Reason! Project Studies, 1999-2002

Angela Donohue

Tim van Gelder

Geoff Cumming

Melanie Bissett

Reason! Project Technical Report 2002/1

The Reason! Project

Department of Philosophy

University of Melbourne

Parkville VIC 3052

tgelder@unimelb.edu.au

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

Acknowledgements

People who assisted with data collection or the preparation of this report include Yanna Rider, Andy Bulka, and Neil Thomason. The research was supported in part by the University of Melbourne (T&LMET), the Australian Research Council, and Austhink (www.austhink.org).

1 Introduction

The Reason! Project has been developing new methods for teaching critical thinking (CT), focusing on general informal reasoning and argument skills. The methods have been designed for use in one-semester first-year undergraduate subjects.

The Reason! approach is based on the Quality Practice Hypothesis, the conjecture that general CT skills will improve to the extent that students engage in quality practice, i.e., practice that is

- Motivated
- Guided
- Graduated
- Scaffolded
- Subject to feedback
- Diverse, i.e., in many different domains or subject matters
- "Practice for transfer," i.e., practice in the transferring of skills from one domain or context to another.

In order to support quality practice, the Reason! Project has developed the Reason!Able practice environment, a software package based on argument mapping. The Reason! approach requires students to engage in extensive practice using Reason!Able.

In order to assess the new methods' effectiveness we have collected data addressing three main questions:

1. How much do students improve in a Reason!-based CT subject, as compared with a traditional CT subject or not taking a CT subject?
2. Is the improvement found in a Reason!-based subject enduring?
3. In a Reason!-based subject, how does practice affect improvement, i.e., how does the amount and nature of practice affect the amount and nature of improvement?

This report summarizes the data. Raw data (Excel files) is available upon request.

2 Question 1: Studies of CT Gains

Every semester the Reason! approach was used we pre- and post-tested students to assess the extent to which their critical thinking skills improved.

For comparison, we also pre- and post-tested other critical thinking subjects using traditional methods, and non-critical thinking subjects.

Table 1: Reason Project Pre/Post Studies, 1999-2002

Study	Semester/ Year	Subject	Teaching method	Test	n	Pre test Mean (SD)	Post test Mean (SD)	Gain (SD)	Gain as % of pre	**Effect Size [a]** **Effect Size [b]**	Lower bound C.I [d]	Upper bound C.I	t [c] (two-tailed)	p [k]
1	1/ 1999	P003 [e]	Tradi-tional	WGCTA[h]	46	63.85 (8.76)	64.03 (6.58)	0.18 (7.10)	0%	**0.02** **0.02**	-0.20 -0.20	0.24 0.24	.15	.884
2	2/ 1999	P003	Reason! prototype	WGCTA	53	61.30 (9.69)	64.89 (7.46)	3.59 (7.17)	6%	**0.37** **0.39**	0.16 0.18	0.58 0.60	3.64	.001
3a	1/ 2000	P003	Reason!	CCTST[i]	50	19.50 (4.74)	23.46 (4.36)	3.96 (3.25)	20%	**0.84** **0.89**	0.64 0.69	1.04 1.09	8.61	.000
3b	1/ 2000	P003	Reason!	Written[j] Marker 1	62	3.05 (1.01)	3.56 (0.86)	0.52 (1.00)	17%	**0.51** **0.54**	0.25 0.28	0.77 0.80	4.05	.001
3c	1/ 2000	P003	Reason!	Written Marker 2	62	2.56 (0.91)	3.65 (0.92)	1.09 (1.14)	40%	**1.20** **1.13**	0.91 0.84	1.49 1.42	7.29	.000
4a	1/ 2001	P003	Reason!	CCTST	114	18.11 (4.86)	22.09 (4.27)	3.98 (5.34)	22%	**0.82** **0.89**	0.60 0.67	1.04 1.11	7.96	.000
4b	1/ 2001	P003	Reason!	Written Marker 1	112	2.72 (0.91)	3.55 (0.91)	0.82 (1.27)	30%	**0.90** **0.85**	0.66 0.61	1.14 1.09	6.82	.000
4c	1/ 2001	P003	Reason!	Written Marker 2	112	2.95 (1.06)	3.48 (1.12)	0.53 (1.43)	18%	**0.50** **0.55**	0.22 0.28	0.77 0.82	3.98	.000
5	1/ 2001	Phil. Monash[f]	Standard philoso-phy in-struction	CCTST	61	19.08 (4.13)	20.39 (4.63)	1.31 (6.57)	7%	**0.32** **0.29**	-0.05 -0.08	0.69 0.66	1.56	.124
6	1/ 2001	Phil. Monash (another subject)	6 weeks philosophy + 6 weeks traditional CT	CCTST	174	19.07 (4.72)	20.35 (5.05)	1.28 (5.45)	7%	**0.27** **0.29**	0.09 0.11	0.45 .047	3.10	.002
7	1/ 2001	HPS Melb. [g]	136-054 Science Life and Mind	CCTST	42	18.76 (4.04)	20.26 (6.14)	1.5 (7.57)	8%	**0.37** **0.34**	-0.14 -0.17	0.88 0.85	1.28	.207
8	1/ 2002	P003	Reason!	CCTST	117	18.85 (4.54)	22.10 (4.66)	3.25 (3.78)	17%	**0.72** **0.73**	0.57 0.58	0.87 0.88	9.3	.000

Notes:

[a] Effect size calculated using the SD of the Pre-test for the group

[b] For groups tested using the CCTST, the effect size was calculated using the SD from the manual for CCTST (SD = 4.457). For groups that were tested using the WGCTA, the effect size was calculated using the SD from the WGCTA manual (SD = 9.2) and for the written test, the SD used was the average of the SD of the four written pre-tests (SD = 0.965)

[c] t is the value obtained from a dependent samples t-test, with n-1 degrees of freedom.

[d] CI lower and upper refer to confidence intervals around the effect size, computed using d ± [1.96 X SD diff / √N]/SD population, where SD population was estimated from the published test manuals (CCTST SD = 4.457; WGCTA SD = 9.2; written tests SD = 0.965).

[e] P003 – Philosophy 161-003, Critical Thinking: The Art of Reasoning, a one-semester 1st year subject in the Department of Philosophy, University of Melbourne. In 1999-2001 it was taught by Tim van Gelder; in 2002, by Charles Twardy.

[f] Phil. Monash – a one-semester, first year Philosophy subject taught at Monash University.

[g] HPS Melb. – 136-054 Science Life and Mind, a one-semester, 2/3rd year subject in History and Philosophy of Science taught at the University of Melbourne by Neil Thomason. This subject had a strong CT emphasis but did not use any CT software.

[h] WGCTA – Watson-Glaser Critical Thinking Appraisal (Watson & Glaser, 1980). The WGCTA is a 40-minute test measuring critical thinking with 5 dimensions (inference, recognition of assumptions, deduction, interpretation and evaluation of arguments) and has a maximum score of 80. The mean for college students in Freshman year was 53.8 (SD = 9.2).

[i] CCTST – California Critical Thinking Skills Test, College Level (Facione & Facione, 1992). The CCTST is a 40-minute test consisting of 34 items, measuring critical thinking skills with 5 dimensions: analysis, evaluation, inference, deductive reasoning and inductive reasoning. The mean with a college sample was 15.89 (SD = 4.457). In CCTST studies, students were pre-tested in the first week of semester and post-tested in the final week. Since semesters were 12 weeks in duration, the time between pre- and post-test was 11 weeks.

[J] Written – 30 minute written test developed by the research team modelled on the Graduate Record Exam Writing Assessment developed by the Educational Testing Service. Written tests were scored by a CT expert who was independent of the Reason! project. The scorer was blind as to factors such as who wrote the answer and whether it was pre- or post. The scale was 0 to 5.

[k] p value computed with respect to a "null" hypothesis that there would be no gain. It provides an upper bound on the probability that the data would have occurred merely by chance had the null hypothesis been true.

Further Notes:

- In Reason! approach studies, 5% of the subject grade was awarded on the basis of the better score of the pre- and post-tests. The goal was to give students equal incentive to perform well on both tests.

- Note that studies 3a, 3b, and 3c were the same group of students, and studies 4a, 4b, and 4c were the same group of students, tested using the CCTST and a written test (which was scored twice by two independent markers).

Chart 1: Effect sizes for Reason! Project Studies

In order to compare gains across different tests, we use a standardized effect size, Cohen's d, the mean gain divided by SD_{pop}, the standard deviation of performance on the test. Here we use the SD as found in the test manual; for the written tests, we use the mean SD across all four studies.

Confidence intervals on effect sizes give the range within which we can be 95% confident the true effect size will be found. They are computed using using $d \pm [1.96 \times SD_{gain}/\sqrt{N}]/SD_{pop}$.

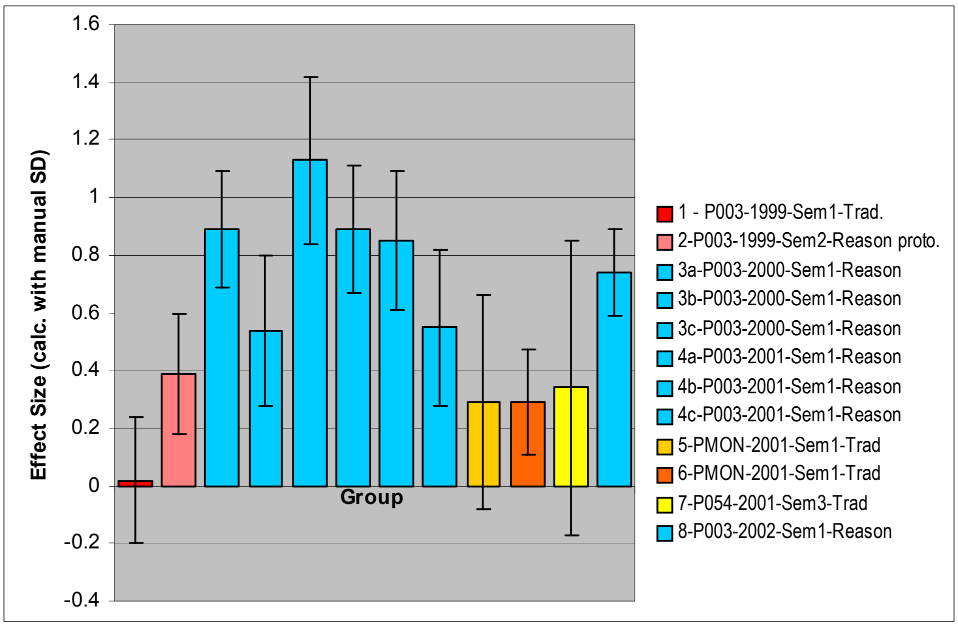
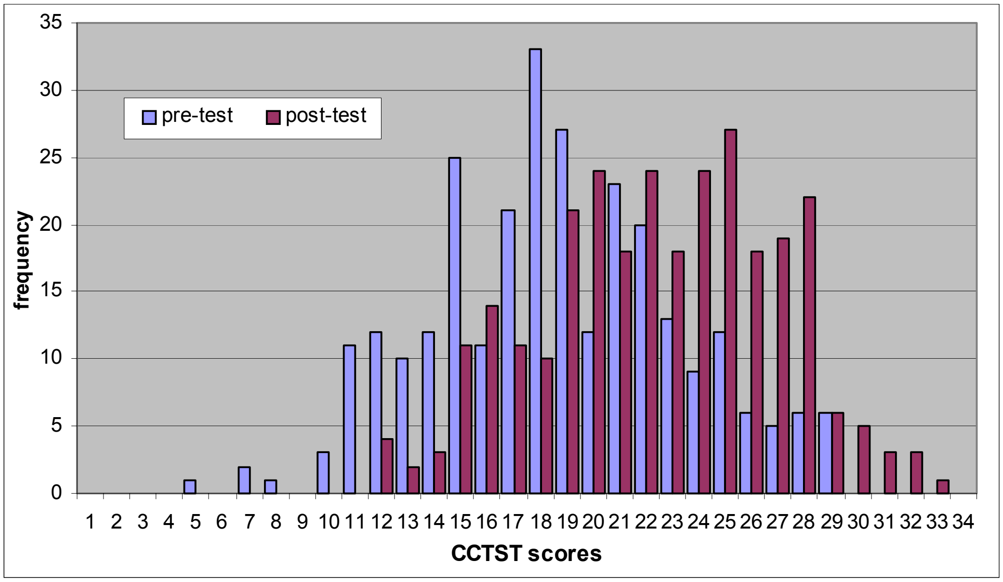


Chart 2: Distribution of CCTST results from Reason! CT students from 3 years combined (2000, 2001and 2002)



Note N = 281 (in 2000 n = 50; in 2001 n= 114 and in 2002 n = 117). CCTST is scored out of a total 34

3 Question 2: Are Reason! Gains Enduring?

In order to determine whether the gains found over one semester are stable, we conducted follow-up testing of a small sample of participants from P003 2001. The results are presented in Table 2.

Table 2: Follow-up CCTST results for 2001 Critical thinking students with the Reason! Method

Semester/ Year	n	Pre-test mean (SD)	Post test Mean (SD)	Follow-up test Mean (SD)	Mean Gain (Post or Follow-up – Pre)	Gain as % of Pre	**Effect size [b]** **Effect size [c]**	Lower Bound C.I [g]	Upper bound C.I	t (two-tailed)	p
1/2001	114	18.11 (4.86)	22.09 (4.27)	-	3.98[d] (5.34)	22%	**0.82** **0.89**	0.60 0.67	1.04 1.11	7.96	.000
1/2001[a]	20	19.30 (5.90)	23.70 (4.84)	-	4.4 [e] (5.50)	23%	**0.75** **0.99**	0.21 0.45	1.29 1.53	3.58	.002
1/2001[a]	20	19.30 (5.90)	-	23.70 (4.45)	4.4 [f] (5.61)	23%	**0.75** **0.99**	0.19 0.44	1.30 1.54	3.51	.002

Notes:

a. The second and third rows give values for the 20 students who participated in the follow-up. These are a self-selected subset of the full group. Follow-up tests were conducted in April 2002, more than 9 months after the students completed the P003 course in critical thinking at Melbourne University.

b. Effect size calculated using the SD of the Pre-test for the group (SD = 4.86 for N = 114, and SD = 5.90 for N = 20)

c. Effect size calculated using the SD from the CCTST test manual (SD = 4.457)

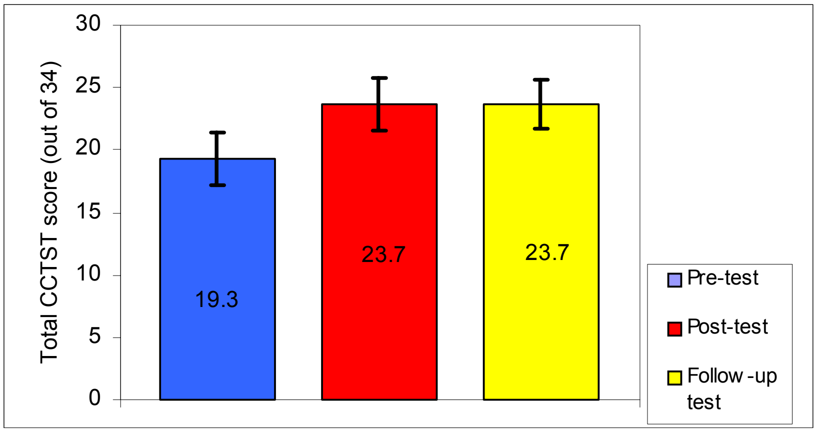
d. Gain measured from pre-test to post-test with n = 114

e. Gain measured from pre-test to post-test with follow-up sample only, n = 20

f. Gain measured from pre-test to follow-up test with sub sample n = 20

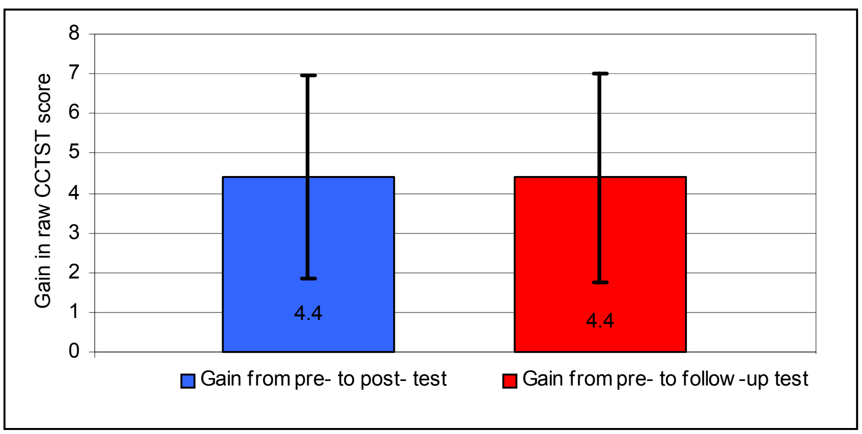
g. CI lower and upper refer to confidence intervals around the effect size, computed using $d \pm [1.96 \times SD_{diff} / \sqrt{N}]/SD_{population}$, where SD population was estimated from the published test manual (CCTST SD = 4.457).

Chart 3: Comparison of Mean CCTST scores on the pre-, post- and follow-up tests (n = 20)



Note. Error bars represent 95% confidence intervals of mean.

Chart 4 The mean gain in CCTST scores from pre- to post- test and from pre- to followup-tests (for sub-sample, N = 20)



Note. Error bars represent 95% confidence intervals of mean.

Chart 5: Pre-, post- and follow-up scores for the 20 participants who completed all three tests; subjects ordered by pre-test score.

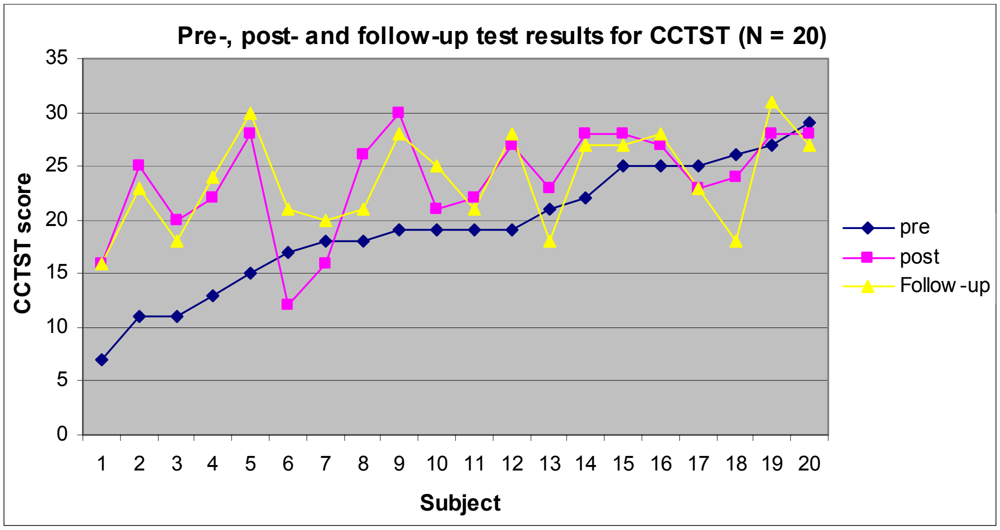


Chart 4 shows follow-up test scores tracking post-test scores. It appears that the CT gain shown immediately after the course is maintained more than 9 months later at approximately the same level.

4 Question 3: How Does Practice Affect Improvement?

One goal of the Reason! project has been investigating the Quality Practice Hypothesis. The strategy is to measure gains in CT, measure the amount of practice students were doing, and look for relationships between the two.

In 2001 and 2002, practice activities with Reason!Able were recorded by the software and the resulting data uploaded to a web server. Information recorded included:

- number of sessions
- time spent using the software
- number of actions of various kinds

The process of recording and uploading Reason!Able activity data is fraught with difficulty. In many cases, we have good reason to believe that a students' data is unreliable. For 2002 students we have selected out reliable student data.

Additionally, in 2002, students' activity with online learning materials (lessons, exercises) was recorded using the online course management system. Information recorded included:

- number of online "hits" (page or file requests)
- time spent online

In 2002 also had students fill out a questionnaire. Information obtained included:

- estimated hours of use of Reason!Able
- motivation to participate and learn

Specifically, the questionnaire included the following questions:

- Please estimate the number of homework hours you spent on each Module. (Note that only estimates for Modules 1, 2 and 3 were analysed, as Module 4 work was incomplete at the time the students completed the questionnaire.)

The following questions used a scale of 1-5:

- How hard did you work in this subject as compared with other subjects?"
- How interesting did you find this subject? (Interest)
- How useful was this subject to you? (Subjective task value)
- Did you do the tasks in this subject to a standard you were happy with? (Efficacy)
- To what extent did you have control over your learning while doing this subject? (Control)
- To what extent did the Reason!Able software help you learn in this subject? (Goals)
- To what extent did you use the Reason!Able software outside this subject? (Intrinsic motivation)
- To what extent do you think you will use the Reason!Able software in the future? (Intrinsic motivation)

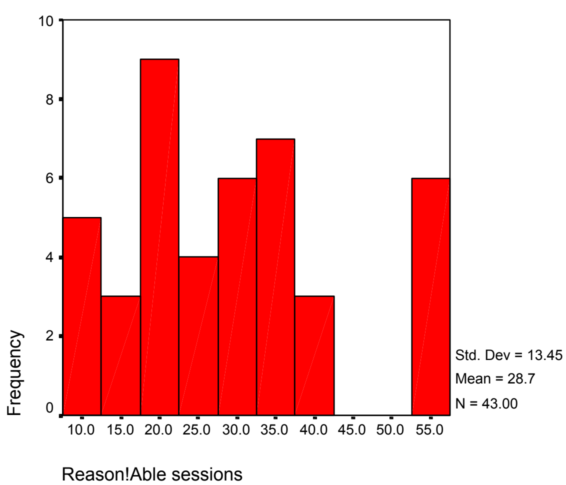
The latter 7 questions sampled 6 motivational constructs. Student responses were all highly correlated (the majority reaching $r = .99$, $p <.000$) indicating that they all measured the same construct, therefore they were combined into one variable, Motivation (Wigfield, Eccles, & Rodriguez, 1998).

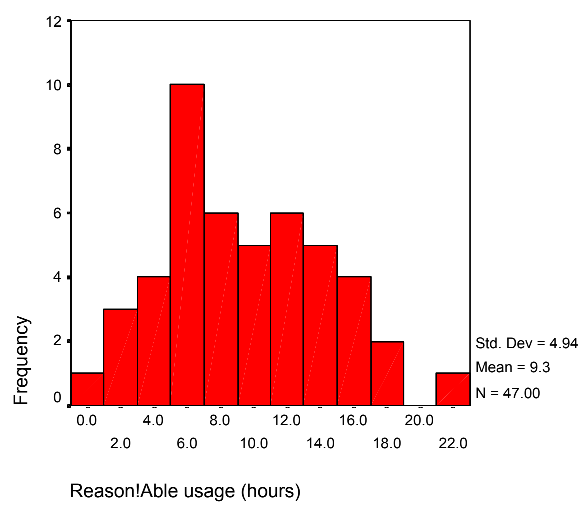
Table 3: Summary of Activity Data, 2001 and 2002

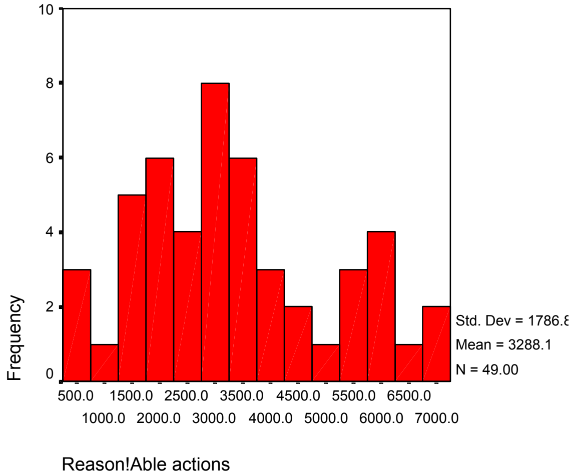
Year/ course	N	Mean number Reason!Able sessions (SD)	Mean Reason!Able Time in hours (SD)	Mean number Reason!Able Actions (SD)	Online hits (page accesses) (SD)	Online Hours (SD)	Mean Practice Participant estimated hours usage (SD)	How hard did you work?	Motivation *combined measure; mean- out of 5* (SD)
2001 P003	83	20.99 (13.53) N = 80	14.47 (10.25) N = 79	2530.91 (1874.59) N = 82	-	-	-	-	-
2001 P003 [a]	50	28.36 (11.80) N = 39	19.24 (8.48) N = 38	3446.07 (1568.92) N = 41	-	-	-	-	-
2002 P003	116	21.71 (17.55) N = 108	6.35 (5.08) N = 112	2246.46 (1824.27) N = 112	343.95 (158.0) N = 109	16.09 (14.23) N = 96	21.08 (8.85) N = 95	3.24 (0.84) N= 98	3.07 (0.57) N = 89
2002 P003 [a]	50	28.65 (13.45) N = 43	9.31 (4.94) N = 47	3288.08 (1786.82) N = 49	298.24 (99.99) N = 41	13.20 (7.96) N = 37	21.34 (8.43) N = 47	3.39 (0.80) N = 46	3.02 (0.59) N = 44

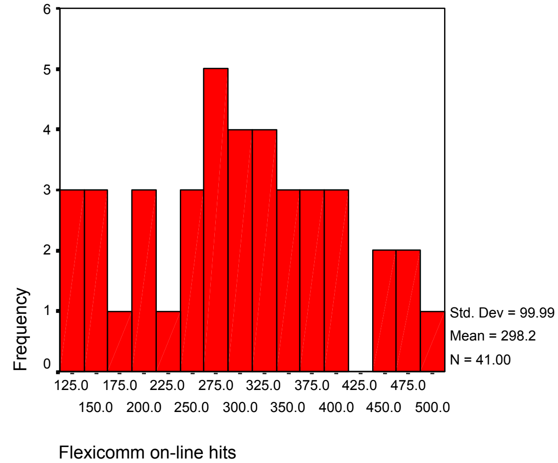
Note. [a] sub-sample of students with reliable Reason!Able data from each year, note that these students may be somewhat self-selected, e.g., better-organised, more compliant, or more motivated than the rest of the students; Outliers and extreme values were deleted based on examination of the box-plots for each variable individually.

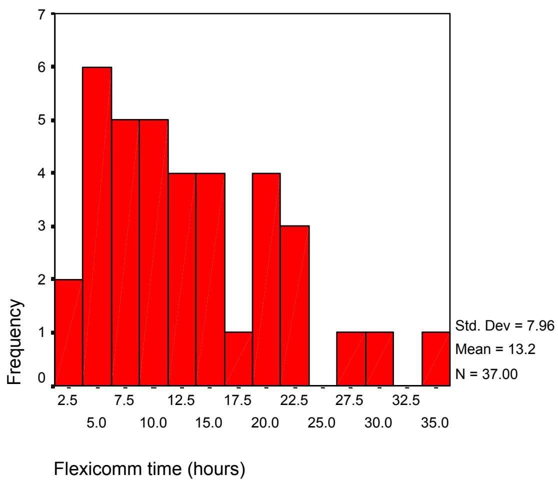
Chart 6: Distributions of Activity Data for 2002 Students with Reliable Activity Data

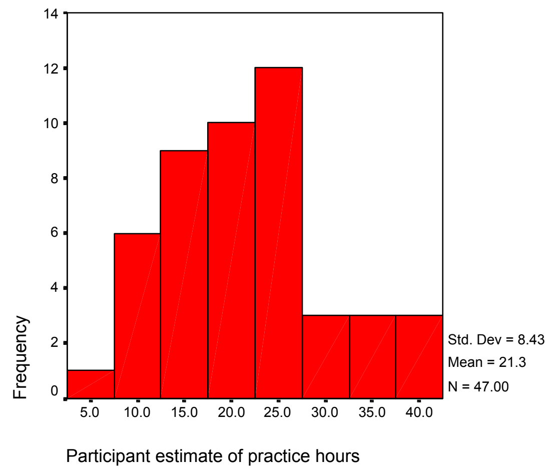


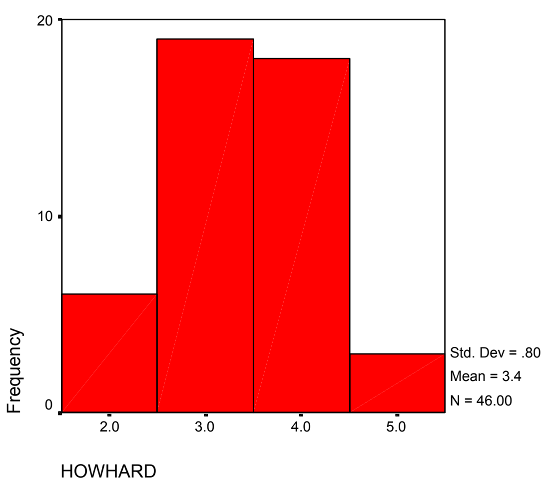


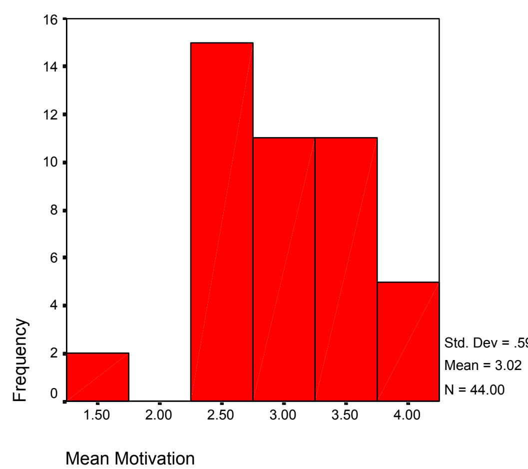

Frequency
Std. Dev = 1786.8
Mean = 3288.1
N = 49.00
Reason!Able actions


Frequency
Std. Dev = 99.99
Mean = 298.2
N = 41.00
Flexicomm on-line hits


Frequency
Std. Dev = 7.96
Mean = 13.2
N = 37.00
Flexicomm time (hours)


Frequency
Std. Dev = 8.43
Mean = 21.3
N = 47.00
Participant estimate of practice hours


Frequency
Std. Dev = .80
Mean = 3.4
N = 46.00
HOWHARD


Frequency
Std. Dev = .59
Mean = 3.02
N = 44.00
Mean Motivation

Note. Outliers and extreme values deleted based on examination of the box plots. N's vary according to whether participants completed the relevant self-report questionnaire items for charts relating to participant estimate of hours practice, mean motivation, and participant estimate of how hard they worked.

Table 4: Correlations for the students with reliable data in 2002 (N = 50)

	Post-test	Gain	Flex. time	Flex. hits	Total actions	Total Sessions	Total Time	Practice [a]	How hard worked	Motivation [b]
Pre-test	**.58*** **	**-.30* **	-.04	.15	.07	.01	-.08	-.18	-.16	.06
Post-test		**.61*** **	.01	.05	.24	.07	.21	-.06	-.14	.03
Gain			.05	-.09	.21	.08	**.32* **	.12	-.01	-.02
Online time				.17	.16	.04	.12	.09	.07	-.01
Online hits					**.49** **	**.50*** **	**.39* **	**.32** **	.08	.16
Total Reason!Able actions						**.78*** **	**.94*** **	**.55*** **	-.01	.32
Total Reason!Able sessions							**.76*** **	**.50*** **	.01	.20
Total Reason!Able time								**.61*** **	.05	.32
Practice									.20	**.37* **
How hard worked										.13

Note. * p <.05, ** p <.01 ***p<.005 (two-tailed). Bold numbers indicate significant correlations.

The following correlations were statistically significant at p < .05 or better:

Low Correlations (approx. 0.3)

- Gain and time using the Reason!Able software
- Negative correlation between pre-test score and gain
- Estimated practice and motivation

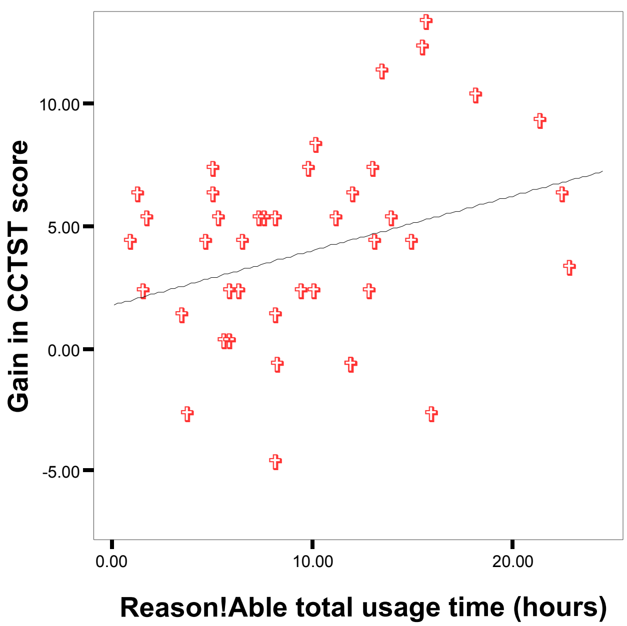
Medium Correlations (approx. 0.5)

- Pre-test score with post-test score
- Estimated practice and Reason!Able activity (sessions, actions and time). These indicate that the participant's estimates of their own practice on the self-report questionnaire did align with the usage data collected
- Online hits with online time and Reason!Able activity (sessions, actions and time). These indicate that their online learning activity was aligned with their Reason!Able software activity.

Large Correlations (approx. 0.7 or larger)

- Gain with post-test score
- Reason!Able actions with Reason!Able sessions and Reason!Able time.

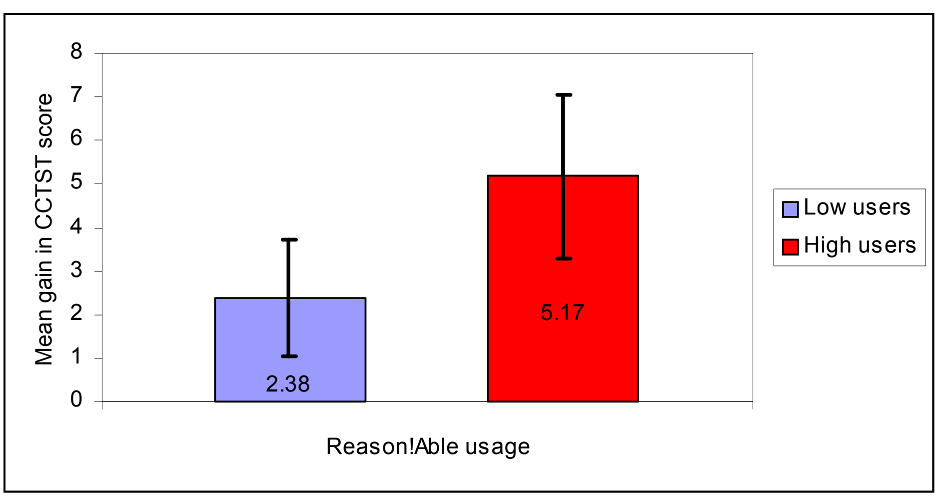
Chart 7: Scatterplot of gain in CCTST score for 2002 sample against total raw time spent using Reason!Able software (N = 42)



Note. r= .32, p <.05

We divided participants into "low users" and "high users", based on the median (8.84) number of hours practice they had completed with Reason!Able. For low-users, the mean gain was 2.38 (SD = 3.19), whereas for the high usage group, the mean gain was 5.17 (SD = 4.36).

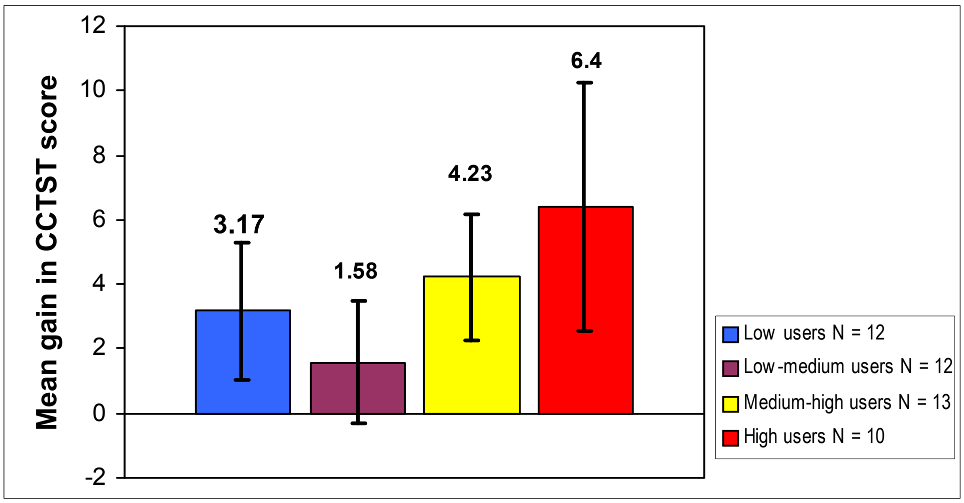
Chart 8 - Mean gain in CCTST score for low (n = 24) versus high (n = 23) Reason!Able users



Note. Error bars represent 95% confidence intervals of mean. The mean (SD) number of hours Reason!Able usage for the 2 groups was 5.29(2.25) for low users; and 13.51(3.11) for high users.

When we divide the students into 4 groups based on the 25th, the 50th and the 75th percentiles, the mean gain for low users was 3.17(SD = 3.33), the mean gain for the low to medium users group was 1.58(SD = 2.97), the mean gain for the medium to high usage group was 4.23 (SD = 3.24) whereas for the high usage group, the mean gain was 6.40 (SD = 5.42).

Chart 9 - Mean gain in CCTST score, with participants divided into 4 usage groups based on Reason!Able use (N = 47)



Note. Error bars represent 95% confidence intervals of mean. The mean (SD) number of hours Reason!Able usage for the 4 groups was 3.29(3.07) for the low users; 7.07(1.07) for the low- to medium users, 11.42(1.35) for the medium to high users and 17.44(3.55) for the high users.

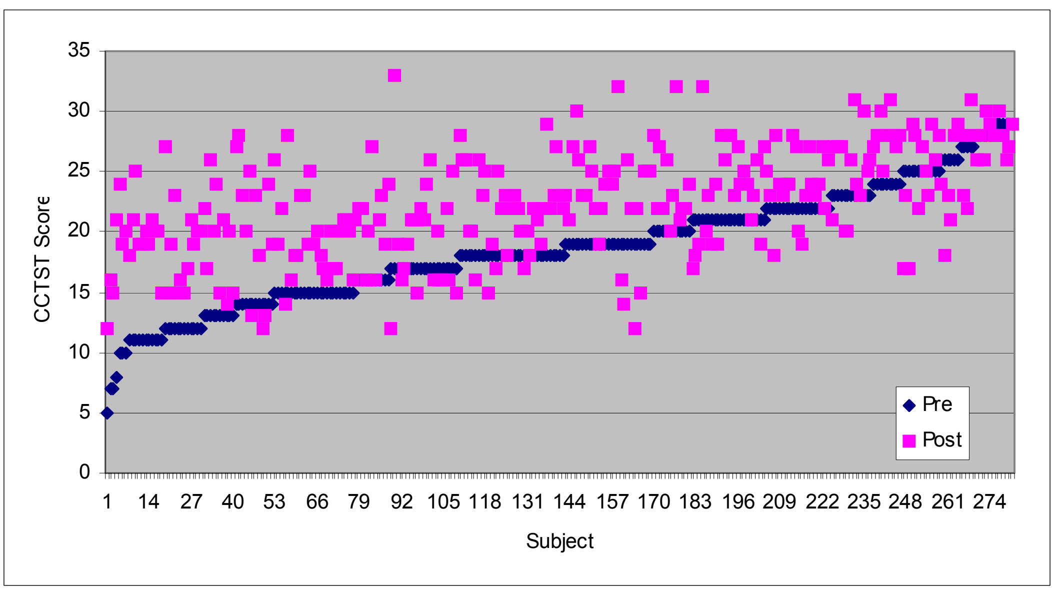
4.1 A ceiling effect?

For students in 2000, 2001 and 2002 combined there was a moderate, significant, negative correlation between pre-test score and gain ($r = -.54$, $p < .005$). That is, low pre-test scorers tended to gain more than high pre-test scorers.

Chart 10: Relationship between pre-test score and gain for 2000, 2001 and 2002 (N = 281)

Note: r = -.54, p<.005; the red-dotted line indicates the highest possible gain from pre-test score (ceiling; ie. If pre test score was 16, then the highest possible gain is 18, since the test is scored out of 34). Some dots represent multiple subjects.

Chart 11: All subjects, pre- and post-test scores, subjects ordered by pre-test score.



Possible explanations of negative correlation include:

- A ceiling effect, whereby high scorers had limited opportunity to demonstrate improvement on the instruments, since their scores were initially close to the maximum possible scores.
- Regression to the mean, arising from inevitable error of measurement. A ceiling effect coupled with such regression could lead lower pretest scores to tend to have larger gain scores than higher pre-test scores.

- The course trained students in skills which were already largely possessed by high pre-test scorers but not by low pre-test scorers; hence high pre-test scorers did not change their skill level as much.

5 Conclusions

These studies suggest the following conclusions:

1. Students in subjects using the Reason! approach improved with an effect size of approximately 0.8 (studies 3, 4 & 8).
2. Gains do not depend on using a specific multi-choice test, and they are not produced by "teaching to" a particular test or scoring rubric (studies 3, 4).
3. There is weak evidence that CT gains produced by the Reason! approach are stable (section 3), and do not depend on the course being taught by the developer of the approach (study 8).
4. CT gains in the Reason! approach appear to be weakly correlated with practice, although we suspect our studies to date understate the size of this relationship.
5. In the Reason! Approach, weaker students gain more than stronger students when tested with the CCTST.
6. Students in first year improve over one semester with an effect size of approximately 0.2 even when not studying CT (study 5).
7. Traditionally-taught CT subjects produce little if any gain over and above that which would occur anyway (studies 1, 6).
8. Therefore, the Reason! approach substantially accelerates CT growth compared with not studying CT or taking a traditional CT subject.

Further research is required to substantiate and elaborate these points.

6 References

Facione, P. A., & Facione, N. C. (1992). The California Critical Thinking Skills Test and Manual. Millbrae, CA, California Academic Press.

Watson, G. B., & Glaser, E. M. (1980). Watson-Glaser Critical Thinking Appraisal Manual. San Antonio, CA: The Psychological Corporation.

Wigfield, A., Eccles, J. S., & Rodriguez, D. (1998). The development of children's motivation in school contexts. In P. D. Pearson & A. Iran-Nejad (Eds.), Review of Research in Education, 23 (pp. 73-118). Washington DC: American Educational Research Foundation.