smoke out all the hidden assumptions, many of them dubious, that advertisers count on consumers unwittingly making. I've captured (some of) those in my map below….



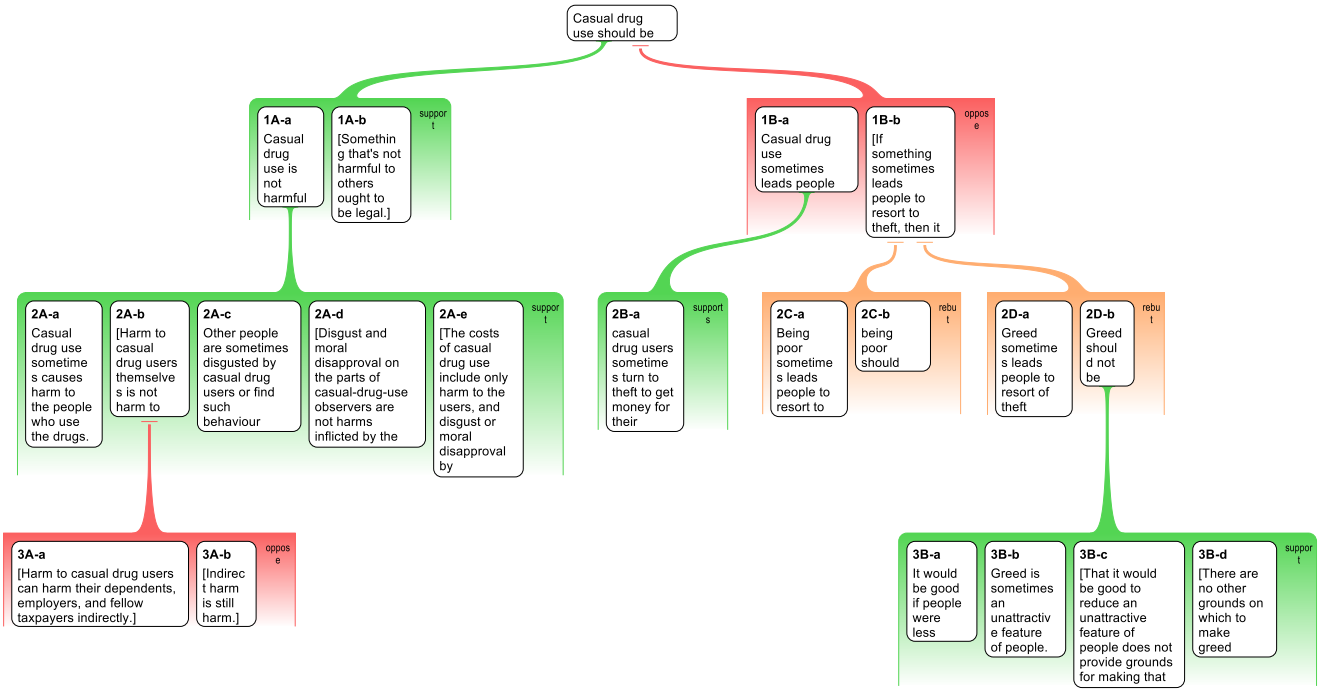
I then presented the similar 'Reshine Moisturiser' argument (Module 7, #5):

> Using Reshine moisturiser is good for your skin. Reshine moisturiser contains a molecular component that encourages collagen formation. As we all know, collagen is good for your skin.



And the 'Drug Use' argument (Module 7, #9)

> Casual drug use should be legal. After all, it is not harmful to others. Perhaps it sometimes causes harm to the people who use the drugs, and maybe others are sometimes disgusted by casual drug users or find such behaviour immoral. But casual drug use is not harmful to others. Yeah, I know, even if it is granted that casual drug use is not harmful to others, some will object to casual drug use being made legal by claiming that casual drug use sometimes leads people to resort to theft, perhaps in order to get money for their drugs. The claim that casual drug use sometimes leads people to resort to theft is questionable but even if we grant its truth, it is a weak objection. Being poor sometimes leads people to resort to theft but it would obviously be silly to claim that being poor should be illegal. Greed sometimes leads people to resort of theft too but no one would think that greed should be illegal. Greed might of course sometimes be an unattractive feature of people but it should not be illegal. But having said that, it would be a good thing if people were less greedy and more considerate of the needs of other people.



I remarked, "it seemed to me natural to capture in the map the putative stated counterarguments against legalizing casual drug use (it disgusts others, it harms the users) all as implicit premises."

The two responses I received to the 13 Apr email both urged a less thorough approach to argument reconstruction and mapping. First, Yanna Rider (on 15 Apr) wrote to "advocate a more minimalist approach to unstated premises."

> Your Centrivium and Reshine examples seem to articulate co-premises by looking for 'truth-makers': What are all the things that would have to be true for R to constitute a (good/reasonable) reason to believe C? Work all those conditions into the map. But then there's a seemingly endless list of such items, pertaining to numerous background conditions that would have to be true, and no obvious way to distinguish between those that must be included and those that must/could be omitted, and that way lies analytic (and mapping) madness.
>
> It seems to me this problem particularly plagues 'bad' arguments, in the sense of 'arguments where there's way too big a leap between the stated premise(s) and the conclusion'. It's not to say that such 'bad' arguments can't be saved, but I don't think it's our job (as mappers) to save them. Apart from anything else, students would need to know about questionnaire design, self-reporting, etc., etc., (or else about molecular structure, clinical testing etc.) in order to articulate an argument with such unstated premises.

> May I be so bold as to suggest that your ability to **already** think critically about the argument is tempting you to map it in this way. But we can't presuppose that ability/knowledge of the students, and no mapping rules or guiding principles are going to bring all those assumptions out without that background knowledge. (I am probably a bit maverick in that I think that *really sophisticated mapping* (of the sort you did) actually depends on the ability to think critically (including knowing about statistical and causal arguments) not vice versa.)
>
> …
>
> The 'casual drug use' argument is, I think, different to the other two. I don't see the statement 'Perhaps it sometimes causes harm to the people who use the drugs, and maybe others are sometimes disgusted by casual drug users or find such behaviour immoral' in the text as being part of the argument. It's kind of a 'psychological', 'consensus-building' concession on the part of the arguer, which I see as irrelevant to the case being made. So I'd definitely omit your 2A from my interpretation of the argument.

A subsequent response from Ashley Barnett (on 29 Apr) included the following comment:

> I avoid this problem all together. As I've been emphasizing for a while, treating mapping as primarily an activity in accurate interpretation leads to all sorts of problems. I treat interpretation as a secondary task to be covered later if at all. People generally write such crappy arguments that it generally isn't worth your while trying to stick closely to the text. Primarily I get students to form the best possible argument that is 'inspired' by a given text. Allowing students to refine claims gives them a lot of freedom to do so. And often I will tell them put the text aside and map out the gist of it in their own words, if they are relying on the text too much. If students want to provide additional reasons and objections, let them go for it. It shows they are thinking about the issue. Just focus on picking them up on the obvious errors – errors in their logical, and gross misreadings of the argument that actually weaken it. This keeps things nice and simple.

Ashley's Normal Scrutiny approach was clearly different from my Extraordinary Scrutiny approach. In retrospect, there are advantages and disadvantages to having the experiments taught with these two approaches.

On one hand, there is an obvious disadvantage to teaching my course with Extraordinary Scrutiny: it probably explains why my results were substantially worse than the general experimental results.

On the other hand, there are benefits as well. There is an important lesson here, one that the Critical Thinking scholarly literature and teaching practices have neglected: "What is the appropriate level of scrutiny for an argument?" "How does one determine what that level is?" The answers will have a large pragmatic element. But answering, "It depends on context" or "This is a matter for judgment," however true, does not help either the theory or the practice. And without straightforward, useful answers to these questions, the practical use of Critical Thinking is limited – as is, I believe, the transfer effect from Critical Thinking courses to applications elsewhere.

Assessment Objective(s) & Results

The central pedagogic objective of the course was for students to obtain unprecedentedly large gains in CT/reasoning skills as demonstrated by performance in highly-respected standardized CT-assessing instruments, the logical reasoning portion of the LSAT and the recently-developed HCTA (Halpern CT Assessment). Those tests do not involve any argument mapping, but the reasoning skills gained via argument mapping were expected to transfer to the kinds of verbal problems offered in the LSAT and HCTA. Because two alternate versions of each test are available, it was possible to administer each test twice – once as a pre-test prior to any course instruction, and once as a post-test, at the conclusion of the course.

The results on the two instruments were astonishingly poor (and also out of step with both end-of-course Student Opinion Forms and instructor observations of individual student skill & performance throughout the semester), raising important questions about the congruence between what is learned by students in the course (as an argument-mapping course) and what is measured by the LSAT and HCTA[19].

On the LSAT, the numbers suggest that the course had no measured effect, within the confidence intervals. The HCTA results are more difficult to assess, because (1) the

[19] Lacking funding for the assessment, and thus needing to grade paper HCTA tests on my own (rather than having students take the tests, and have them graded online), I was able to administer only the short-answer portion of the HCTA. Results from a complete version of the HCTA might perhaps have been different.

official grading guide is unclear or ambiguous in places; (2) the official grading guide reflects questionable implicit judgments of correctness in places; (3) some test questions cover material not included in the course. As a consequence, the tests were regarded using a set of adjustments, thereby reducing (though not eliminating) the problems. The adjusted results show a slight average improvement over the semester, by contrast to the unadjusted scores, whose mean is unchanged but whose variance increases substantially from pre-test to post-test.

The test results are inconsistent in many places with compelling observations I made of individual subjects and their performance throughout the course. For the LSAT, there are numerous post-test scores that are remarkably low, including one that is shockingly low (given that it was a middling score earned by one of my most impressive students), and three pre-test scores that are remarkably high. For the HCTA, there were also some very surprising results that seem to track only poorly what I myself saw from students in the course. But, such anomalies are common in all critical thinking testing, including the other experiments of this project. They probably are not the primary cause of my course results.

The test results are surprising globally, and certainly make a *prima facie* case that the course was poorly designed, or poorly taught, or both. But there is a good reason to doubt such dire conclusions.

My students were taught Intensive Scrutiny and neither the LSAT nor the HCTA measures the ability for Extraordinary Scrutiny. Whatever improvements or otherwise my students had in their Extraordinary Scrutiny ability was not being picked up by these tests. Further, because Extraordinary Scrutiny necessarily takes more time than Normal Scrutiny, insofar as the students were in an Extraordinary Scrutiny mindset, they were slowed down in their test taking, and perhaps more fatigued cognitively as they proceeded.

If this explanation is correct, it opens a range of interesting pedagogic and methodological questions.

Assessment Objective(s) & Results

The central pedagogic objective of the course was for students to obtain unprecedentedly large gains in CT/reasoning skills as demonstrated by performance in highly-respected standardized CT-assessing instruments, the logical reasoning portion of the LSAT and the recently-developed HCTA (Halpern Critical Thinking Assessment). Those tests do not involve any argument mapping, but the reasoning skills gained via argument mapping were expected to transfer to the kinds of verbal problems offered in the LSAT and HCTA. Because two alternate versions of each test are available, it was possible to administer each test twice – once as a pre-test prior to any course instruction, and once as a post-test, at the conclusion of the course.

The results on the two instruments were astonishingly poor (and also out of step with both end-of-course Student Opinion Forms and instructor observations of individual student skill & performance throughout the semester), raising important questions about the congruence between what is learned by students in the course (as an argument-mapping course) and what is measured by the LSAT and HCTA[20].

On the LSAT, the numbers suggest that the course actually *dumbed down* the students a little bit, on average. The HCTA results are more difficult to assess, because (1) the official grading guide is unclear or ambiguous in places; (2) the official grading guide reflects questionable implicit judgments of correctness in places; (3) some test questions cover material not included in the course. As a consequence, the tests were regarded using a set of adjustments, thereby reducing (though not eliminating) the problems. The adjusted results show a slight average improvement over the semester, by contrast to the unadjusted scores, whose mean is unchanged but whose variance increases substantially from pre-test to post-test.

[20] Lacking funding for the assessment, and thus needing to grade paper HCTA tests on my own (rather than having students take the tests, and have them graded online), I was able to administer only the short-answer portion of the HCTA. Results from a complete version of the HCTA might perhaps have been different.

The test results are surprising globally, and certainly make a *prima facie* case that the course was poorly designed, or poorly taught, or both. But there are reasons for hesitating to accept such dire conclusions. Most important, the results are inconsistent in many places with very compelling observations I made of individual subjects and their performance throughout the course. For the LSAT, there are numerous post-test scores that are remarkably low, including one that is shockingly low (given that it was a middling score earned by one of my most impressive students), and three pre-test scores that are remarkably high. The significant declines in individual LSAT performances are also mystifying. For the HCTA, there were also some very surprising results that seem to track only poorly what I myself saw from students in the course.

Student evaluations at end of course

Several students volunteered that they had developed greater capacities for considering differing points of view, for curiosity, for intellectual humility, and for argument analysis. A good number remarked that they consider themselves now to be more careful in thinking, reading, and articulating thoughts. Of course, students (and their instructors) notoriously overestimate the learning that occurs in their courses, and in their CT courses in particular, so this evidence must be taken with due care. But it is striking how out of step it is with the pre-test-post-test outcomes.

Student suggestions for improvement

- offered by multiple students
    - less time devoted to discussing individual analytical issues (or even individual words) in detail; "more mapping, less talking"
    - more variety, incl. harder maps, maybe non-mapping activities
    - more extended mapping exercises, more independent/student-chosen maps
    - "Better MLMs that are not so absolute in the answers that it causes us to resort to bad habits in order to complete them." "The MLMs were frustrating and open to interpretation … perhaps hands-on grading rather than electronic."
- offered by single students
    - *less* group mapping in class, *more* plenary debate & analysis

- start with a pure logic chapter
- don't require printing of homework to be brought to class
- don't insist on continual switching up of partners in group work
- do some argument evaluation
- review textbook material in class before asking for homework on a new chapter

Textbook

Seemed very clear to all (though not last student suggestion above). Seems to be an excellent learning tool.

Some sense that some of the students were just skimming the text chapters, maybe looking primarily at the diagrams, despite the brevity of the assigned chapters.

The original plan to start with basic *argument forms* was a good one, but unfortunately was not implemented.

Never got to any material on argument evaluation – indeed, I don't know whether such material was ever added.

Student Mapping Problems I Noted During the Course

Struggles with knowing how far to go in supplying unstated assumptions, how suspiciously to read the arguments, whether to include reasons that are entirely unstated (and not superordinate to any stated premises).

Struggles with demarcating 'cheap' co-premises from non-cheap ones of like form.

By Module 8, about 75% of the way through the course, I found "an unprecedented discrepancy between the maps I was seeking & expecting, on the one hand, and their HW submissions, on the other, which were way too abbreviated, quick, and heedless of the Rabbit Rule and the warnings about Danglers. I began by restating that our current maps require all the skills we've been using thus far, with only the addition of converting some of the explicit claims into forms that are more clear and accurate. Some of them had taken liberties with this 'revision' permission to rewrite the arguments quite radically, thereby eliding most of the details (including the unstated

assumptions) of the arguments offered. I also reminded them not to vertically compress the reasoning depicted in their maps, as some are wont to do, but to break it down into discrete steps."

In reviewing their second exam, also about 75% of the way through, I "[e]mphasized the need to break down the reasoning into distinct, vertically-layered steps, something that some of the students didn't/don't do – which trips them up as a result and makes it much more cognitively taxing to understand and appraise the argument's reasoning (thereby undermining a major purpose of mapping in the first place)."

8.2 Instructor Report B

8.2.1 Overall assessment.

8.2.1.1 General

Overall, I would say that the course was surprisingly successful, even though it was not the virtually unprecedented success for which we'd hoped. We needed to hit it out of the park on our first at bat. Honestly, I think we hit something close to a triple or a stand-up double – remarkable, though short of our extremely ambitious goal.

I was a sceptic when I agreed to be part of the project. After teaching the class, I now think that argument mapping is considerably more effective than I'd previously thought. I now think that a mapping-heavy approach to teaching reasoning is promising, and worth pursuing. I've integrated more mapping into my standard CT course, and I now require students to purchase and use *Rationale*.

There were several problems with the course, which I detail below, but in my opinion, most of them had to do with either factors extrinsic to the course concept, or having to do with the mastery learning aspect of the course.

8.2.2 Strengths

8.2.2.1 Mapping

In my opinion, the strongest and most promising aspect of the course was also its most salient feature: the emphasis on argument mapping. Mapping turned out to be more effective, interesting and engaging than I had predicted. I do think that the course could have been better with less emphasis on argument reconstruction and more on argument

evaluation. And I acknowledge that we'd need more empirical evidence to definitively attribute the success of the course primarily to mapping. But, in my estimation, the mapping itself was the most effective aspect of the course.

8.2.2.1.1 Active Use of Class Time

The second most important aspect of the course, in my opinion, was what I'll call the active use of class time. When I say that class time was used actively I mean that there was very little lecturing; instead, students worked through mapping problems, generally collaboratively, and we then discussed them as a class.

8.2.3 Weaknesses

8.2.3.1 Novelty of the Course/No Dry Run

The most serious weakness of the course was that it was new. In my experience, courses rarely go terribly well the first time through. Teaching a given type of material, like almost any other activity, can be perfected (or something like it) only with practice. Unfortunately, I didn't have an opportunity for a dry run. (There was a dry run at elsewhere. We tried to put together a small class for a dry run, but to no avail.) Start-up costs were, in my opinion, especially pronounced for this course because it was nonstandard in many ways – I'd never used mapping so extensively before, I had no experience whatsoever with mastery learning, and I had not taught a class that focused on LAMP, nor any other primarily active use of class time.

8.2.3.2 Insufficient Emphasis on Evaluation

With respect to the content, in my view the course should have focused less on argument interpretation and reconstruction and more on evaluation. We were hindered, in my opinion, by the decision to use the LSAT as a pre-/post-test. The LSAT is a rich source of argument evaluation exercises; these exercises were, however, unavailable to us since we could not risk "teaching to the test." In my view, however, the exercises in most critical thinking books are typically of poor quality. Consequently, I spent a fair bit of time trying to write new exercises. I believe we would have been significantly more successful had we emphasized evaluation more, and had a good source of exercises (such as the LSAT).

8.2.3.3 Extrinsic Factors/Bad Luck

My course also suffered greatly from a large number of problems extrinsic to the course. Not to make excuses, but I suffered from a particularly nasty bout of insomnia while teaching the class. We also faced an inordinate number of technical problems; for example: students were unable to use *Rationale* on their Macs; our computer classroom did not have the capacity to project student work on the class screen until almost halfway through the semester (and then the first program we were given did not work well; we didn't have a capacity to do this reliably until about two-thirds of the way through the course); I did not have the capacity to monitor student work from my computer until two-thirds of the way through the semester; the classroom turned out to be the slowest computer lab on campus; the classroom was extremely uncomfortable, primarily because it was extremely hot (alarms prevented us from even propping open the doors until we were able to have them disarmed halfway through the semester); one quarter of the students lost their answers to the first exam because of a peculiarity of the system in our classroom; several glitches with the mastery learning quizzes early in the semester caused considerable confusion among the students. These problems, combined with the problems mentioned above, were extremely detrimental to the course. I was constantly playing catch-up and, as they say, putting out brush fires. Obviously these problems are extrinsic to the concept of the course.

8.2.4 Specific Aspects of the Course

8.2.4.1 Classes/students

I taught two courses, a freshman critical thinking course of 20 students and a junior-level course of 12 students. The material in the two classes was largely the same, though the advanced class tackled much more difficult material toward the end of the course. One student dropped the critical thinking course early in the semester; no students dropped the intelligence analysis course. Classes lasted fifty minutes and met three times per week. There was relatively little lecturing. Most class time was devoted to working on and discussing mapping problems.

8.2.4.2 Student Feedback

Student written comments on course evaluations were unusually positive, especially from the [redacted] students. In fact, the comments from the students were among the most positive I have ever received. Verbally, students expressed great enthusiasm for the class, insisted that the class should be taught again, and many of them said that they would take another, more advanced mapping course were it offered. My sense was the morale in both classes was good, and I suspect that this had something to do with the active nature of the class. Lectures can become rather deadening, especially in the age of PowerPoint.

8.2.4.3 Mastery Learning and MLMs

I was not a fan of the mastery learning approach. I realize that there is evidence of its effectiveness, but I also think that there's a limit to the amount of novelty one can build into any one class. Students did not care for the mastery learning quizzes, and quickly started to semi-game the system. They realized that if they missed more than one question as they worked through the quiz, it was most efficient for them to simply quit and start over later instead of spending more time and energy on a quiz they had already failed. There were also some technical and content glitches (e.g., typos, unrecognized alternate correct answers) with the quizzes that exacerbated the problems. In short, my tentative conclusion was that the mastery learning aspect of the course was not particularly successful, though, without a mapping-based, non-mastery-learning-based class with which to compare the class, this conclusion must be tentative.

8.2.4.4 Materials

8.2.4.4.1 Materials: General

Class materials were generally notably good. I believe such things tend to improve significantly, however, when refined over time; so, in keeping with one of my general thoughts about the class, let me say that I think they could be improved were the course taught again.

8.2.4.4.2 Textbook

In my opinion, most critical thinking textbooks are not terribly good, but I have a high opinion of the text for this class. Its biggest weakness was a lack of a completed chapter on argument evaluation – though that is a very significant weakness indeed. In fact, let me repeat that I believe that one of the main failings of the class was that it overemphasized argument interpretation/reconstruction and underemphasized argument evaluation.

8.2.4.4.3 Practice Activities/Exercises

The mapping practice exercises were absolutely crucial – about as important as the text, and perhaps even more important. I thought the mapping exercises were very good, but not excellent. In fact, in my opinion, this is one of the most salient ways in which the course could/would be better if taught again: there could and should, in my opinion, be fairly extensive revisions of the exercises. In my opinion one of the errors we made was to rely almost exclusively on one person to produce the exercises. This was simply too big a task for one person. Although he did yeoman's work, I don't think that the exercises were as strong as they could have been. Furthermore, I disagree with the strategy of insisting on co-premises at every point. I believe that there are cases in which the inclusion of explicit co-premises (and that's what the early exercises were like) is extremely unnatural and confusing, and this approach does not teach students to rely on the content of the premises/conclusions in constructing their maps. In my opinion, too many of the mapping exercises had the same rather unnatural style. I think this was, in part, because of the insistence on co-premises and, in part, on the fact that all the exercises were written by the same person.

8.2.4.4.4 Mastery Learning Milestone Quizzes (MLMs)

(See above *8.2.4.3)*.

9 Appendix: Other ways to use argument maps to teach critical thinking

While Dr Adajian's JMU seminar on Kant's *Critique of Aesthetic Judgment* was not part of the experiment proper and has no pre/post test data, his observations are illuminating and indicate some possible ways that future argument-mapping based courses could be developed. His experience was that the argument maps were clearer and more understandable than the written or verbal discussions presented by students, and that students understood the text much better than when using traditional methods:

> We spent the semester working through the first Part (Critique of Aesthetic Judgment) of the *Critique of Judgment*. Each week students had to bring in maps of whichever sections we read that week. In most classes I spent a significant amount of time going over my own map of the relevant section(s). For their first graded assignment, they turned in a short paper, a map of the paper, and maps of the relevant sections of Kant. The maps were markedly better than the papers. I have just gotten the second papers. The assignment was similar. Their final assignment will be in two stages (i) doing presentations the last week of class, which include presenting a short paper, its map, and maps of the relevant sections of Kant, for class critique; (ii) turning in final versions of those papers and maps, improved, we hope, in the light of the class critique.
>
> I spent very little time explicitly teaching the maps in class, other than talking about co-premises, the Rabbit Rule, and adding implicit premises. Discussion of their maps in class was usually brisk; I didn't do it every class. Nevertheless, the students' maps have improved markedly over the course of the semester. I'm hoping that the last papers come up to the quality of the maps. Many students seem to think that they need to sound "educated" or "literary" or something like that. The papers are often badly written, wordy, hard to follow, while the maps of those very papers are clear and to the point.

So ever since the first paper, I've instructed them to write the maps of their papers before they write the papers. (In fact I told them that I debated with myself as to whether to have them write papers at all, but decided to since writing papers is still regarded as the norm in academia and elsewhere.)

In prepping them to write their next papers/maps, I went over my own map of a fairly complex, but not too technical, section, up on the screen. We went through an exercise in which I asked them "If this was your map, how would you write your paper from it?" It was something of a revelation to them that the gross structure of the map was a very good guide to organizing the paper, and also a revelation that they can say things in their papers like: **I will argue for claim C. The basic argument for C has 3 premises. In section 1, I explain two arguments for the first premise, and consider one reply to each of those arguments. In section 2, I explain an argument for the second premise. In section 3…** and don't have to sound "literary" and "intellectual" (i.e., polysyllabic and vague).

I've also done a fair amount of grouping in my own maps – the ones I present the material with in class – and they have in their own as well. I didn't do grouping in every map, or even every week. Probably about 40% of the time, depending on where in Kant we were and whether or not the need to do so leaped out of the text, so to speak. Oftentimes he is at pains to provide some sort of taxonomy of his own concepts (e.g., different kinds of necessity, different kinds of pleasure, different varieties of aesthetic judgment, etc.), and then it's very useful; other times this concern isn't in the foreground. The students find constructing their own groupings, based on Kant's text, very helpful. (I just told them to make their own decisions, based on the text, about whether to do groupings in addition to maps.) They have found mine very helpful too.

I should also add that most of the students (granted, there were only 4 – 3 are good, 1 a slacker) have said, with great conviction, that mapping the arguments was the only way that they would have understood Kant on their own, and described the mapping as invaluable. The quality of their maps is really quite good (again, it's not only much better than their writing – it's much better than their speaking in class). The

third *Critique* is very dense and difficult, as you know, and I don't really have any doubt that they wouldn't be reading as carefully and comprehending as much without the maps.

It's also been a great exercise for me, philosophically, to have mapped most of the first half of the *Critique*. It's certainly made the structure of it much clearer to me – the structure of the individual sections, the structure of the individual arguments, and the gross structure of the whole first Part. It's made me eager to go back, re-do, polish. I can't say for sure that it's given me insights into Kant, or stimulated ideas, that I wouldn't have gotten from teaching without maps, but I suspect it has.

The class was very small, but I got some very positive written comments from the students.

> *Question*: Did your mapping of Kant's CJ help you to understand it? Please explain.
>
> *Student*: "Yes! I would read a passage and think I understood what he was saying, but the maps actually made me focus on his specific words and what they meant. The maps also helped me see and pick out his arguments."
>
> *Question*: How would you compare (a) the process of reading Kant while trying to map his arguments with (b) the process of reading Kant without trying to map the arguments?
>
> *Student*: "I would not have understood it [Kant's text]. Or, I would have read it and come to a conclusion completely different than the one he intended. This class without maps would have been impossible for me."

Adajian's class, and Eliana Horn's forthcoming short-intervention study at Monash should give us a far better understanding of the strengths and weaknesses of providing maps to students to help them understand and discuss texts.

10 Appendix: Statistical Data, with additional analyses

10.1 Data

Immediately below is the data for each individual for each test in each experiment, followed by the untrimmed data by experiment, by test, and by gender. This is followed by a somewhat shorter analysis of the trimmed data, since trimming did not prove to be produce much difference, except in cases where the trimming procedure produced lopsided trims within a test within an experiment. Since there is good reason to hold that the Intensive Scrutiny was a different intervention, in addition to analysing all seven experiments together, there is a comparison of Experiment 7 to the six Normal Scrutiny experiments.

In the raw data for each test for each experiment, we have indicated with a plus ("+") those data points that were trimmed from the top, and a minus ("–") those data points that were trimmed from the bottom, in the trimming procedures. Below each test, the number of data points removed from the top (e.g., a pre/post difference of 25) or from the bottom (e.g., a pre/post difference of –8). As one would expect, many of the test results for each experiment had an unequal number of data points removed, some more from the top, others more from the bottom. For example, Experiment 1 had 3 data points removed from the top, and none from the bottom, thereby substantially decreasing average change in raw score and so, not surprisingly, the raw score effect size.

Experiment 1, Test: HTCA

ID	Gender	Pre-Test Form	Post-Test Form	Pre-Test Score	Post-Test Score	Difference	Trimming
1	F	B	A	42	67	25	+
2	F	B	A	53	62	9	
3	M	B	A	75	67	-8	–
4	F	A	B	68	68	0	
5	F	A	B	55	74	19	+
6	M	B	A	32	73	41	+
7	M	A	B	62	64	2	
8	M	A	B	63	67	4	
9	F	A	B	73	71	-2	
10	M	A	B	59	64	5	
11	F	B	A	65	72	7	
12	F	A	B	56	65	9	
13	F	A	B	67	73	6	
14	M	B	A	71	67	-4	

3 from the top; 1 from the bottom

Experiment1, Test: LSAT

ID	Gender	Pre-Test Form	Post-Test Form	Pre-Test Score	Post-Test Score	Difference	Trimming
1	F	B	A	13	14	1	
2	F	B	A	12	13	1	
3	M	B	A	12	16	4	
4	F	A	B	9	12	3	
5	F	A	B	14	15	1	
6	M	B	A	14	21	7	+
7	M	A	B	17	18	1	
8	M	A	B	13	20	7	+
9	F	A	B	14	13	-1	
10	M	A	B	9	10	1	
11	F	B	A	6	10	4	
12	F	A	B	7	8	1	
13	F	A	B	11	14	3	
14	M	B	A	15	16	1	

2 from the top; 0 from the bottom

Experiment 2

Due to a bureaucratic error, all subjects in this experiment were given the same CCTST pre-tests and post- tests and many were given the same LSAT post-test as pre-test.

Test: CCTST

ID	Gender	Pre-Test Form	Post-Test Form	Pre-Test Score	Post-Test Score	Difference	Trimming
1	F	A	A	16	13	-3	_
2	F	A	A	17	25	8	+
3	F	A	A	14	19	5	
4	M	A	A	17	20	3	
5	F	A	A	14	21	7	
6	F	A	A	14	22	8	
7	F	A	A	15	20	5	
8	M	A	A	18	20	2	
9	F	A	A	16	19	3	
10	F	A	A	15	20	5	
11	F	A	A	21	24	3	
12	F	A	A	23	23	0	
13	M	A	A	18	23	5	
14	F	A	A	13	20	7	
15	F	A	A	14	25	11	+
16	F	A	A	19	19	0	
18	M	A	A	21	27	6	
19	M	A	A	19	20	1	
20	M	A	A	23	27	4	

2 from the top; 1 from the bottom

Experiment 2, Test: LSAT

ID	Gender	Pre-Test Form	Post-Test Form	Pre-Test Score	Post-Test Score	Difference	Trimming
1	F	A	A	10	12	2	
2	F	B	B	14	19	5	
3	F	A	A	7	6	-1	
4	M	B	B	9	17	8	+
5	F	B	B	12	9	-3	
6	F	A	A	11	15	4	
7	F	B	A	11	13	2	
8	M	A	B	15	17	2	
9	F	B	A	14	14	0	
10	F	A	B	9	10	1	
11	F	A	B	16	17	1	
14	F	A	A	10	9	-1	
15	F	B	B	11	14	3	
16	F	A	A	13	15	2	
17	M	A	A	19	22	3	
18	M	B	B	19	18	-1	
20	M	A	A	15	18	3	

1 from the top; 0 from the bottom

Experiment 3,

Due to a bureaucratic error, all subjects in this experiment were given the same CCTST pre-tests and post- tests and many were given the same LSAT post-test as pre-test.

Test: CCTST

ID	Gender	Pre-Test Form	Post-Test Form	Pre-Test Score	Post-Test Score	Difference	Trimming
M30	M	A	A	22	24	2	
M31	M	A	A	21	26	5	
M33	F	A	A	25	22	-3	–
M34	M	A	A	23	28	5	
M36	M	A	A	21	20	-1	
M37	M	A	A	26	25	-1	
M39	M	A	A	15	20	5	
M40	M	A	A	12	19	7	
M41	M	A	A	22	29	7	
M43	M	A	A	23	26	3	

0 from the top; 1 from the bottom

Experiment 3, Test: LSAT

ID	Gender	Pre-Test Form	Post-Test Form	Pre-Test Score	Post-Test Score	Difference	Trimming
M30	M	A	A	11	8	-3	
M31	M	B	B	18	18	0	
M32	M	A	A	11	10	-1	
M33	F	A	A	15	16	1	
M34	M	A	A	16	18	2	
M35	M	B	B	11	15	4	
M36	M	A	B	13	9	-4	–
M37	M	B	B	22	24	2	
M39	M	A	A	10	13	3	
M40	M	B	B	7	14	7	+
M41	M	B	B	15	19	4	
M43	M	B	B	15	18	3	

1 from the top; 1 from the bottom

Experiment 4, Test: CCTST

ID	Gender	Pre-Test Form	Post-Test Form	Pre-Test Score	Post-Test Score	Difference	Trimming
1	F	A	B	22	27	5	
2	M	B	A	12	15	3	
3	M	B	A	18	21	3	
4	M	B	A	17	20	3	
5	M	A	B	16	13	-3	–
6	M	A	B	19	23	4	
7	M	A	B	17	24	7	

0 from the top; 1 from the bottom

Experiment 4, Test: LSAT

ID	Gender	Pre-Test Form	Post-Test Form	Pre-Test Score	Post-Test Score	Difference	Trimming
1	F	B	A	20	17	-3	
2	M	A	B	12	16	4	
3	M	A	B	13	14	1	
4	M	B	A	17	17	0	
5	M	A	B	12	13	1	
6	M	A	B	17	16	-1	
7	M	B	A	9	10	1	

0 from the top; 0 from the bottom

Experiment 5, Test: CCTST

ID	Gender	Pre-Test Form	Post-Test Form	Pre-Test Score	Post-Test Score	Difference	Trimming
1	M	B	A	23	24	1	
2	M	B	A	26	29	3	
3	M	B	A	27	30	3	
4	F	B	A	27	28	1	
5	M	A	B	20	29	9	+
6	F	B	A	29	28	-1	
7	M	B	A	30	31	1	
8	M	B	A	25	29	4	
9	M	B	A	24	28	4	
11	M	B	A	30	30	0	
12	M	A	A	22	29	7	
15	M	A	B	22	28	6	
16	M	B	A	22	23	1	

1 from the top; 0 from the bottom

Experiment 5, Test: LSAT

ID	Gender	Pre-Test Form	Post-Test Form	Pre-Test Score	Post-Test Score	Difference	Trimming
1	M	A	B	16	22	6	
2	M	A	B	15	21	6	
3	M	B	A	23	24	1	
4	F	B	A	17	17	0	
5	M	A	B	10	14	4	
6	F	B	A	18	17	-1	
8	M	A	B	19	18	-1	
9	M	A	B	13	15	2	
10	F	B	A	13	13	0	
11	M	B	A	16	20	4	
12	M	A	B	19	19	0	
13	M	B	A	19	22	3	
14	M	A	B	16	18	2	
15	M	A	B	20	19	-1	
16	M	B	A	15	12	-3	

0 from the top; 0 from the bottom

Experiment 6, Test: HCTA

ID	Gender	Pre-Test Form	Post-Test Form	Pre-Test Score	Post-Test Score	Difference	Trimming
1	F	B	A	71	69	-2	
2	M	B	A	70	71	1	
3	M	A	B	74	80	6	
4	M	A	B	75	80	5	
5	F	A	B	64	69	5	
6	M	B	A	72	76	4	
7	M	B	A	77	78	1	
8	M	A	B	53	77	24	+
9	F	B	A	60	67	7	
10	M	A	B	73	76	3	
11	F	A	B	72	69	-3	
12	F	B	A	62	73	11	
13	F	B	A	68	73	5	
14	F	B	A	63	68	5	
15	M	B	A	69	70	1	
16	M	A	B	71	81	10	
17	F	A	B	70	70	0	
18	F	A	B	70	76	6	
19	F	A	B	75	77	2	
20	M	A	B	71	75	4	
21	M	A	B	76	69	-7	
22	M	A	B	64	77	13	
23	F	B	A	70	70	0	
24	F	A	B	65	76	11	
25	M	B	A	75	65	-10	
26	M	A	B	66	64	-2	
27	F	A	B	70	72	2	
28	F	B	A	70	71	1	
29	F	B	A	64	63	-1	
30	M	A	B	41	52	11	
31	M	B	A	72	78	6	
32	M	A	B	63	71	8	
33	F	B	A	68	77	9	
34	F	B	A	64	67	3	
35	M	A	B	61	66	5	
36	M	B	A	50	63	13	
37	F	B	A	63	73	10	
38	F	B	A	76	75	-1	
39	M	A	B	71	75	4	

1 from the top; 0 from the bottom

Experiment 6, Test: LSAT

ID	Gender	Pre-Test Form	Post-Test Form	Pre-Test Score	Post-Test Score	Difference	Trimming
1	F	B	A	14	10	-4	–
2	M	A	B	18	22	4	
3	M	A	B	16	21	5	
4	M	B	A	22	22	0	
5	F	B	A	10	14	4	
6	M	A	B	12	17	5	
7	M	B	A	10	14	4	
8	M	A	B	14	10	-4	–
9	F	A	B	8	11	3	
10	M	A	B	17	20	3	
11	F	B	A	19	16	-3	
12	F	B	A	8	15	7	+
13	F	B	A	19	18	-1	
14	F	B	A	8	11	3	
15	M	B	A	23	24	1	
16	M	B	A	17	20	3	
17	F	A	B	18	15	-3	
18	F	A	B	19	20	1	
19	F	A	B	16	18	2	
20	M	A	B	15	21	6	
21	M	A	B	18	17	-1	
22	M	A	B	12	16	4	
23	F	B	A	11	17	6	
24	F	A	B	19	18	-1	
25	M	A	B	14	14	0	
26	M	B	A	11	9	-2	
27	F	A	B	17	16	-1	
28	F	B	A	13	16	3	
29	F	A	B	10	11	1	
30	M	A	B	15	19	4	
31	M	A	B	16	17	1	
32	M	B	A	15	15	0	
33	F	A	B	13	14	1	
34	F	B	A	15	14	-1	
35	M	A	B	10	10	0	
36	M	B	A	9	12	3	
37	F	B	A	6	9	3	
38	F	A	B	17	18	1	
39	M	A	B	14	10	-4	–

1 from the top; 3 from the bottom

Experiment 7, Test: HCTA

ID	Gender	Pre-Test Form	Post-Test Form	Pre-Test Score	Post-Test Score	Difference	Trimming
1	M	A	B	69	81	12	
2	F	A	B	70	70	0	
3	M	A	B	73	77	4	
4	M	B	A	72	73	1	
5	M	B	A	75	65	-10	–
6	M	B	A	70	57	-13	–
7	M	A	B	74	78	4	
8	M	A	B	78	71	-7	
9	M	A	B	72	70	-2	
10	F	A	B	66	65	-1	
11	M	B	A	78	75	-3	
12	M	B	A	72	70	-2	
13	M	B	A	72	77	5	
14	M	A	B	75	74	-1	
15	M	B	A	68	69	1	
16	M	B	A	71	62	-9	–
17	M	A	B	75	80	5	
18	F	A	B	69	69	0	
19	F	A	B	65	68	3	
20	M	A	B	68	71	3	
21	M	B	A	74	75	1	
22	M	B	A	76	78	2	
23	M	B	A	70	69	-1	
24	M	B	A	62	71	9	

0 from the top; 3 from the bottom

Experiment 7, Test: LSAT

ID	Gender	Pre-Test Form	Post-Test Form	Pre-Test Score	Post-Test Score	Difference	Trimming
1	M	A	B	13	13	0	
2	F	A	B	15	16	1	
3	M	A	B	18	21	3	
4	M	B	A	17	20	3	
6	M	B	A	11	9	-2	
7	M	B	A	12	12	0	
8	M	A	B	21	22	1	
9	M	A	B	23	20	-3	
10	F	B	A	15	15	0	
11	M	A	B	22	17	-5	–
12	M	B	A	16	16	0	
13	M	A	B	11	12	1	
14	M	B	A	16	16	0	
15	M	A	B	18	12	-6	–
16	M	B	A	11	12	1	
17	M	B	A	14	15	1	
18	F	A	B	20	18	-2	
19	F	A	B	15	22	7	+
20	M	A	B	19	17	-2	
21	M	A	B	16	13	-3	
22	M	B	A	17	23	6	
23	M	B	A	23	20	-3	
24	M	B	A	13	10	-3	

1 from the top; 2 from the bottom

10.2 Analysis of Untrimmed Data

10.2.1 Experiment Data

Experiment	Test	N	% of improved cases	Mean			SD		
				Pre-Test	Post-Test	Difference	Pre-Test	Post-Test	Average Pre/post
1	HCTA	14	71.4%	60.1	68.1	8.1	3.8	12.0	8.9
	LSAT	14	92.9%	11.9	14.3	2.4	3.8	3.1	3.5
2	CCTST	19	84.2%	17.2	21.4	4.2	3.3	3.1	3.2
	LSAT	17	70.6%	12.6	14.4	1.8	4.2	3.4	3.8
3	CCTST	10	70.0%	21.0	23.9	2.9	3.5	4.3	3.9
	LSAT	12	66.7%	13.7	15.2	1.5	4.7	4.0	4.4
4	CCTST	7	85.7%	17.3	20.4	3.1	5.0	3.0	4.1
	LSAT	7	57.1%	14.3	14.7	0.4	2.6	3.8	3.3
5	CCTST	13	84.6%	25.2	28.2	3.0	2.3	3.3	2.8
	LSAT	15	53.3%	16.6	18.1	1.5	3.5	3.2	3.4
6	HCTA	38	81.6%	67.2	72.0	4.8	5.7	7.4	6.6
	LSAT	39	61.5%	14.3	15.7	1.4	4.0	4.1	4.0
7	HCTA	24	50.0%	71.4	71.5	0.0	5.8	4.0	5.0
	LSAT	23	39.1%	16.3	16.1	-0.2	4.1	3.7	3.9

Experiment	Test	Using Pre-Test SD			Using Pre/Post Av SD		
		d	d_{unb}	95% CI d_{unb}	*d*	d_{unb}	95% CI d_{unb}
1	HCTA	0.674	0.634	[0.064, 1.723]	0.908	0.854	[0.064, 1.723]
	LSAT	0.775	0.729	[0.245, 1.143]	0.703	0.661	[0.245, 1.143]
2	CCTST	1.350	1.293	[0.663, 1.923]	1.303	1.248	[0.663, 1.923]
	LSAT	0.515	0.490	[0.094, 0.815]	0.460	0.439	[0.094, 0.815]
3	CCTST	0.671	0.614	[0.071, 1.373]	0.737	0.674	[0.071, 1.373]
	LSAT	0.372	0.346	[-0.092, 0.766]	0.344	0.320	[-0.092, 0.766]
4	CCTST	1.034	0.899	[0.048, 1.442]	0.764	0.664	[0.048, 1.442]
	LSAT	0.112	0.098	[-0.368, 0.622]	0.132	0.115	[-0.368, 0.622]
5	CCTST	0.906	0.848	[0.341, 1.745]	1.057	0.989	[0.341, 1.745]
	LSAT	0.452	0.427	[-0.007, 0.863]	0.435	0.411	[-0.007, 0.863]
6	HCTA	0.646	0.633	[0.417, 1.021]	0.722	0.707	[0.417, 1.021]
	LSAT	0.334	0.327	[0.094, 0.573]	0.336	0.329	[0.094, 0.573]
7	HCTA	0.010	0.010	[-0.453, 0.470]	0.008	0.008	[-0.453, 0.470]
	LSAT	-0.058	-0.056	[-0.382, 0.272]	-0.055	-0.054	[-0.382, 0.272]

10.2.2 Level of Scrutiny

Using Pre-Test SD

Effect Size Measure	Test	Extraordinary Scrutiny Experiment 7		Normal Scrutiny (Experiments 1 to 6)		All Expts	
		Effect Size (ES)	95% CI	Wgtd Avg (ES)	95% CI	Wgtd Avg (ES)	95% CI
D	CCTST	-	-	0.984	[0.676, 1.291]	0.984	[0.676, 1.291]
	HCTA	0.010	[-0.453, 0.470]	0.649	[0.392, 0.905]	0.494	[0.271, 0.717]
	LSAT	-0.058	[-0.382, 0.272]	0.412	[0.272, 0.552]	0.339	[0.210, 0.467]
	All	-0.035	[-0.299, 0.230]	0.538	[0.424, 0.652]	0.448	[0.343, 0.553]
d_{unb}	CCTST	-	-	0.896	[0.615, 1.177]	0.896	[0.615, 1.177]
	HCTA	0.010	[-0.453, 0.470]	0.633	[0.383, 0.883]	0.480	[0.263, 0.698]
	LSAT	-0.056	[-0.382, 0.272]	0.388	[0.255, 0.521]	0.321	[0.199, 0.443]
	All	-0.034	[-0.289, 0.222]	0.509	[0.401, 0.617]	0.427	[0.327, 0.526]

Using Pre/post Test SD

Effect Size Measure	Test	Extraordinary Scrutiny (Experiment 7)		Normal Scrutiny (Experiments 1 to 6)		All Expts	
		Effect Size	95% CI	Wgtd Avg	95% CI	Wgtd Avg	95% CI
D	CCTST	-	-	0.938	[0.635, 1.240]	0.938	[0.635, 1.240]
	HCTA	0.008	[-0.453, 0.470]	0.738	[0.475, 1.001]	0.554	[0.327, 0.782]
	LSAT	-0.055	[-0.382, 0.272]	0.393	[0.254, 0.533]	0.324	[0.196, 0.452]
	All	-0.034	[-0.298, 0.231]	0.536	[0.422, 0.650]	0.446	[0.342, 0.551]
d_{unb}	CCTST	-	-	0.847	[0.571, 1.123]	0.847	[0.571, 1.123]
	HCTA	0.008	[-0.453, 0.470]	0.721	[0.464, 0.977]	0.539	[0.317, 0.760]
	LSAT	-0.054	[-0.382, 0.272]	0.370	[0.238, 0.502]	0.307	[0.185, 0.428]
	All	-0.033	[-0.288, 0.223]	0.505	[0.397, 0.613]	0.424	[0.324, 0.523]

10.2.3 Gender

Experiment	Test	Gender	N	% of improved cases	Mean			SD	
					Pre-Test	Post-Test	Difference	Pre-Test	Post-Test
1	HCTA	F	8	75.0%	59.9	69.0		10.1	4.2
		M	6	66.7%	60.3	67.0	9.1	15.1	3.3
	LSAT	F	8	87.5%	10.8	12.4	6.7	3.1	2.3
		M	6	100.0%	13.3	16.8	1.6	2.7	3.9
2	CCTST	F	13	76.9%	16.2	20.8	3.5	3.0	3.2
		M	6	100.0%	19.3	22.8	4.5	2.3	3.4
	LSAT	F	12	66.7%	11.5	12.8	3.5	2.5	3.7
		M	5	80.0%	15.4	18.4	1.3	4.1	2.1
3	CCTST	F	1	0.0%	25.0	22.0	3.0	-	-
		M	9	77.8%	20.6	24.1	-3.0	4.3	3.7
	LSAT	F	1	100.0%	15.0	16.0	3.6	-	-
		M	11	63.6%	13.5	15.1	1.0	4.2	4.9
4	CCTST	F	1	100.0%	22.0	27.0	1.5	-	-
		M	6	83.3%	16.5	19.3	5.0	2.4	4.4
	LSAT	F	1	0.0%	20.0	17.0	2.8	-	-

		M	6	66.7%	13.3	14.3	-3.0	3.1	2.6
5	CCTST	F	2	50.0%	28.0	28.0	1.0	1.4	0.0
		M	11	90.9%	24.6	28.2	0.0	3.3	2.5
	LSAT	F	3	0.0%	16.0	15.7	3.5	2.6	2.3
		M	12	66.7%	16.8	18.7	-0.3	3.5	3.6
6	HCTA	F	18	72.2%	67.2	71.1	1.9	4.2	3.7
		M	20	90.0%	67.2	72.8	3.9	9.5	7.1
	LSAT	F	19	63.2%	13.2	14.4	5.6	4.0	3.0
		M	20	60.0%	15.4	16.9	1.2	4.0	4.5
7	HCTA	F	4	25.0%	67.5	68.0	1.5	2.4	2.2
		M	20	55.0%	72.2	72.2	0.5	3.8	6.0
	LSAT	F	4	50.0%	16.3	17.8	-0.1	2.5	3.1
		M	19	36.8%	16.4	15.8	1.5	4.0	4.3

Pre-post differences

Test	Gender	N	Pre-Post Mean Difference	95% CI	Female-Male Mean Difference (Standardized ES)	*p*-value
CCTST	F	17	3.59	[1.61, 5.57]	0.18 (0.06)	0.85
	M	32	3.41	[2.54, 5.85]		
HCTA	F	31	4.68	[2.42, 6.93]	1.66 (.21)	0.37
	M	46	3.02	[0.36, 5.68]		
LSAT	F	48	1.08	[0.34, 1.43]	-0.22 (-0.08)	0.67
	M	79	1.30	[0.62, 1.99]		

Pre-test differences

Test	Gender	N	Pre-score Mean	95% CI	Female-Male Mean Difference (Standardized ES)	*p*-value
CCTST	F	17	18.5	[16.0, 21.0]	-2.9 (-0.54)	0.15
	M	32	21.0	[19.4, 22.5]		
HCTA	F	31	65.6	[63.1, 68.1]	-2.5 (-0.34)	0.079
	M	46	68.5	[65.8, 71.2]		
LSAT	F	48	13.2	[12.0, 14.3]	-1.97 (-0.51)	0.006
	M	79	15.1	[14.3, 16.0]		

10.3 Trimmed Data

Since there are many ways of trimming data and there is not a standard one or two, the trimmed data below should be read with caution. The choice of trimming data technique

allows for a considerable amount of experimenter degrees of freedom. In exploring different approaches and criteria, we did not find way of trimming that made much of a difference to the conclusions of this project. Thus, we recommend that the untrimmed data be used, but present the trimmed data and analyses for readers interested in our rationale and our results.

10.3.1 The Rationale for examining trimmed data and the trimming technique

We were somewhat hesitant about presenting the untrimmed data only – not that the untrimmed data was inaccurate but that it may mislead, in some cases seriously mislead. While the untrimmed data is an accurate record of the subjects' results, some of those results are themselves implausible. Some subjects had amazing improvements – improvements so good that we do not believe them. For example, one obviously very bright subject went from a 32 pre-test on the HCTA (lower by 10 points than anyone else in that group) to a 73 on the post-test (the second highest in that group) – about 6 SD measured against the rest of the group's SD. We regard that result as highly suspect, flattering though it is to the project. From conversations, it was clear that this person, with a Masters degree in philosophy, initially had disliked the HCTA so much that they had not really tried and indeed had not bothered finishing the pre-test. But, on the post-test, that person understood the point of the experiment and tried much harder. Although we regard that subject's pre- and post- HCTA data as an accurate record of how well he/she did on the tests, we believe they should not count as evidence that the project's approach worked amazingly well for this subject. The pre-test data is not an accurate measure of her/his pre-course critical thinking ability. It should not be counted.

There are also subjects whose post-test results fell dramatically. For example, one particularly bright, enthusiastic student's HCTA score went down more than 2 SD, although his/her post-exam self-report was that he had become a better thinker; the teacher agrees. While they may have been wrong in thinking s/he had improved, it almost certain that s/he had not suffered serious cognitive damage. This is far from an isolated case.

Those who view increases of 6 SD and decreases of more than 2 SD as cognitively plausible can view the following data as the course's effects on the middle 90% of subjects. Those, who view such changes as cognitively implausible can view the trimmed data as a more accurate reading of the effects of the project.

So, we trimmed the top 5% and the bottom 5% (including any negative scores). Then the various experiments' effect sizes were re-calculated. Here, in more detail, is how we derived the trimmed data.

10.3.2 Trimming Methodology

The results of all of the experiments were combined by test (CCTST, HCTA, and LSAT) into three separate datasets, one for each test. They were ordered by the difference between each subjects' the pre- and post-test results. That is, for each test the results were ranked from largest improvement on the test to those who did worst on the post-test than they had done on the pre-test.

We decided upon using 5% as the threshold of exclusion in each direction, because (i) we deemed a total of 10% of the data as a tolerable amount of data to lose while getting rid of the most egregious cases and (ii) we had no a priori reasons to base a threshold on a particular value (e.g. ±10 score points).

We calculated the number of cases *N* that would correspond to 5% of the total cases in for each test in each direction. Then, for each dataset, we removed *N* cases with the largest positive pre-post test differences and *N* cases with the largest negative pre-post test differences. Hence, approximately 10% of the cases were removed from each test's combined dataset.

If the difference score of the *N*th case was not unique (and if the cases with the same difference score were not already marked to be removed), we decided that if the other cases were from a different experiment than the *N*th case, then those cases ought to be removed too (taking the total number of removed cases to more than *N*), but if the other cases were from the *same* experiment, then we would remove only up to the *N*th case and leave the other cases in the dataset:

	Number of cases (Untrimmed cases)	Number of cases (Trimmed cases)
CCTST	49	43
HCTA	76	68
LSAT	127	115

After the largest positive and negative differences had been removed, the data were then split up by experiment, and each experiment was analysed separately. Because the removal of the outliers was based on the combined data for each test, some individual experiments did not have any cases removed from them, while others had several. The next table shows the number of data points removed from each test in each experiment, with "*m*+" indicating that the top m scores were removed, and "*n*–" indicating that the bottom n scores were eliminated.

Experiment	Test	Number of Cases (Untrimmed Data)	Number of Cases (Trimmed Data)	Number of Cases Removed
	HCTA	14	10	3+; 1–
1	LSAT	14	12	2+; 0
	CCTST	19	16	2+; 1–
2	LSAT	17	16	1+; 0
	CCTST	10	9	0; 1–
3	LSAT	12	10	1+; 1–
	CCTST	7	6	0; 1–
4	LSAT	7	7	0; 0
	CCTST	13	12	1+; 0
5	LSAT	15	15	0; 0
	HCTA	38	37	0; 1–
6	LSAT	39	35	0; 0
	HCTA	24	21	0; 3–
7	LSAT	23	20	1+; 2–

As one would expect, for each individual experiment the untrimmed and trimmed data are sometimes distinctly different, generally in large part because of the asymmetry of the trimming procedure when applied to tests in specific experiments. E.g., Experiment 1's and Experiment 7's HCTA results. While the total standardized effect size for the entire project did not change much, the elimination of the 5% implausibly good and 5% implausibly bad changes in scores reduced the inter-group effect size differences.

For example, the raw score for Experiment I's pre-test SD on the HCTA was a very large 12, largely due to that bright subject who initially disliked the test and tested so poorly on the pre-test (32) and did so well on the post test (73) for a 41 point improvement. This was much larger than the SD of the other two HCTA experiments (7.0 and 4.0). After trimming, Experiment 1's HCTA pre-test SD was a more reasonable 6.4. Of course, removing such a 41 point raw score improvement, substantially lowered the average raw score improvement since 1/3 of that improvement was due to this one person's evolving attitude toward the HCTA. On the other hand, since final reports are in standardized effect sizes (Cohen's d_{unb}), it also substantially reduced the pre-test and pre/post average standard deviations. There is no general rule as to whether trimming outliers will increase or decrease the standardized effect size.

Unfortunately, our trimming procedure includes in the trimmed data those subjects who ran close to the ceiling at pre-testing (e.g., 23 out of a possible 25 on LSAT). Their data would be eliminated only if they had they done much worst on the post-test. Such people necessarily are not measured as having improved much when up against the ceiling and so lower the average raw effect size. Further, since their pre-test scores are outliers, they increase the pre-test SD and thereby lowered the standardized effect size in this way as well.

Also, our trimming procedure includes those subjects whose English was so poor that they could not understand a fair amount of what was being said. Despite our best efforts to dissuade them, some students with rather rudimentary English participated in some experiments. These were not ESL classes and so, unsurprisingly, several such subjects did poorly on the pre-test and post-test. For example, one foreign national with very poor English got a 6 on the LSAT pre-test and a 6 on the post-test; randomly answering the LSAT questions would give an expected score of 5. Such subjects are included in the trimmed data set, decreasing the average pre-post test difference and increasing the size of the pre-test standard deviation.

Because the statistical and methodological issues are so complex, we have not attempted to estimate the size of the ceiling effect or the not-fluent-in-English effect. Both would result in a subject having very similar pre and post-test results and in larger

standard deviations for both raw and standardized scores. We do not know how large these effects are, but are confident that they would not substantially change the results of this project.

10.3.3 Trimmed Data and Analyses

10.3.3.1 Trimmed Data (excluding 5% of most positive and 5% of most negative pre-post differences from each test result)

				Mean			SD		
Experiment	Test	N	% of improved cases	Pre-Test	Post-test	Difference	Pre-Test	Post-test	Average Pre/post
1	HCTA	10	70.0%	63.7	67.3	3.6	6.4	3.7	5.2
	LSAT	12	91.7%	11.6	13.3	1.7	3.3	2.9	3.1
2	CCTST	16	87.5%	17.5	21.5	4.0	3.3	2.6	3.0
	LSAT	16	68.8%	12.9	14.3	1.4	3.4	4.3	3.9
3	CCTST	9	77.8%	20.6	24.1	3.6	4.3	3.7	4.0
	LSAT	10	70.0%	14.4	15.9	1.5	3.8	4.7	4.2
4	CCTST	6	100.0%	17.5	21.7	4.2	3.3	4.1	3.7
	LSAT	7	57.1%	14.3	14.7	0.4	3.8	2.6	3.3
5	CCTST	12	83.3%	25.6	28.1	2.5	3.1	2.4	2.7
	LSAT	15	53.3%	16.6	18.1	1.5	3.2	3.5	3.4
6	HCTA	37	81.1%	67.6	71.9	4.3	7.1	5.8	6.5
	LSAT	35	65.7%	14.5	16.2	1.7	4.2	3.9	4.0
7	HCTA	21	57.1%	71.3	72.9	1.6	4.2	4.4	4.3
	LSAT	20	40.0%	16.1	16.0	-0.1	3.8	4.1	3.9

The most dramatic change in average post-pre test difference was in Experiment 1's HCTA results, with one person improving his/her score by 41 points (= about 6 SD for that group, when SD is calculated excluding that person). This was clearly due to a substantial change in attitude toward that test. In addition, because of our choice of trimming procedure, that Experiment 1 had three data points eliminated from the most largest positive change side and only 1 eliminated from the largest negative change for that test. This asymmetry obviously decreased the raw average effect size which, for the untrimmed data was 8.1 questions more questions answered correctly, whereas in the trimmed data it was 3.6.

In every case where there was a substantial difference in the average raw difference, there was such an asymmetry. In particular, with Experiment 7, the untrimmed raw

average difference was 0.0, the trimming removed the three lowest negative scores but none of the positive scores, and the trimmed average difference was 1.6. If the three top positive scores are also trimmed, the average effect difference is 0.5.

Pre-test differences between experimental groups

Experiment	Test	N	% of improved cases	Mean Pre-Test	SD Pre-Test
1	HCTA	14	71.4%	60.1	3.8
	LSAT	14	92.9%	11.9	3.8
2	CCTST	19	84.2%	17.2	3.3
	LSAT	17	70.6%	12.6	4.2
3	CCTST	10	70.0%	21.0	3.5
	LSAT	12	66.7%	13.7	4.7
4	CCTST	7	85.7%	17.3	5.0
	LSAT	7	57.1%	14.3	2.6
5	CCTST	13	84.6%	25.2	2.3
	LSAT	15	53.3%	16.6	3.5
6	HCTA	38	81.6%	67.2	5.7
	LSAT	39	61.5%	14.3	4.0
7	HCTA	24	50.0%	71.4	5.8
	LSAT	23	39.1%	16.3	4.1

Trimmed

Experiment	Test	N	% of improved cases	Mean Pre-Test	SD Pre-Test
1	HCTA	10	70.0%	63.7	6.4
	LSAT	12	91.7%	11.6	3.3
2	CCTST	16	87.5%	17.5	3.3
	LSAT	16	68.8%	12.9	3.4
3	CCTST	9	77.8%	20.6	4.3
	LSAT	10	70.0%	14.4	3.8
4	CCTST	6	100.0%	17.5	3.3
	LSAT	7	57.1%	14.3	3.8
5	CCTST	12	83.3%	25.6	3.1
	LSAT	15	53.3%	16.6	3.2
6	HCTA	37	81.1%	67.6	7.1
	LSAT	35	65.7%	14.5	4.2
7	HCTA	21	57.1%	71.3	4.2
	LSAT	20	40.0%	16.1	3.8

The untrimmed LSATs are close to the trimmed LSATs. The largest difference is the increased difference between Experiment 1 and Experiment 2; the difference between their untrimmed LSAT scores is 0.7 and their trimmed scores is 1.3. The only other notable change is the slight reversal of ranking between Experiments 3 and 4; because Experiment 3's pre-test trimmed mean was 0.7 higher than the untrimmed mean and the trimmed and untrimmed mean of Experiment 4 are the same, Experiment 3 has a slightly higher average number of questions answered correctly (0.1 questions) than Experiment 4.

Standardized Effect Sizes (trimmed data)

Experiment	Test	Using Pre-Test SD			Using Pre/post Av SD		
		d	d_{unb}	95% CI d_{unb}	*d*	d_{unb}	95% CI d_{unb}
1	HCTA	0.561	0.513	[0.056, 1.291]	0.687	0.628	[0.056, 1.291]
	LSAT	0.503	0.468	[0.178, 0.879]	0.535	0.498	[0.178, 0.879]
2	CCTST	1.217	1.155	[0.710, 1.958]	1.343	1.275	[0.710, 1.958]
	LSAT	0.404	0.383	[0.053, 0.648]	0.356	0.337	[0.053, 0.648]
3	CCTST	0.821	0.741	[0.217, 1.526]	0.887	0.801	[0.217, 1.526]
	LSAT	0.397	0.363	[-0.023, 0.716	0.354	0.324	[-0.023, 0.716]
4	CCTST	1.274	1.073	[0.360, 1.878]	1.126	0.949	[0.360, 1.878]
	LSAT	0.112	0.098	[-0.368, 0.622	0.132	0.115	[-0.368, 0.622]
5	CCTST	0.817	0.760	[0.273, 1.533]	0.916	0.852	[0.273, 1.533]
	LSAT	0.452	0.427	[-0.007, 0.863	0.435	0.411	[-0.007, 0.863]
6	HCTA	0.600	0.587	[0.395, 0.917]	0.659	0.645	[0.395, 0.917]
	LSAT	0.398	0.389	[0.185, 0.635]	0.412	0.403	[0.185, 0.635]
7	HCTA	0.376	0.362	[-0.070, 0.798	0.368	0.354	[-0.070, 0.798]
	LSAT	-0.01	-0.01	[-0.278, 0.252	-0.01	-0.01	[-0.278, 0.252]

10.3.4 Level of Scrutiny (trimmed data)

		Using Pre-Test SD					
		Extraordinary Scrutiny (Experiment 7)		Normal Scrutiny (Experiments 1 to 6)		All Expts	
Effect Size Measure	Test	Effect Size	95% CI	Wgtd Avg ES	95% CI	Wgtd Avg ES	95% CI
d_{unb}	CCTST	-	-	0.970	[0.747, 1.193]	0.970	[0.747, 1.193]
	HCTA	0.362	[-0.070, 0.798]	0.573	[0.378, 0.769]	0.535	[0.358, 0.712]
	LSAT	-0.013	[-0.278, 0.252]	0.374	[0.263, 0.485]	0.311	[0.209, 0.413]
	All	0.088	[-0.128, 0.304]	0.509	[0.421, 0.598]	0.448	[0.366, 0.531]

		Using Pre/post Av SD					
		Extraordinary Scrutiny (Experiment 7)		Normal Scrutiny (Experiments 1 to 6)		All Expts	
ES Measure	Test	Effect Size	95% CI	Wgtd Avg	95% CI	Wgtd Avg ES	95% CI
d_{unb}	CCTST	-	-	0.966	[0.744, 1.188]	0.966	[0.744, 1.188]
	HCTA	0.354	[-0.070, 0.798]	0.642	[0.443, 0.841]	0.589	[0.409, 0.768]
	LSAT	-0.012	[-0.278, 0.252]	0.368	[0.257, 0.480]	0.307	[0.205, 0.408]
	All	0.086	[-0.129, 0.302]	0.519	[0.430, 0.608]	0.456	[0.374, 0.538]

In none of the All Experiments results did the effect size go down, and in some it was substantially higher.

10.3.5 Comparison of experiment results by gender, trimmed data

Experiment	Test	Gender	N	% of improved cases	Mean			SD	
					Pre-Test	Post-Test	Difference	Pre-Test	Post-Test
1	HCTA	F	6	66.7%	63.7	68.5	4.8	4.3	7.6
		M	4	75.0%	63.8	65.5	1.8	1.7	5.1
	LSAT	F	8	87.5%	10.8	12.4	1.6	2.3	3.1
		M	4	100.0%	13.3	15.0	1.8	3.5	3.5
2	CCTST	F	10	80.0%	16.4	20.7	4.3	1.8	3.4
		M	6	100.0%	19.3	22.8	3.5	3.4	2.3
	LSAT	F	12	66.7%	11.5	12.8	1.3	3.7	2.5
		M	4	75.0%	17.0	18.8	1.8	2.2	2.3
3	CCTST	F	0	-	-	-	-	-	-
		M	9	77.8%	20.6	24.1	3.6	3.7	4.3
	LSAT	F	1	100.0%	15.0	16.0	1.0	-	-
		M	9	66.7%	14.3	15.9	1.6	4.9	4.0
4	CCTST	F	1	100.0%	22.0	27.0	5.0	-	-
		M	5	100.0%	16.6	20.6	4.0	3.5	2.7
	LSAT	F	1	0.0%	20.0	17.0	-3.0	-	-
		M	6	66.7%	13.3	14.3	1.0	2.6	3.1
5	CCTST	F	2	50.0%	28.0	28.0	0.0	0.0	1.4
		M	10	90.0%	25.1	28.1	3.0	2.6	3.1
	LSAT	F	3	0.0%	16.0	15.7	-0.3	2.3	2.6
		M	12	66.7%	16.8	18.7	1.9	3.6	3.5
6	HCTA	F	18	72.2%	67.2	71.1	3.9	3.7	4.2
		M	19	89.5%	67.9	72.6	4.6	7.3	9.2
	LSAT	F	17	64.7%	13.5	14.6	1.2	3.0	4.0
		M	18	66.7%	15.5	17.6	2.1	4.1	4.2
7	HCTA	F	4	25.0%	67.5	68.0	0.5	2.2	2.4
		M	17	64.7%	72.2	74.1	1.8	3.9	4.0
	LSAT	F	3	33.3%	16.7	16.3	-0.3	1.5	2.9
		M	17	41.2%	15.9	15.9	0.0	4.4	4.0

11 References

Abrami, P.C.; Bernard, R.M; Borokhovski, E.; Wade, A.; Surkes, M.A.; Tamim, R.; Zhang, D. 2008. "Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis." *Review of Educational Research* no. 78 (4):1102-1134.

Alvarez, C. 2007. *Does Philosophy Improve Reasoning Skills?* Masters Thesis, University of Melbourne, Melbourne, Australia.

Arum, Richard, and Josipa Roksa. 2011. *Academically adrift : limited learning on college campuses*. Chicago: University of Chicago Press.

Bessick, Sherlynn C. 2008. *Improved Critical Thinking Skills as a Result of Direct Instruction and their Relationship to Academic Achievement*, Education, Indiana University of Pennsylvania.

Butchart, S.; Forster, D.; Gold, I.; Bigelow, J.; Korb, K.; Oppy, G.; Serrenti, A. 2009. "Improving critical thinking using web based argument mapping exercises with automated feedback." *Australasian Journal of Educational Technology* no. 25 (2):268-291.

Carrington, M.; Chen, R.; Davies, M.; Kaur, J.; Neville, B. 2011. "The effectiveness of a single intervention of computer-aided argument mapping in a marketing and a financial accounting subject." *Higher Education Research and Development* no. 30 (3):387-403.

Carwie, Lisa. 2009. *The Effect of Computer-Supported Argument Mapping on the Critical Thinking Skills of Undergraduate Nursing Students*, Instructional Design and Development, University of South Alabama.

Cohen, Jacob. 1969. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale NJ: Lawrence Erlbaum Associates.

Crouch, C., and R. Mazur. 2001. "Peer Instruction: Ten Years of Experience and Results." *Americal Journal of Physics* no. 69:970-977.

Crouch, Catherine H, Jessica Watkins, Adam P Fagen, and Eric Mazur. 2007. "Peer instruction: Engaging students one-on-one, all at once." *Research-Based Reform of University Physics* no. 1 (1):40-95.

Cumming, Geoff. 2012. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York: Routledge.

Cumming, Geoff., and Sue Finch. 2005. "Inference by Eye." *American Psychologist* no. 60 (2):170-180.

Davies, M. 2011. "Concept Mapping, Mind Mapping, Argument Mapping: What are the Differences and Do They Matter?" *Higher Education* no. 62 (3):279-301.

Davies, M. 2012. "Can Universities Survive the Digital Revolution?" *Quadrant* no. 56 (12).

Donohue, Angela, Tim van Gelder, Geoff Cumming, and Melanie Bissett. 2002. Reason! Project Studies, 1999-2002. Parkville, Australia: Department of Philosophy, University of Melbourne.

Dunkin, Michael J, and Jennifer Barnes. 1986. "Research on teaching in higher education." *Handbook of research on teaching* no. 3:754-777.

Dwyer, C.P., M.J. Hogan, and I. Stewart. 2011. "The promotion of critical thinking skills through argument mapping." In *Critical Thinking*, edited by C.P. Horvath and J.M. Forte, 97-122. New York: Nova Science Publishers.

Dwyer, C.P., M.J. Hogan, and I. Stewart. 2012. "An evaluation of argument mapping as a method of enhancing critical thinking performance in e-learning environments." *Metacognition & Learning* no. 7:219-244.

Ericsson, K.A., R.T. Krampe, and C. Tesche-Römer. 1993. "The role of deliberate practice in the acquisition of expert performance." *Psychological Review* no. 100:363-406.

Facione, Peter A. 1990. The Delphi report. Committee on pre-college philosophy, American Philosophical Association.

Fisher, Alec. 1988. *The logic of real arguments*. Cambridge England ; New York: Cambridge University Press.

Giuliodori, Mauricio J, Heidi L Lujan, and Stephen E DiCarlo. 2009. "Student interaction characteristics during collaborative group testing." *Advances in physiology education* no. 33 (1):24-29.

Govier, Trudy. 1988. *A practical study of argument*. 2nd ed. Belmont, Calif.: Wadsworth Pub. Co.

Halpern, D.F. 2010. Halpern Critical Thinking Assessment. Vienna: Schuhfried.

Harrell, M. 2011. "Argument diagramming and critical thinking in introductory philosophy." *Higher Education Research and Development* no. 30 (3):371-385.

HERI. 2009. The American college teacher: National norms for 2007-2008. Los Angeles: Higher Education Research Institute, University of California.

Hitchcock, David. 2004. "The Effectiveness of Computer Assisted Instruction in Critical Thinking." *Informal Logic* no. 24 (3):183-217.

Kulik, C., J. Kulik, and R. Bangert-Drowns. 1990a. "Effectiveness of mastery learning programs: A meta-analysis." *Review of Educational Research* no. 60 (2):265-306.

Kulik, Chen-Lin C, James A Kulik, and Robert L Bangert-Drowns. 1990b. "Effectiveness of mastery learning programs: A meta-analysis." *Review of educational research* no. 60 (2):265-299.

Law School Admission Council. 2007. *The Next 10 Actual, Official LSAT PrepTests*. Newtown, PA: Law School Admission Council.

Lehman, Darrin R, Richard O Lempert, and Richard E Nisbett. 1988. "The effects of graduate training on reasoning: Formal discipline and thinking about everyday-life events." *American Psychologist* no. 43 (6):431.

Liu, Ou Lydia, Brent Bridgeman, and Rachel M. Adler. 2012. "Measuring Learning Outcomes in Higher Education: Motivation Matters." *Educational Researcher* no. 41 (9):352-362. doi: 10.3102/0013189x12459679.

Macagno, Fabrizio, Chris Reed, and Douglas Walton. 2007. "Argument Diagramming in Logic, Artificial Intelligence, and Law " *Knowledge Engineering Review* no. 22 (1):87-109.

Mazur, Eric. 1997. *Peer instruction: a user's manual*, *Prentice Hall series in educational innovation*. Upper Saddle River, N.J.: Prentice Hall.

McCoy, J., Spurrett, D. unpublished. "Developing critical thinking skills in South African undergraduates through computer-supported argument mapping."

McMillan, James H. 1987. "Enhancing College Students' Critical Thinking: A Review of Studies." *Research in Higher Education* no. 26 (1):3-29.

Mencken, H. L. 1997. *Minority report : H.L. Mencken's notebooks*. Johns Hopkins paperbacks ed, *Maryland paperback bookshelf*. Baltimore, Md.: Johns Hopkins University Press.

Mercier, Hugo, and Dan Sperber. 2010. "Why Do Humans Reason? Arguments for an Argumentative Theory." *Behavioral and Brain Sciences, Vol. 34, No. 2, pp. 57-74, 2011.*

Pascarella, Ernest, and Patrick Terenzini. 2005. How college affects students: Findings and insights from twenty years of research. Volume 2. A third decade of research. San Francisco: Jossey-Bass.

Rider, Yanna, and Neil Thomason. 2008. "Cognitive and Pedagogical Benefits of Argument Mapping: L.A.M.P. Guides the Way to Better Thinking." In *Knowledge Cartography: Software Tools and Mapping Techniques*, edited by A. Okada, S. Buckingham Shum and T. Sherborne, 113-130. Springer.

Smith, Michelle K, William B Wood, Wendy K Adams, Carl Wieman, Jennifer K Knight, Nancy Guild, and Tin Tin Su. 2009. "Why peer discussion improves student performance on in-class concept questions." *Science* no. 323 (5910):122-124.

Spencer, Ken. 1991. "Modes, media and methods: the search for educational effectiveness." *British Journal of Educational Technology* no. 22 (1):12-22.

ter Berg, T. unpublished.

Twardy, Charles. 2004. "Argument maps improve critical thinking." *Teaching Philosophy* no. 27 (2):95–116.

van Gelder, T.J. 2007. "The Rationale for Rationale™ ." *Law, Probability and Risk* no. 6 (23-42).

van Gelder, T.J. 2013. "Argument Mapping." In *Encyclopedia of the Mind*, edited by H. Pashler. Thousand Oaks, CA: SAGE.

van Gelder, Tim, Melanie Bissett, and Geoff Cumming. 2004. "Cultivating Expertise in Informal Reasoning." *Canadian Journal of Experimental Psychology* no. 58 (2):142-152. doi: 10.1037/h0085794.

Walton, Douglas. 2000. "Problems and useful techniques: My experiences in teaching courses in argumentation, informal logic and critical thinking." *Informal Logic* no. 20 (2).

Whitney, C.R. 2005. *The WMD Mirage: Iraq's Decade of Deception and America's False Premise for War : Featuring the Report to the President from the Commission on the Intelligence Capabilities of the United States Regarding Weapons of Mass Destruction*: PublicAffairs.