11

Using Argument Mapping to Improve Critical Thinking Skills

Tim van Gelder

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 Roksa 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" (2011, 35), citing (HERI 2009).

However a major message of their work is that college education generally makes little progress toward this goal: "Many students are only minimally improving their skills in critical thinking, complex reasoning, and writing during their journeys through higher education" (35). 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" (36).

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, 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" (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" (1120).

Therefore an important challenge in improving critical thinking is clearly identifying the types of CT instruction that have the most impact on AM skills. One type of instruction that seems to be showing significant promise in this regard is argument mapping (AM). This chapter briefly reviews AM-based instruction and the evidence that such instruction is an effective way to improve CT skills.

Argument mapping

Argument mapping, also known as argument diagramming or argument visualization, 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. For an example see figure 11.1.

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

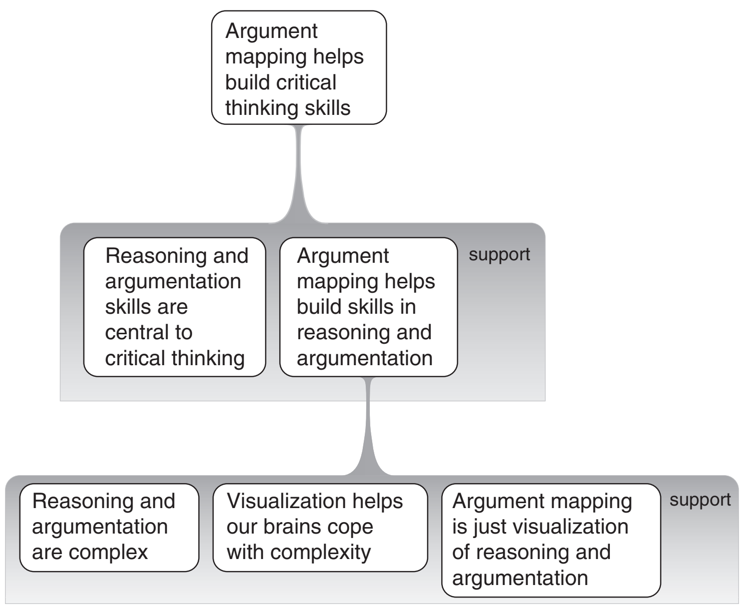

Figure 11.1 A map of an argument for the proposition that argument mapping helps build critical thinking skills. Map produced using the Rationale software (van Gelder 2007).

reasoning and CT skills. Indeed its immediate precursor was the kind of argument diagramming found in many introductory textbooks (see e.g., Fisher 1988; Govier 1988).

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

Using AM in CT instruction

As mentioned, argument diagramming of some sort is frequently found in introductory critical thinking textbooks, though it is generally treated as just one useful technique among many. AM-based instruction goes further in making argument mapping the primary or central method used to develop CT skills. Typically, this involves setting lots of AM exercises using dedicated AM software (though see Harrell 2008). A range of concepts and principles have been developed to help students map arguments properly (e.g., ter Berg, van Gelder, Patterson, and Teppema 2013).

The most common type of exercise involves providing a short text and requiring the student to identify and map out the argument it contains, that is, to produce an argument diagram faithfully representing the reasoning in the text. This can be surprisingly difficult, even for philosophers and others with prior training in argument analysis.

Another common type of exercise is requiring students to develop an AM representing an argument of their own creation, which may be preparatory to drafting an argumentative text. A third type of exercise is taking an argument map and rendering it into fluid argumentative prose.

Good AM-based instruction, like good instruction generally, presents a carefully graduated sequence of exercises of increasing difficulty. Also, as with instruction generally, good AM-based instruction requires students to receive good-quality feedback on their work. This requires human instructors with strong AM skills. Such people are in short supply, so this is a key obstacle to wider uptake of AM-based instruction.

Does it work?

It is prima facie plausible that learning and practicing AM would help students build their critical thinking skills. Reasoning and argumentation are not the entirety of critical thinking, but they are central to it; and AM should help build skills in reasoning and argumentation. There is a simple and compelling, almost syllogistic, argument for the latter point, with two uncontroversial premises: reasoning and argumentation are complex, and visualization, in general, helps our brains cope with complexity.

Note that in figure 11.1 the overall argument has two component arguments (reasons), each of which is made up of multiple premises. This argument has prima facie plausibility but requires buttressing with empirical evidence.

For many AM instructors this kind of general argument draws apparent support from what they see in the classroom. Students' attempts to think critically are frequently confounded by an inability to disentangle the threads of inference running through disputation on any given topic. To AM instructors it seems obvious that the diagrams help students grasp what is going on, and over time, help build logical acuity and facility.

However, as good critical thinkers we also know informal observations are suspect on matters of any subtlety or complexity. After all, medical practitioners for centuries thought they could see the benefits of bleeding patients. What we'd like is more empirically rigorous substantiation or validation of the "prima facie" case.

We also want to better understand just how well AM works. The claim that AM "helps build" CT skills is disturbingly vague. We would much prefer to have rigorous quantitative insight into the extent to which AM builds critical thinking skills, and how it compares with other instructional approaches. But how can this be obtained?

Empirical research on impact of AM instruction

Most readers would be familiar with the idea that the "gold standard" in social scientific research is the large, randomized controlled trial. Applied to evaluating the effectiveness of AM as an instructional method for CT, this would mean taking a large number of students and randomly assigning them to two CT subjects. One subject would make substantial use of AM, and the other would be similar in all significant respects except that it does not use AM, using instead some more traditional form of instruction. At the end of the instruction period, students in both subjects would be tested for their CT skills, using the same good-quality test for both groups. AM would then be deemed effective just to the extent that the students in the AM-based subject score more highly.

Unfortunately, no such study has ever been conducted. There are a number of reasons. Numerous practical challenges stand in the way, such as the bureaucratic difficulties involved in setting up two versions of a subject and making a genuinely random assignment of students to one or the other. There is also the difficulty of ensuring that the two subjects are sufficiently close to identical in all significant respects other than the use of AM. For example, do the two subjects cover essentially the same content, despite the difference in method? Are the students equally motivated to perform?

For these reasons, most efforts to rigorously evaluate the impact of AM on CT have taken a different approach, seeking to understand the impact of AM-based instruction by testing students at the start (pre-testing) and at the end (post-testing) of the instruction, and comparing the results.

Although pre- and post-testing is far more feasible than conducting a full-scale RCT, it is not without practical challenges of its own. For example, there is the problem of ensuring that students put proper effort into the tests, and that they are equally motivated to perform on both tests. Degree of motivation can make a huge difference to performance (Liu, Bridgeman, and Adler 2012), and if students slack off on the post-test, the real gain might be seriously underestimated (or vice versa).

Examples of the pre- and post-testing approach are the studies reported by van Gelder and colleagues (van Gelder, Bissett, and Cumming 2004). Starting in the late 1990s, the "Reason Project" at the University of Melbourne developed a radical alternative to traditional CT instruction, based on extensive deliberate practice (Ericsson, Krampe, and Tesche-Römer 1993) using AM. A dedicated AM software package, *Reason!Able* (van Heuveln 2004), was developed to support the approach. (Note that *Reason!Able* was the precursor to *Rationale* [van Gelder 2007].) Aided by a grant from the Australian Research Council, in two subjects students exposed to the approach were pre- and post-tested using the California Critical Thinking Skills Test (Facione 1991). To handle the motivation issue, students were assigned 5% of their overall score for the subject for their best performance on the two tests.

The data indicated that students had improved their CT skills by around 20%. Since CT is a generic cognitive skill, and since it is notoriously difficult to raise performance in such skills, this sounds like a substantial and worthwhile improvement, suggesting that the new approach works quite well. However this conclusion would be a bit hasty. For proper interpretation, the results need to be put in context.

First, we have to consider how much the students' CT skills would have improved anyway, due to factors such as maturation and just being at university. Second, and similarly, we have to consider how much the students would have improved due to the fact that they were taking a CT subject. Perhaps all or most of the 20% gain was due to experiencing CT instruction of some kind, rather than AM-based instruction specifically.

Given the simple pre- and post-test study design without a control group, both these issues need to be addressed by looking at the results of other studies. We need a good general estimate of how much we would normally expect students to gain in CT over one semester at university, and a good general estimate of how much we would expect students to gain over one semester in a CT subject.

Now there are plenty of studies that address these questions, particularly the former (i.e., typical gain over a university semester). But here we face more problems. First, there are quite a few tests of critical thinking, and some are more difficult than others. A 20% gain on one test may not be equivalent to a 20% gain on another. Second, the available studies are heterogeneous, differing in many key aspects such as the size of the study (number of participants), the size of the gains (or losses) they found, the quality of the instruction in the subject being assessed, and level of care and rigor involved in the assessment. In the face of all this, how does one know, or estimate, what the "true" gains are?

Taking a meta-analytic approach

The best approach to handling these issues is to use a procedure called meta-analysis. In essence, meta-analysis is a way of pooling studies together to identify common trends or effects. Meta-analysis is a complex topic, but fortunately there are excellent introductions available—for example Cumming (2012). Here I'll describe the bare minimum required to understand the empirical results we have obtained for studies of AM-based instruction.

For current purposes, meta-analysis has three main steps:

1. *Select studies*. The first step is to determine which studies should be pooled together. This involves searching far and wide for potentially relevant studies, including unpublished studies, then using a set of criteria to determine which of these "make the cut," that is, are included in the data analysis.
2. *Convert results to effect sizes*. As indicated above, studies use a variety of different tests of critical thinking. They also report their results in a variety of ways. To enable pooling, these results need to be made commensurable. A common way to do this is to express the gain (or loss) found in a particular study as a proportion of the extent of the variability in performance on the test (technically, "Cohen's *d*" [Cohen 1969]; see also [Cumming 2012], Chapter 11) a figure often referred to as the "effect size." For example, in the van Gelder et al. studies mentionedearlier, a gain of around 20% converted to an effect size of around 0.87.
3. *Calculate pooled effect sizes*. Finally, results are thrown into the pool. This is not just a matter of finding the average effect size. Rather, effect sizes of individual studies are weighted by the size of the study (i.e., the number of participants), then the average is calculated. This gives the results of larger studies more weight, on the grounds that they are less susceptible to statistical noise and so more likely to indicate the true gain.

What do we find if we apply meta-analysis to studies of AM-based approaches to CT instruction? For some years, we have been gathering relevant studies and pooling their results. At time of writing, we obtained results twenty-six pre- and post studies of AM-based instruction in a one-semester CT subject, from institutions in Australia, Europe, and the United States. Many of these are unpublished, but published studies include those found in Butchart (2009), Dwyer, Hogan, and Stewart (2011; 2012), Harrell (2011), Twardy (2004), and van Gelder, Bissett, and Cumming (2004).

This is work in progress, but it currently appears that the weighted effect size for AM-based CT instruction is around 0.7. This effect size is based on all studies that meet the basic criteria for inclusion in the meta-analysis. However within that set there are clear differences in what we call the "intensity" of AM-based learning activities. Dividing the studies into high, medium, and low-intensity groups, we find a clear relation between intensity and amount of gain. In a high-intensity study, students took a subject in which AM was the primary or central activity, with lots of homework activities, and with instructors with high proficiency in AM. Fifteen of the twenty-six studies were high-intensity, and the weighted average effect size for these studies is 0.85.

However, as compared with a 20% gain, talking of an effect size of 0.85 means little to most people and may sound negligible. How do we gauge its significance? One approach is to use the rule of thumb recommended by one of the pioneers of meta-analysis, according to which 0.2 is a small effect, 0.5 is medium, and 0.8 is large (Cohen 1969).

This makes it seem like AM-based instruction has a "large" effect, but we haven't yet taken into account how much students would have gained anyway, even without AM. To estimate these, we turn to other meta-analyses. For example Alvarez conducted a meta-analysis of studies of gains in CT over one semester at college or university (Alvarez 2007), and found an effect of around 0.11 over one semester just due to maturation and being at college generally. This is a little larger than the 0.18 gain over two years of undergraduate education found identified in Arum and Roksa's large study (2011). The conclusion we can safely draw from these numbers is that the "value add" of AM-based CT instruction, relative to just being at college, is around 0.6 (or 0.7 for high-intensity AM), which is somewhere between a medium and a large effect size. Or, put another way, AM-based CT instruction yields many times the gain in CT skills over one semester than is normally achieved by just being at college.

Using a similar approach we can estimate the benefit of AM over other forms of CT instruction. In their meta-analysis (mentioned earlier) Abrami et al. found an effect size of 0.34 for college-level CT instruction generally (Abrami 2008), so AM-based instruction appears substantially more effective than other forms of CT instruction generally. (This doesn't rule out the possibility of some other particular form of instruction being at least as effective as AM-based instruction.)

Future directions

The upshot of the previous section is that increasing amounts of empirical research have been lending convergent support to the intuitively plausible idea that AM can substantially enhance critical thinking skills. Indeed, at this stage it seems fair to say that high-intensity AM-based instruction is one of the most effective techniques we know for accelerating CT skill gains in higher education.

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 AM software package might facilitate better thinking *performance* (van Gelder 2007). He canvassed three potential causal mechanisms:

1. that such software is more "usable" than the standard technologies we use for representing and manipulating reasoning;
2. that such software complements the strengths and weaknesses of our inbuilt cognitive machinery; and
3. that AM represents a semiformal "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 also learning of CT skills. Another potential causal mechanism is that working with argument maps builds, in the learners' minds, mental templates or schemas for argument structures, making it easier for them to critically evaluate argumentation.

What dimensions of CT are being enhanced?

CT is multidimensional. For example, the Halpern Critical Thinking Assessment 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.

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 intense training?

Even the most intense 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 0.85 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 intense instruction.

What would it take to achieve gains of this order?

1. *Combining AM with other general approaches*. AM techniques should be used in conjunction with other techniques known to enhance learning, such as mastery learning (Kulik, Kulik, and Bangert-Drowns 1990) and peer instruction (Crouch and Mazur 2001), as suggested by Neil Thomason.
2. *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).
3. *Improved mapping tools*. The AM 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-length instruction using AM are very good.

References

Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M. A., Tamim, R., and Zhang, D. 2008. "Instructional Interventions Affecting Critical Thinking Skills and Dispositions: A Stage 1 Meta-Analysis." *Review of Educational Research* 78 (4): 1102–1134.

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

Arum, R., and Roksa, J. 2011. *Academically Adrift : Limited Learning on College Campuses*. Chicago: University of Chicago Press.

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* 25 (2): 268–291.

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

Crouch, C., and Mazur, R. 2001. “Peer Instruction: Ten Years of Experience and Results.” *American Journal of Physics* 69: 970–977.
Cumming, G. 2012. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York: Routledge.
Davies, M. 2011. “Concept Mapping, Mind Mapping, Argument Mapping: What Are the Differences and Do They Matter?” *Higher Education* 62 (3): 279–301.
Dwyer, C. P., Hogan, M. J., and Stewart, I. 2011. “The Promotion of Critical Thinking Skills through Argument Mapping.” In *Critical Thinking*, edited by C. P. Horvath and J. M. Forte. New York: Nova Science Publishers. 97–122.
Dwyer, C. P., Hogan, M. J., and Stewart, I. 2012. “An Evaluation of Argument Mapping as a Method of Enhancing Critical Thinking Performance in E-Learning Environments.” *Metacognition & Learning* 7: 219–244.
Ericsson, K. A., Krampe, R. T., and Tesche-Römer, C. 1993. “The Role of Deliberate Practice in the Acquisition of Expert Performance.” *Psychological Review* 100: 363–406.
Facione, P. A. 1991. “Using the California Critical Thinking Skills Test in Research, Evaluation, and Assessment.” ED337498.
Fisher, A. 1988. *The Logic of Real Arguments*. Cambridge, England; New York: Cambridge University Press.
Govier, T. 1988. *A Practical Study of Argument*. (second edition). Belmont, CA: Wadsworth Pub. Co.
Halpern, D. F. 2010. *Halpern Critical Thinking Assessment*. Vienna: Schuhfried.
Harrell, M. 2008. “No Computer Program Required.” *Teaching Philosophy* 31 (4): 351–374.
Harrell, M. 2011. “Argument Diagramming and Critical Thinking in Introductory Philosophy.” *Higher Education Research and Development* 30 (3): 371–385.
HERI. 2009. *The American College Teacher: National Norms for 2007–2008*. Los Angeles: Higher Education Research Institute, University of California.
Kulik, C., Kulik, J., and Bangert-Drowns, R. 1990. “Effectiveness of Mastery Learning Programs: A Meta-Analysis.” *Review of Educational Research* 60 (2): 265–306.
Liu, L., Bridgeman, B., and Adler, R. 2012. “Measuring Learning Outcomes in Higher Education: Motivation Matters.” *Educational Researcher* 41 (9): 352–362.
Macagno, F., Reed, C., and Walton, D. 2007. “Argument Diagramming in Logic, Artificial Intelligence, and Law.” *Knowledge Engineering Review* 22 (1): 87–109.
Mencken, H. L. 1997. *Minority Report : H. L. Mencken's Notebooks*. (Johns Hopkins paperbacks edition). *Maryland Paperback Bookshelf*. Baltimore, MD: Johns Hopkins University Press.
ter Berg, T., van Gelder, T. J., Patterson, F., and Teppema, S. 2013. *Critical Thinking: Reasoning and Communicating with Rationale*. Seattle, WA: CreateSpace.
Twardy, C. 2004. “Argument Maps Improve Critical Thinking.” *Teaching Philosophy* 27 (2): 95–116
van Gelder, T. J. 2007. “The Rationale for Rationale™.” *Law, Probability and Risk* 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, T. J., Bissett, M., and Cumming, G. 2004. “Cultivating Expertise in Informal Reasoning.” *Canadian Journal of Experimental Psychology* 58: 142–152.