# Argument Maps Improve Critical Thinking

CHARLES TWARDY
*Monash University*

I've been skeptical about claims for various approaches to teaching critical thinking, including those for argument maps coming from nearby University of Melbourne. Indeed, confident in our skepticism, we at Monash Philosophy accepted a challenge to compare our methods with theirs on pre- and post-test gains on the California Critical Thinking Skills Test (CCTST) developed by Peter Facione (1990, 1992). The Monash students did a bit better than theirs on the pre-test, raising our hopes. But when Melbourne University's post-test results showed far higher performance gains, I thought their method worth a closer look. In short, I think computer-based argument mapping is the key to Melbourne's success, and that it should feature centrally in any critical thinking course. In fact, it is useful quite beyond critical thinking courses. I will describe the method, my experiences with it, and the results.

## *Introduction*

Being soundly trounced on improvement scores only convinced us that *something* over there was working. I thought the approach was fine, and argument maps were certainly helpful, but suspected that most of the gains were due to founder effect: the Melbourne teacher (Tim van Gelder) developed the Reason! approach to critical thinking, the Reason!Able software,[1] and the course materials used at Melbourne. In addition, he had been the sole teacher since they adopted this approach, so it was very likely that a lot of the result was because of his founder's enthusiasm, a well-known effect in education. Tim also wanted to see how much was founder effect.

So we replaced Tim with a relatively inexperienced postdoc from an outside university. Me. Tim got more time to develop materials and do research, and Monash got a "spy" to learn more about the Reason! method and why it worked so well.

 0145-5788 pp. 95–116

The quick answer is that with my novice teaching skills thrown for the first time into an unfamiliar curriculum, student improvement on the CCTST was still 90 percent of Tim's previous gains, and about three times the gross gain typical of other endeavors. The students improved noticeably despite what I must candidly regard as a poorly taught first-half of the subject.

I'll report on my experiences teaching the Reason! method, including common student difficulties and misconceptions, and benefits and drawbacks of the course materials and the Reason!Able software. I'll start with some results, to show that there really is a difference, and move on to discuss the method in detail. However, as a quick introduction, I should say that the Reason! method consists mainly of intensive practice on computer-supported argument mapping in a highly structured setting.

## *Numbers and Results*

This section highlights some of the results presented in Donohue et al. (2002). With a first-round entrance cutoff at the 93rd percentile of entrance score for students in Arts, Melbourne University is roughly equivalent to an elite private university in the United States. (There are no private secular universities in Australia.) It takes only the top 5–7 percent. Monash University, by comparison, has a first-round cutoff of about 88 for Arts.

I lectured in a first-year introduction to critical thinking with about 135 students. I ran one of that class's nine tutorials ("discussion sections" in the U.S.). Mine had twelve students; Tim had two tutorials and the remainder were taken by philosophy graduate or honors students. I also ran a tutorial in a third-year History and Philosophy of Science class that used the Reason!Able software and argument maps to guide discussions. All tutorials used a computer-equipped classroom with an interactive whiteboard.

In the first-year introduction, we measured the average gain (improvement) in performance on the CCTST. At the beginning of the semester, we randomly gave half the students one version of the CCTST, and half another version, and at the end of the semester, gave them the other version. Raw gains would apply only to this test, but we want a common measure we can use across tests. On common standardized effect size measures the average gain in terms of the variability in the scores: divide the average gain by that variability.

The usual measure of variability is the standard deviation. The standard deviation is calculated so that a central slice of the bell-shaped curve one standard deviation to either side of the mean contains about 68 percent of the scores. Two standard deviations contain about 95 percent, and three standard deviations contain about 99 percent. So a

tall, narrow bell will have a small standard deviation, while a wide flat bell will have a large standard deviation, but no matter the width of the bell, one standard deviation will capture 68 percent of the scores.

If we divide the gain by the standard deviation in raw test scores, we get a standardized effect size called Cohen's d (Cohen, 1988). It measures the number of standard deviations' improvement, and so gives a measure that takes inherent variability into account.[2]

The result was that my first-year class had a gain of about 0.72 standard deviations, down slightly from Tim's average of about 0.8. On the other hand, an effect size of 0.7 compares favorably with the 0.3 typical of other endeavors, and the 0.5 in a critical-thinking subject at McMaster University. (van Gelder, 2001; Donohue et al. 2002; Hitchcock, 2003) More results are shown in Figure 1. Because the confidence intervals do not overlap the line at 0.34, we know that the Reason! results are clearly above the expected gain for one whole year in university, and so represent a real improvement.[3] The other subjects are not clearly adding anything to this baseline "maturation" effect. Indeed, our Monash CT subject is up with the rest, despite being only six weeks of lecture time in the elapsed twelve weeks.

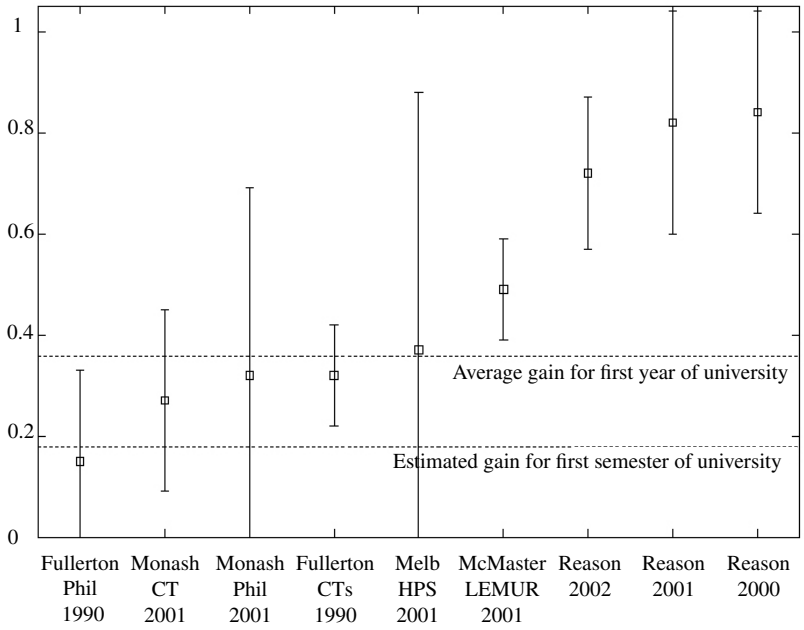



Figure 1. Effect sizes in standard deviations (Cohen's d) on the CCTST for various subjects.

To put the numbers in further perspective, consider general educational testing results. A full three years of undergraduate education has an effect size of about 0.55 on the CCTST, with about 0.34 occurring in the first year. (Pascarella and Terenzini, forthcoming).

Or consider the SAT college entrance exam, which has had a standard deviation of about 115 in recent years. (Educational Statistics, 1997) Therefore an effect size of 0.8 would correspond to an 90-point increase on SAT score. In contrast, major (and expensive) test-prep companies claim to provide a 120–140-point average increase but according to a College Board study[4] they actually provide an average increase of only 29 points on the verbal and 40 points on the math. (Powers and Rock, 1998) And there is no reason to believe SAT coaching transfers to anything other than the SAT (because it often consists largely of test-taking strategy specifically for the SAT, plus some crash review), while there is good reason to believe that Reason!-based critical thinking does transfer (perhaps even to verbal SAT scores, because they rely in part on logical thinking and analysis similar to what is found on the CCTST).

But anyone can get large effects by teaching to the test. However, a CCTST question is typically something like:

> When Mark is late for dinner, both Susan and Andrew eat out. Which of the following must be true?
> 1. Mark is late, Susan eats out, Andrew eats out
> 2. Mark is not late, Susan eats in, Andrew eats in

In contrast, as we will see below, an argument mapping question is typically a passage of text and a request to produce and evaluate a map.

For comparison, interactive methods of physics education such as those of Eric Mazur (Mazur, 1997) get a stronger effect. In a survey of 62 introductory physics courses (6,500 students) using the standard Force Concepts Inventory pre/post test, Richard Hake (Hake, 1998, 1999) found that interactive instruction yielded a gain of 2.18(!!) standard deviations, versus 0.88 for traditional methods.[5] On the other hand, they are teaching not just a skill but domain knowledge, and in class they frequently discuss conceptual questions like those which would appear on the FCI.[6]

## *Limitations of the Empirical Study*

The previous section aims to show that the Reason! method is much more effective than other methods at teaching critical thinking. There are two main avenues of criticism: results on the CCTST may not be the right measure, and there may be problems with this kind of study design and analysis.

The CCTST is not perfect, but there is good reason to think it is relevant. Nevertheless, it is good to know that in the years where other tests were also given, independent assessments of written essays similarly favored the Reason! approach, as did the Watson-Glaser Critical Thinking Appraisal (Donohue et al., 2002). Furthermore, the fact that the class exercises are so far removed from CCTST questions indicates that students are learning something general, not test-specific.

The pre-test / post-test is far better than merely reporting post-test results, but it lacks explicit control groups. All we have are comparisons between classes which differ on material, approach, lecturer, class size, class aptitude and motivation, and any number of other variables. Monash is now attempting a more controlled study, but it will take several years. But we can already evaluate alternate explanations of the existing data.

For example, we could just be looking at maturation. If we only had data for our own classes, that would be quite plausible. However, if it were maturation, then we should see similar effects in other subjects, but as the last section discussed, we don't. Furthermore, other studies have estimated the expected gain due to maturation, and our gain is far above that.

By randomly assigning test versions, we control for merely getting an easier version of the CCTST on the post-test. Still, students should do better the second time merely because they gain practice with that kind of test. However, that does not explain the difference between the Reason! method and others.

In the rest of the paper I shall discuss various features of the method, and present an argument that efficient argument mapping is a key element.

## *Argument Analysis with Maps*

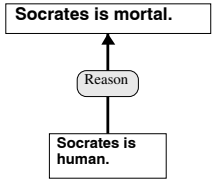
The most distinctive feature of the Reason! method is the use of argument maps. Argument mapping was proposed at least as early as Scriven (1976), but revising pen-and-paper maps just isn't practical, and the method never caught on. However, computer software takes care of the layout and redrawing for you, making it easy to rearrange and rethink the argument and add new considerations. The semester revolves around argument mapping, and consists largely in argument analysis. So what is an argument map?

### Argument Maps

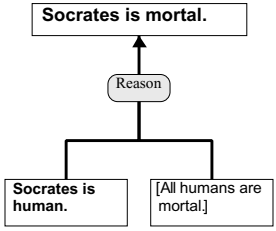
An argument map is a two-dimensional representation of argument structure. It is usually a box-and-arrows diagram which resembles a

tree. (In fact, it has to be slightly more complex than a tree or it would conflate separate arguments with helping premises.)

The boxes are claims, which are arranged so that some are reasons for believing (or disbelieving) others. At one extreme is the final conclusion, supported (and opposed) by its reasons and objections. At the other extreme are the unsupported claims you take as basic. For example, to map the explicit part of the argument, "Socrates is mortal because Socrates is human," we would draw:

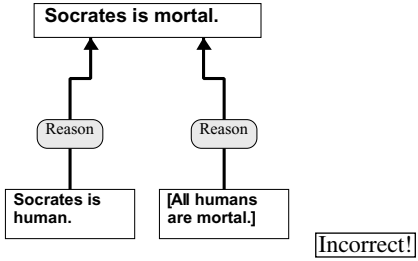



Of course, this argument is not valid as it stands: there is a missing (or enthymematic) premise, "All humans are mortal." By convention, when mapping someone else's argument, we put brackets around any claims we have had to supply. So the argument map becomes:




A reason is a collection of claims which help each other, rather than a single claim. For example, in the argument map above, we argue that "Socrates is mortal" with the two claims "All humans are mortal" and "Socrates is human." But those two claims together are one reason. One claim helps the other, and together they yield the conclusion.

The most common error students make is to confuse such multi-premise reasons with multiple reasons. For example, they might incorrectly represent the Socrates argument like this:



But these are not independent reasons. "Socrates is human" does not yield the conclusion except with the help of "All humans are mortal." Yet by drawing them as independent reasons, we assert that even should one reason fail, we still have another to support our argument. In this case, that is false.

There are many versions of argument maps around, and they do not all share the same conventions. However, as you have just seen, the Reason! maps use the convention that separate arguments have separate lines of support, with separate arrows. Claims that help each other, and so together make up a single reason, fall under a single arrow.

## More on the Reason! Approach

The Reason! approach to teaching critical thinking does not include symbolic logic, formal probability, causal reasoning, or other technical specialties. Instead students spend most of the semester analyzing arguments, and some on producing them. I admit this sounds unimaginative. After all it would seem that informal logic should certainly include something like probabilistic reasoning and causal inference. As a discipline, it certainly does, and causal inference is my own area of expertise. But as a first-year subject, it should not.

The literature suggests that critical thinking students overwhelmingly do not learn basic argument analysis. For example, Doug Walton has (in)famously written:

> I wish I could say that I had a method or technique [for teaching Introductory Critical Thinking] 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)

Teaching extra flourishes does not mean they are being learned, and every such addition takes away time that could be used to help students master the fundamental critical thinking skill of argument analysis. Indeed, most other critical thinking skills centrally depend upon argument analysis: a student must be able to identify claims and lines of reasoning before they can possibly engage them critically, and then must be able to assess evidence and support "in the wild" of informal arguments.

As we have already seen, argument maps force us to make a distinction we normally would not even consider: do two claims form part of a single reason, or are they parts of separate reasons? Even very bright students get it wrong surprisingly often.

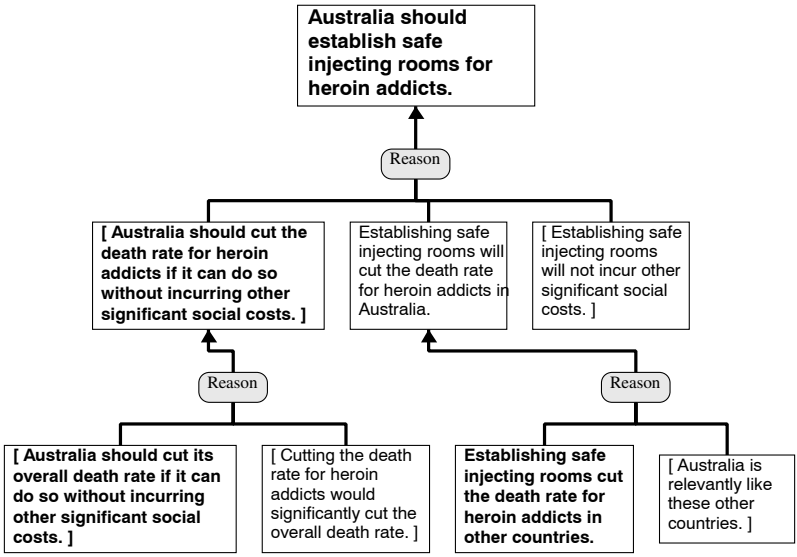
Early in the semester the mistake is entirely expected: students have to learn the convention that separate branches should be fully independent reasons for believing the conclusion. However, although students do get much better at this, it continues to be a problem. In

normal informal reasoning, we just do not make this distinction, even though it is crucial for understanding argument structure. Fortunately, there are some largely mechanical rules Neil Thomason devised for checking argument structure. They are simple, quick to apply, and incredibly helpful. Properly used, with an argument map in front of you, they help overcome many mistakes made in representing argument structure.

## Dr. Neil's Helping Rules

> RABBIT RULE.
> "You can't conclude something about rabbits if you haven't been talking about rabbits."
> Therefore:
> "Every meaningful term in the conclusion must appear at least once in each reason."

This applies to every independent reason separately. The reason cannot get you to the conclusion "Socrates is mortal" unless in its premises you have both "Socrates" and "mortal." Since most arguments are enthymematic,[7] the Rabbit Rule prompts students to find the hidden premises and state them explicitly. For example, an editorial might argue that Australia should adopt safe injection rooms (and related programs) for heroin addicts because such programs reduce heroin deaths. Plausible hidden premises are that these programs do not increase other deaths, nor incur other prohibitive costs, and that Australia should adopt policies which reduce overall deaths without incurring other prohibitive costs.



This rule is simple, clear, and easy to apply, and all the students know it cold. Yet on any given argument map they produce, for the first half of the semester they probably fail to use it, and in a significant way. Usually they will fail to articulate the hidden premise doing the most work ("All humans are mortal."). Yet any twelve-year-old can see if an argument map follows the Rabbit Rule: all you have to do is circle matching words. Nevertheless, bright university students often have difficulty, probably because they let their own mind fill in the missing premises and definitions.

In our own heads we perform so many semantic inferences that it is easy to let a Rabbit violation slip past, though it is trivially easy to see in someone else's argument map, or when asked to look at our own again.

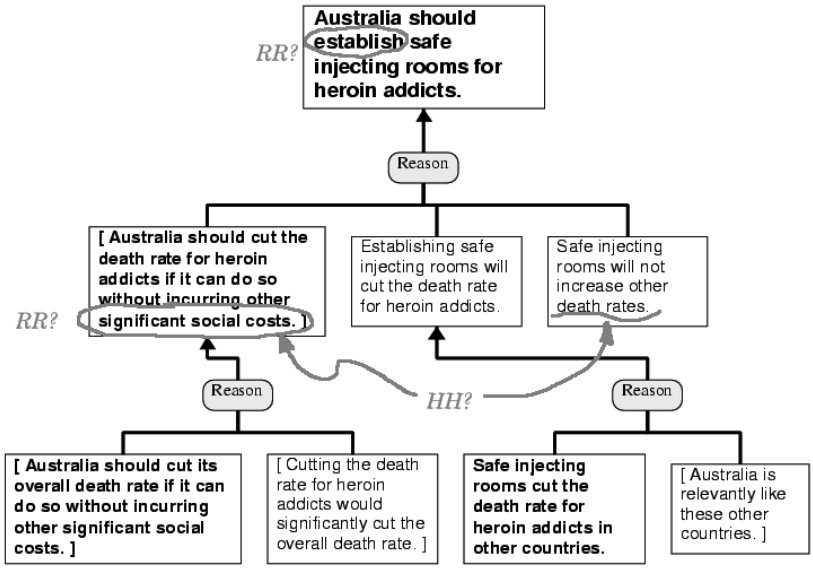
Of course, merely having all the meaningful terms in your premises does not make it a good argument. For example, you might have "Socrates is a man" and "All elephants are mortal." We've covered "Socrates" and "mortal," but the premises do not work together.

> HOLDING HANDS.
> "We can't be connected if we're not holding hands."
> Therefore:
> "Every meaningful term in one premise of a reason must appear at least once in another premise of *that* reason, or in the conclusion."

The premises "hold hands" by sharing terms. So "Socrates is human" shares "human" with "All humans are mortal."[8] Just making the terms link up in this simple way helps prevent a lot of irrelevant "reasons" that fail to get you to the conclusion even though all the "ingredients" seem to be there.

When grading assignments it is sufficient to write "RR" for "Rabbit Rule" or "HH" for "Holding Hands," and students quickly work out the problem for themselves. After a few assignments, most of the red arrows on the example would be unnecessary.

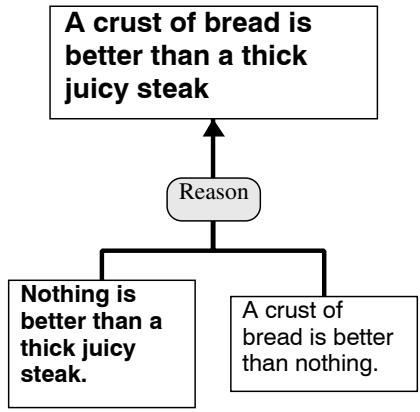
For example, we might have received the following map for the injection rooms argument, and commented on it this way:




But while RR and HH help assure that each reason is sufficient for its conclusion, they are still heuristic rules.

## Limits of Dr. Neil's Helping Rules

The Helping Rules are merely syntactic, and so are not sufficient to guarantee a good argument. For example, you can satisfy them all in the Socrates example by offering a single reason "Socrates is mortal." You could get around that with a principle (also suggested by Neil) like "Walking in a circle gets us nowhere," but for any such syntactic rule set, you can probably construct a counterexample. My favorite is the following, which I first heard from my wife:



Equivocation errors are necessarily errors of semantics, so mechanical checks cannot find them. So rather than complicating our rule set, it is much more instructive to present such examples in class, after students have become proficient at applying the rules. Such examples helps the students really see that the rules are merely quick helpful guides, and to intuit what we're really after.

## Relaxing Dr. Neil's Helping Rules

Finally, while I am fairly strict with RR and HH early in the semester, there are a lot of cases where you are either adding a lot of trivial premises like "A hare is a rabbit," or fighting with simple facts of English grammar like tenses and plurals. This is OK to learn and practice the idea, and to get students to see what good, tight, deductive arguments look and feel like. But drop these more pedantic applications as soon as possible. Later in the semester (or in the third-year class), you are trying to handle big arguments like inferences from experiments to public policy, and too strict an adherence to syntactic rules overly complicates and slows down the mapping.

Part way through the semester you can say, "OK, from now on, if you're sure that it's a trivial matter, and the meaning is clear, don't worry too much about adhering strictly to the RR." Inevitably students then start asking a lot of questions about where to draw the line. The basic rule is, "if you're unsure, work it out." But partly what they are asking is, "when will you take off points for not following RR"? My answer is, "If you make a real mistake that the RR would have caught." This seems fair enough to them and to me.

## Quality Practice

Because students are to acquire a skill, the structure of the Reason! subject more closely resembles a laboratory or mathematics class than a traditional philosophy class. Like learning algebra, critical thinking requires practice rather than merely study or thought. I believe that lecturing by itself does little good.

Consequently, the weekly discussion sections are hands-on tutorials. Melbourne University has some exceptionally nice electronic classrooms, so students can work in groups at computer stations, and then in turn project their maps up in front of the whole tutorial for discussion. Each module has a set of practice exercises with answers, and then a set of exercises for homework. Since there are four modules, students have to turn in homework about every three weeks. That homework is typically six questions, nominally requiring five to ten hours of work,[9] not counting the time spent on the practice exercises. (Neil requires an argument map each week, handed in at the beginning of the tutorial.)

Each module also has an in-class exam which happens after the final lecture for that module. The goal is both to provide plenty of practice and plenty of feedback.

In particular, Quality Practice is, according to van Gelder (2001), practice which is:

- Motivated—the student should be deliberately practicing in order to improve skills
- Guided—the student should have some way of knowing what to do next
- Scaffolded—particularly in early stages, there should be structures preventing inappropriate activity
- Graduated—tasks should gradually increase in complexity
- [Providing] Feedback—the student should have some way of telling whether a particular activity was successful or appropriate

The bottleneck, unfortunately, is feedback.

The grading burden is already quite high, and as a result we were able to grade only two (randomly chosen)[10] problems from each homework set. I set up the schedule so that students turned in homework on Friday of the last week for a module, and we had the graded homework returned by Monday at noon. All the tutors and I held various office hours Monday and Tuesday, and we gave the exam Tuesday afternoon during class. We then graded the exams on Wednesday and discussed the exam on Thursday. That made for an intense week of grading (and exam-writing!), but at least provided rapid feedback for the students. We decided that rapid feedback was very important, because even though the online materials provide a great deal of feedback in one sense (especially the practice exercises), students the previous year had complained that they did not have enough, meaning that they were not getting evaluations of their homework. The fact is that most students do not really test themselves until they have to turn assignments in for a grade.[11]

Ideally, the students would get feedback on ALL the homework problems, and would do two of them per week, gradually moving from simpler to more complex. But the grading burden was already hitting the limits of what was practical and legal to require of tutors. We did get faster as the semester went on, and most of us did find that grading on paper was faster than grading electronic assignments, but efficient grading is the ultimate limit on how much students can learn. I think if we could get them rapid feedback on graded assignments every week, students would break the 1.0 standard-deviation improvement barrier. Two per student per week may not be possible without some sort of automated grading system, or veteran tutors grading problems they have graded many times in the past. One per student may well be possible.

## Author's Intent

Argument mapping can be so demanding that it is very easy to saddle the author with an uncharitable interpretation when filling in hidden premises. The usual error is to make the hidden premise too strong. For example, when the third-years were mapping arguments for cognitive bias, they would often write the helping premise, "People see what they expect." Indeed, the author had said this at some point, and the map went through just fine. The problem is that the author cannot possibly have meant something so strong—it is a description of delusion, not ordinary cognitive bias! But a more nuanced claim like, "People interpret ambiguous evidence to favor their own position" makes the argument map more complex. Nevertheless, it is crucial to make such claims explicit, and to do so in a way that captures author's intent. In lecture I talked about the Principle of Charity and the deleterious consequences of ignoring it, in particular the ubiquitous tendency for academic debates to drag on by attacking straw-man positions.

## Evaluating Arguments

So far I have talked only about argument structure. Once you have reconstructed a written argument, or developed your own argument, you evaluate it. The evaluation system built in to the Reason!Able software is rudimentary, but it serves the purpose. It does almost nothing for you, but provides helpful hints and questions, and allows you to display your evaluation in color. A strong supporting argument is bright green. Weaker ones become paler. Likewise a strong opposing argument is bright red.

The software requires students to see that evaluation is a recursive process, and to appreciate that the recursion must in practice bottom-out somewhere. In Reason!Able, the lowest-level reasons are given "grounds" rather than further reasons. Grounds for belief include: necessary truth, expert opinion, eyewitness testimony, common knowledge, personal experience, and considered opinion. The Reason!Able program has cute icons for each ground, along with rules for application. For example, the help screens say that to count as common knowledge a claim must be:

1. widely believed in the community
2. well-founded—*i.e.*, the community's belief must itself be solidly based
3. free from any serious dispute

Similarly, to be expert opinion, a claim must be:

1. the opinion of a genuine expert who is sincere, trustworthy, and has no conflict of interest
2. within that expert's area of expertise
3. free from any serious dispute among the relevant experts

If the argument map is done properly, then all co-premises of a reason are necessary for the reason to be any good. So if one of them fails, that reason must give no support. For example, suppose we judge the third premise in the heroin argument ("Establishing safe injecting rooms will not incur other significant social costs.") as having no grounds. Then, because there are no other arguments for the main conclusion, we decide that we have in this argument map no good reason to believe the conclusion. In practice, we would evaluate the third premise by considering various supporting and opposing argument.

Although the Reason!Able software separates construction and evaluation modes, in practice building a good argument requires frequently switching modes, at least in your head. Quite often you only notice gaps during evaluation.

## Maps, Practice, Structure

Having already seen that the Reason! results are repeatable, even with a novice instructor, the three major components of the method are argument mapping, quality practice, and scaffolded, structured learning.

Practice is clearly important: argument mapping without practice would not much improve critical thinking. Likewise, clear structure and expectations will improve any subject. Nevertheless, I suspect that argument mapping is the key—that if a traditional critical-thinking class matched the amount of practice and graduated structure of the Reason! method, it would not show the same level of improvement. I could be wrong, but I see at least three reasons to emphasize mapping.

First, mapping is the most distinct feature of the Reason! method. There are a lot of traditional critical thinking classes with varying levels of practice and structure. I fully expect that some of them—especially those based on formal logic—have matched or nearly matched our practice hours and clear, graduated goals. So while important in any method, I think they are not really unique to the Reason! method.

Second, argument mapping precisely targets argument structure, and students really need help learning this. Prose does not force students to know the structure, but maps do. It helps many students see that there is structure.

Third, critical thinking as generally conceived, and as measured by the CCTST, consists of tacking arguments "in the wild." Argument mapping is a general skill which can be applied to all kinds of arguments, and so its improvements will help on all arguments, rather than just specialized subtypes. Similarly, argument mapping is always applied to "wild" arguments, so students do not have to translate back and forth between formal systems, and so what practice they have applies directly.

It is hard to separate the role of argument mapping from that of the software. But in general, merely adding computers does not tend to improve educational methods, making it unlikely that it is the software *per se*, as opposed to mapping itself. Furthermore, the key role of the software is quite clearly to make mapping *easy* or even *practical*. To that extent, the software is not an "extra" component, though apparently large sheets of paper and sticky notes make a decent substitute.

## *Student Difficulties*

If you still think a whole semester is too much to spend on argument analysis, just pick any two-paragraph argument and attempt to map it. Then see if you have followed RR and HH. Are you sure that you have captured the author's intended structure? By the time you are sure of yourself, you have probably spent thirty minutes or more producing your map, even though in some sense you understood the argument fine beforehand. For some reason, no one believes this until they try.

It took me over an hour to map Hume's argument against necessary connection, even though I knew it and felt I understood it quite well. But I had never asked myself whether the structure was a chain of reasons, a bunch of parallel reasons working together, or something in between: a more tree-shaped structure. In ordinary listening and thinking, we fill in all sorts of inferences without noticing what they are. When mapping, we must decide specifically whether the claim was a co-premise for the present reason, or part of a separate reason.

In order to really appreciate this, you should try it now. But just as an example, Figure 2 is a portion of that argument which, once mapped, reveals itself to be mostly a chain. The selection of quotations from Hume's Treatise is due in part to M. J. Garcia-Encinas (Garcia-Encinas, 2003 forthcoming).

Figure 2 shows a simple chain of reasoning for a simple, clear conclusion: we have no such impression. It also pinpoints the helping premises doing all the work, most notoriously, "Whatever we can conceive is possible," which ties the discussion to logical possibility when we should probably be talking physical possibility.[12]

Even so, a quick check shows that this does not strictly follow RR and HH. I happen to think it is reasonable to expect readers to equate "There is no" with "does not exist," and to parse placeholders like "something," "A," and "B." But I wouldn't use this map in class early in the semester.

### Toying With Arguments

Students tended to toy with arguments rather than engaging them. Consequently, their objections were often throw-away scenarios: "Maybe

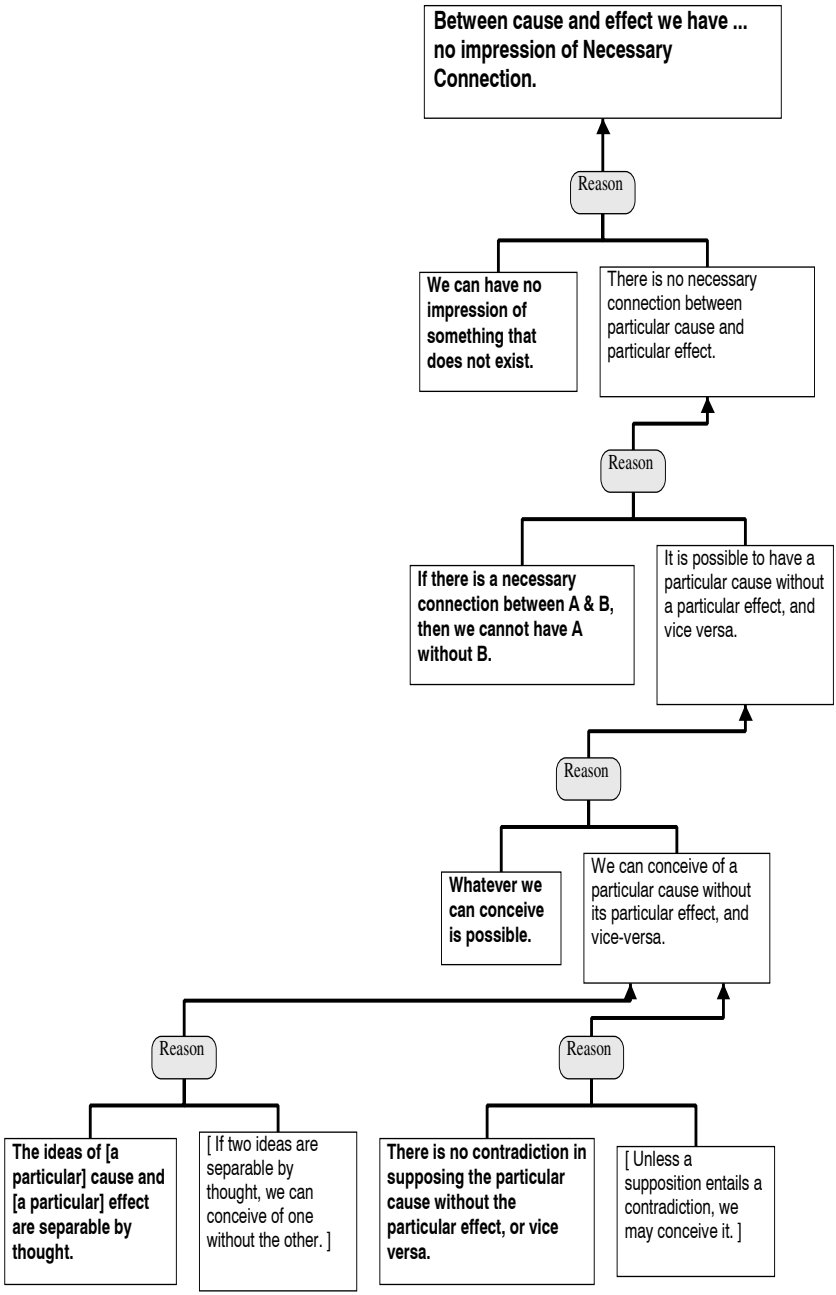



Figure 2. Argument map of part of Hume's argument against necessary connection. This part argues that we can have no impression of necessary connection. "Impression" is a technical term for Hume

this could happen," without suggesting why, or how, or how likely that was. This is like explaining your lab results as "measurement error" without analyzing whether your likely measurement errors could possibly have explained your discrepancies.

Students were either just being lazy and throwing out minimal-effort responses, or they were learning to play a game of Boxes-and-Arrows based on our rules, without thinking about what they were doing. In part this happened as a natural consequence of focusing on such simple arguments early in the semester. By the end of the semester I saw much less of this, partly because I addressed it specifically in lecture, and partly because we were dealing with more complex arguments.

### Multi-Premise Reasons Versus Multiple Reasons

I've already mentioned that this is the most common error in making argument maps, either in mapping other people's arguments, or in constructing your own.

But I want to reiterate that while the error persists, students did improve. By the second module, I wrote to the tutors list:

> Students are still unsure when something is a separate reason and when something is a helping premise. But I know that now mostly because they are ASKING. Their finished maps don't make this confusion as often. Progress!

### Just What is a Reason in the Map?

In part this was my mistake: I delved right into mapping and structure without stopping to make clear what our reasons were. While the maps presume that reasons are evidence for a conclusion, in real life reasons are just as likely to be explanations. I noticed this most dramatically in the third-year (senior) tutorial for "Science, Life, and Mind" (SLM) which used argument maps but did not spend three weeks introducing them.

Students in SLM kept putting theories down as reasons for the facts which were being used to support the theory! For example, one chapter offered several experiments to demonstrate that people often act quite inconsistently with their stated beliefs. In one experiment, Richard LePierre and two Chinese friends visited about 250 restaurants and hotels across the southern U.S. in 1930, and received cordial and even exceedingly friendly service in all but one. Six months later, he sent out surveys to each establishment asking whether they would serve Chinese customers. Almost all of the 130 responses were negative! The author provided the results of the experiment as reason to believe his relatively speculative theory of inconsistency.

Nevertheless, most of the SLM students created maps which offered the theory of inconsistency as a reason to believe the results of the experiment. While this kind of reasoning does happen sometimes, it

was clearly not the case here. This theory was offered as an explanation of the facts, not a reason to believe them! We believe the facts because of experimental results plus assumptions about representativeness and generalizability (to say nothing of the honesty of the experimenter).

The students agreed with my diagnosis that they were thinking of reasons as explanations, and most of them were able to switch within another map or two. I think the maps helped to point out a real ambiguity in student thinking, that would have been much harder to understand if they were merely arguing in prose. Because in prose, "reason" really does have all these meanings.

## Lousy Maps Versus Lousy Thinking

In both classes I often saw maps that were just unaccountably bad, at least assuming the students spent any time at all on them! I continue to believe that students were not having nearly as much trouble understanding the argument as their maps indicated. I suspect that had I asked students to make a map of an argument, and also to write up what they took the argument to be, their essay would show greater understanding. But that may not be the case. Tim has done plenty of in-class exercises with groups of all ages where the participants were asked to identify the main point and argument in short letters to the editor (three sentences!). Here is an example from the homework, an entire letter responding to an academic's claim that Shakespeare was actually an Italian.

> SO Shakespeare was an Italian because almost half of his plays are set in Italy. Almost all of Isaac Asimov's novels are set in outer space—does that mean that he was a Martian?
>
> Graham Simpson
> Cairns, Qld

There was tremendous disagreement among the students, even as to the conclusion. The bad news is that this remained difficult. The good news is that most students were noticeably better by about halfway through the semester.

There are two explanations for the improvement. First, their thinking and real comprehension was in fact getting better. They knew they had to distinguish structure, and were learning that it was not all arbitrary, and that there were ways to see which interpretation was better. And they were learning that they had previously been misinterpreting arguments, or failing really to interpret them at all.

Second, they were getting practice at using maps. Make no mistake: while maps may be the natural way to represent arguments, mapping is not natural. While the best way to interpret "reason" in a map may well be "evidence for believing," that is itself a convention you have

to choose (or accept) and learn to apply strictly. By the time they take this subject, these bright students have had at least twelve school years' writing and reading prose: whatever their shortcomings, they are well-practiced at what they do. Most of them have never seen an argument map before, and have to learn how to read it, write it, manipulate it, and evaluate it. That takes practice.

Nonetheless, they learn it and thereby get much better at thinking.

## *Conclusion*

In addition to the strong empirical data in favor of the argument-mapping-based Reason! approach, I have the strong impression that students really were learning to understand the structure of arguments better, and forcing themselves to think clearly. Several said they were using maps to plan other essays and felt it helped. Despite my own training in analytic philosophy, I feel that mapping helps me with my own thinking.

In summary I believe that the Reason! method and argument mapping in particular really does work. Figure 3 shows why.

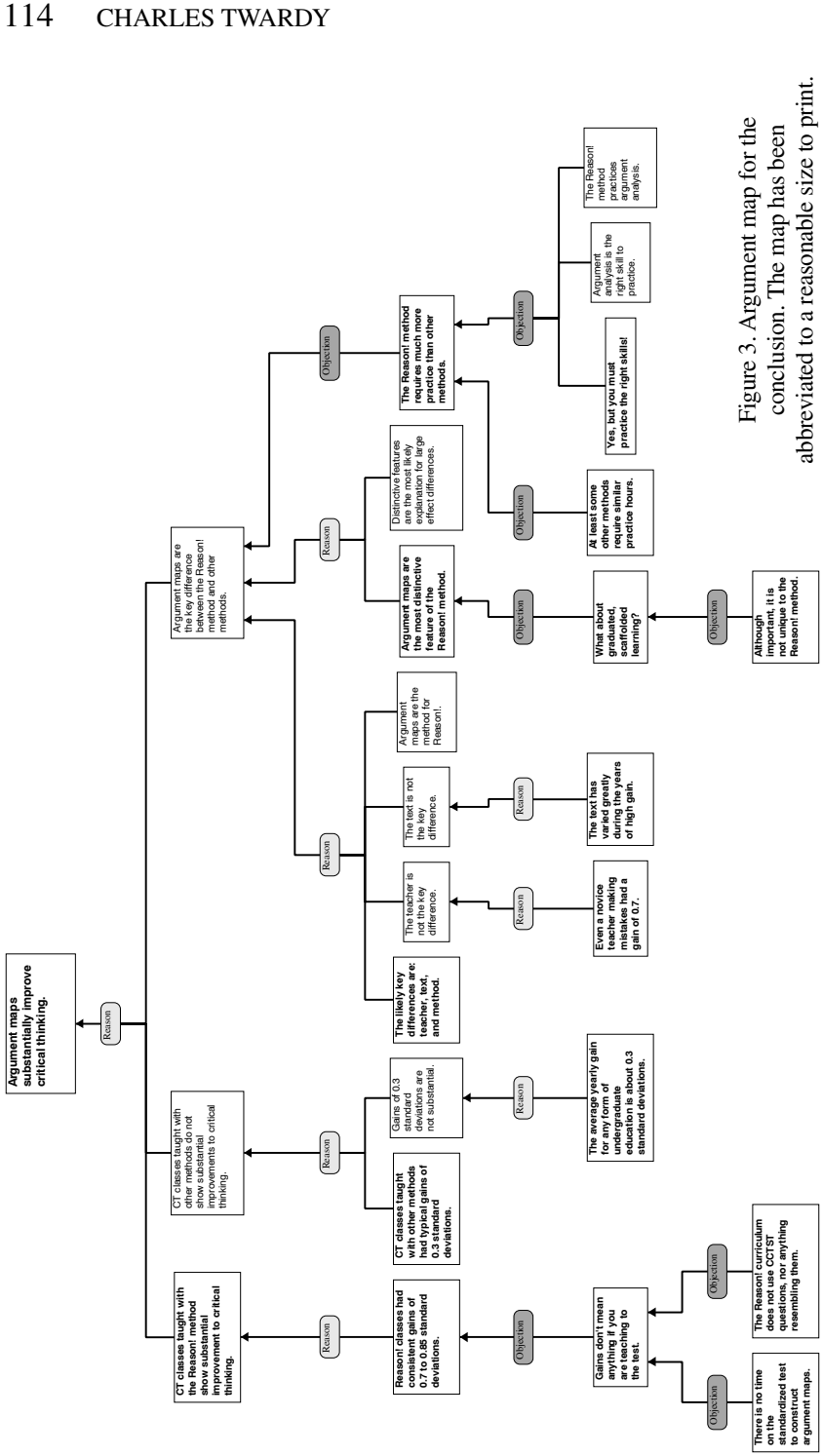



Figure 3. Argument map for the conclusion. The map has been abbreviated to a reasonable size to print.

## Notes

Thanks to Neil Thomason, Tim van Gelder, Kevin Orb, Ian Gold, and an anonymous referee for very helpful comments. NSF grant SES 99-06565 supported me during the time I taught the subjects described here.

1. For an overview of the software and how to use it, see van Gelder (2002), also available on the Reason!Able website http://www.goreason.com/papers.

2. We could use either the standard deviation in our raw pre-test scores, or that quoted the CCTST test manual. Donohue et al. (2002) did both. The resulting effect sizes were always within 0.07 of each other.

3. If the 95 percent confidence interval does not overlap the value of interest (0 in a traditional null-hypothesis test), then $p < .05$ relative to that value. For the Reason! results, the large distance between 0 and the confidence interval corresponds to p values of about .001. But a large enough study could have $p < .001$ with a tiny effect size, just because the confidence interval is small. For these and other reasons, the American Psychological Association's Publication Manual now recommends using confidence intervals on effect sizes, rather than p values.

4. Admittedly, the College Board is affiliated with ETS, the maker of the SAT and has an interest in showing that their test is not so coachable, but they had a large survey (3100 students) and results comparable to many previous independent studies.

5. A less selective meta-analysis of science, medical, and engineering subjects (Springer, Stanne, and Donovan 1999) reports smaller effects. It reported that small groups had much higher gains (0.55 standard deviations) than subjects which did not. But then it also claims not to have included any studies which did not use small-group methods, and does not report an effect size for such methods. However, it does report an effect size of 0.87 for science subjects generally, which is in line with Hake's result for traditional methods, as we might expect for subjects loosely-defined to use "small groups," some of which did not use pre/post-test methods at all and so were not nearly as carefully selected as those in Hake's study.

6. Hake prefers a measure called <g>, the average normalized gain, which is the average gain divided by the average total possible gain. The Reason! subjects had a <g> of 0.21 to 0.27, and the other subjects in Figure 1 had a <g> of 0.03 to 0.13. Hake was reporting <g> of about 0.5 on the FCI for the interactive teaching methods in physics.

7. See Grice collected in Grice (1989), passim.

8. Students will ask whether we should require the second premise to be something like, "Any human is mortal" since "humans" and "human" aren't really the same. But even at my pedantic best (the first three weeks of the semester, when it pays to be stricter and clearer), I did not insist on such superficial changes.

9. Tim doubts students spend anywhere near this much time on it.

10. Within reason. In practice we usually rule out one or two problems as too difficult or ambiguous for grading. But we do not tell the students this.

11. For a very interesting, if depressing, look at how we teach students to behave this way, and precisely how it destroys their intrinsic interest, read Alfie Kohn's *Punished by Rewards* (Kohn, 1993).

12. Garcia-Encinas (citing Kripke) argues that we may doubt Hume's claim even in mathematics: what are mathematical conjectures if not conceptions of possibility? Yet many are, often after decades, shown to be false. So conceivability is a fallible guide to possibility, even in pure thought.

## Bibliography

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Donohue, A., T. J. van Gelder, G. Cumming, and M. Bissett. 2002. *Reason! Project Studies, 1999–2002* (Tech. Rep. No. 2002/1). Department of Philosophy: University of Melbourne.

Educational Statistics, N. C. Forthcoming. *The Condition of Education*. Government Printing Office, 1997, NCES 97-388 (http://nces.ed.gov/pubs/ce).

Facione, P. A. 1990, 1992. *California Critical Thinking Skills Test*. Available from California Academic Press.

Garcia-Encinas, M. 2003. "*A Posteriori* Necessity in Singular Causation and the Humean Argument" *Dialectica* 57, 2003 forthcoming.

Grice, H. P. 1989. *Studies in the Way of Words.* Cambridge, Mass.: Harvard University Press.

Hake, R. R. 1998. "Interactive-Engagement vs Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses." *American Journal of Physics* 66: 64–74 (http://www.physics.indiana.edu/~sdi/ajpv3i.pdf visited 27 Feb. 2003).

_____ . 1999. *Analyzing Change/Gain Scores*. (An analysis of the data of Hake 1998 in terms of "effect sizes.") Available at http://www.physics.indiana.edu/~sdi/AnalyzingChange-Gain.pdf. Originally posted on 13 March 1999 to AERA-D—American Educational Research Association's Division D, Measurement and Research Methodology.)

Hitchcock, D. L. 2003. "The Effectiveness of Computer-Assisted Instruction in Critical Thinking." (Unpublished draft manuscript accepted for a conference in May 2003. Available from http://www.humanities.mcmaster.ca/~hitchckd/effectiveness.pdf, cited with permission.)

Kohn, A. S. 1993. *Punished By Rewards.* Boston: Houghton-Mifflin.

Mazur, E. 1998. *Peer Instruction: A User's Manual*. Upper Saddle River, N.J.: Prentice-Hall.

Pascarella, E. T., and P. Terenzini. Forthcoming. *How College Affects Students Revisited: Research from the Decade of the 1990s*. San Francisco: Jossey-Bass.

Powers, D. E., and D. A. Rock. 1998. *Effects of Coaching on SAT I: Reasoning Scores* (Report No. 98-6). The College Board.

Scriven, M. 1998. *Reasoning*. New York: McGraw-Hill.

Springer, L., M. E. Stanne, and S. S. Donovan. 1999. "Undergraduates in Science, Mathematics, Engineering, and Technology: A Meta-Analysis." *Review of Educational Research* 69:1: 21–51 (http://www.aera.net/pubs/rer/abs/rer691-3.htm).

van Gelder, T. J. 2001. "How to Improve Critical Thinking Using Educational Technology," in *Meeting at the Crossroads. Proceedings of the 18th Annual Conference of the Australian Society for Computers In Learning In Tertiary Education (ASCILITE 2001)*, ed. G. Kennedy, M. Keppell, C. McNaught, and T. Petrovic, pp. 539–548. (Melbourne: Biomedical Multimedia Unit, The University of Melbourne) (http://www.medfac.unimelb.edu.au/ascilite2001/pdf/papers/vangeldert.pdf).

_____ 2002. "Argument Mapping with Reason!Able." *American Philosophical Association Newsletter on Philosophy and Computers*.

van Gelder, T. J., and G. Cumming. Forthcoming. *Change in Critical Thinking Skills During College: A Meta-Analysis.*

Walton, Douglas. 2000. "Problems and Useful Techniques: My Experiences in Teaching Courses in Argumentation, Informal Logic and Critical Thinking." *Informal Logic* 20, Teaching Supplement No. 2.

*Charles R. Twardy, Computer Science and Software Engineering, Monash University, VIC 3800, Australia, ctwardy@alumni.indiana.edu*