The development of reasoning skills during compulsory 16 to 18 mathematics education.

Nina Attridge

Centre for Pain Research, University of Bath, UK

Maria Doritou

Kalogeropoulou Gymnasium, Cyprus

Matthew Inglis

Mathematics Education Centre, Loughborough University, UK

Acknowledgements.

This research was partially supported by a Royal Society Worshipful Company of Actuaries Research Fellowship to MI.

Disclosures.

The authors have no conflicts of interest to disclose.

Abstract

The belief that studying mathematics improves reasoning skills, known as the Theory of Formal Discipline, has been held since the time of Plato. Research evidence supports this idea, at least in the context of students who had chosen to study post-compulsory mathematics. Here we examined the development of reasoning skills in 16- to 18-year-old Cypriot students, who are required to study mathematics until age 18. One hundred and eighty eight students, studying high- or low-intensity mathematics, completed the abstract Conditional Inference Task and the contextual Belief Bias Syllogisms task at ages 16, 17 and 18. While the high-intensity group improved on the conditional inference task and showed a reduction in belief bias, the low-intensity group did not change on either measure. This is promising for the Theory of Formal Discipline, but suggests that a certain level of mathematical study may be necessary for students' general reasoning skills to develop.

Keywords: Theory of Formal Discipline, reasoning, international comparison

Many philosophers, mathematicians, and policy-makers have claimed that studying mathematics improves reasoning skills. For example, John Locke stated that mathematics should be taught to "all those who have time and opportunity, not so much to make them mathematicians as to make them reasonable creatures" (1706/1971, p.20), while Oakley (1949) suggested that "the study of mathematics cannot be replaced by any other activity that will train and develop man's purely logical faculties to the same level of rationality". These views reflect the Theory of Formal Discipline (TFD), which suggests that studying formal subjects, such as mathematics or Latin, improves thinking skills more broadly.

Views about the veracity of the TFD may influence educational curricula. For example, Stanic (1986) and Stanic and Kilpatrick (1992) reported that changes to the US school-level mathematics curriculum in the twentieth century were related to views about the TFD in the mathematics education community. It is therefore important that views on the TFD are based on research evidence, which historically has been rather sparse. The aim of the current paper is to investigate the development of reasoning skills in 16- to 18-year-old mathematics students in Cyprus. Before describing the study in detail, we review the existing evidence on the relationship between mathematical study and improvements in reasoning skills.

Lehman and Nisbett (1990) measured the conditional reasoning skills of US undergraduates in their first and fourth years of study. They found a correlation between the number of mathematics courses that natural science students had taken and the extent of improvement on their conditional reasoning task. Inglis and Simpson (2009) did not find any improvement in abstract conditional reasoning skills in English mathematics undergraduates, but the mathematics students did outperform arts students on their conditional reasoning task at the beginning of their university

studies. They argued that studying mathematics at A level (a non-compulsory qualification taken at the ages of 16-18 in England and Wales) may have been responsible for the initial difference between groups. This was supported by Attridge and Inglis (2013) who studied abstract conditional reasoning skills in mathematics and literature A level students, and found no difference between groups in reasoning behaviour at the start of post-compulsory education. While the literature students' reasoning did not change during their first year of undergraduate studies, the mathematics students' did.

However, it should be noted that the change Attridge and Inglis (2013) observed was not quite as the TFD would predict. Proponents of the TFD suggest that mathematics teaches students to reason with 'logic', implying that their reasoning should become more normative. While the mathematics students did reason more in line with the normative model of the conditional over the year (the so-called 'material' conditional), the change was better described as a move towards the 'defective' conditional. Under the material interpretation of the conditional, 'if *p* then *q*' is true in all cases except where *p* is true and *q* is false. Under the defective interpretation of the conditional, the conditional is only relevant when *p* is true, so that the conditional is true when *p* and *q* are true, false when *p* and not-*q* are true, and irrelevant when *p* is false.

The reason for the increase in conformity to the defective conditional, as Houston (2009) suggested, may be that many mathematical statements are of the form 'if statement A is true, then statement B is true' (even if heavily disguised), and that A is assumed to be true (even if it is clearly not) and the truth or falsity of B is then deduced. Since the AS level curriculum does not include any explicit reference to conditional logic or the normative model of conditionals, it is plausible that exposure

to implicit ‘if then’ statements, where the antecedent is assumed to be true, could induce a defective interpretation of the conditional, where false antecedent cases are considered irrelevant.

Despite its name, the defective conditional is considered to be psychologically sophisticated (Evans et al, 2007), and so the change could still be viewed as a positive development. In fact, Hoyles and Küchemann (2002) argued that the defective conditional is actually more appropriate for mathematics learning than the material conditional (which leads to normative responses) because students need to understand the consequences of an implication when the antecedent is taken to be true. On the other hand, Durand-Guerrier (2003) argued that a defective interpretation of the conditional hinders students from drawing the valid modus tollens (if *p* then *q*, not *q*, therefore not *p*) inference. Contrary to this, and in support of Hoyles and Küchemann, Inglis and Simpson (2009b) found that a group of undergraduate mathematics students tended to have a more defective than normative interpretation of conditionals, and that this did not in fact prevent them from drawing modus tollens inferences or from being successful in their university-level mathematics examinations.

While Attridge and Inglis’s (2013) findings seemed to support the suggestion that studying mathematics develops at least one type of reasoning skill, the students in their sample had chosen to study mathematics at the post-compulsory level, suggesting that on the whole they enjoyed it or considered it important, and would likely have been engaged with it. If someone were compelled to study a reduced syllabus of mathematics, it is not necessarily the case that they would see the same benefits. Thus an important question is whether the compulsory study of mathematics

at a lower level than that found in the A level curriculum develops similar reasoning skills.

In most educational jurisdictions it is already compulsory to study mathematics until the age of 18 (Hodgen, Pepper, Sturman & Ruddock, 2010; Hodgen, Marks & Pepper, 2013). In Cyprus, students have the option to study mathematics as a low- or high-intensity subject from the age of 16 to 18. This provides a good opportunity to see whether the changes found in the mathematics group in Attridge and Inglis's (2013) study are also seen in groups who study a reduced mathematics syllabus, and who are studying mathematics as a requirement rather than as an option.

Another important difference between our study and that by Attridge and Inglis (2013) is that the mathematics curricula in Cyprus at the 16-18 stage has a substantial logic (deductive geometry) component, whereas the English and Welsh curriculum does not. In fact, Attridge and Inglis could identify no explicit logic component to the English and Welsh A level curriculum. The Cypriot curriculum is discussed in more detail in the methods section.

The study

We aimed to investigate the development of reasoning skills in mathematics students when it is compulsory to study some level of mathematics and where there is an explicit logic component to the curriculum. We used a similar method to that of Attridge and Inglis (2013) but with a different manipulation – rather than comparing mathematics to non-mathematics students, we compared students studying high- or low-intensity mathematics. This allowed us to examine the effect of intensity of mathematics study on the development of reasoning skills, as opposed to the

presence/absence of mathematical study. In this sense our study can be seen as a conceptual replication and extension of Attridge and Inglis's study.

We tested Cypriot students at the beginning, middle and end of their two years of study beginning at age 16. We gave the students two reasoning tasks: the abstract conditional inference task used by Attridge and Inglis (2013) and Inglis and Simpson (2009), and a belief bias syllogisms task (Sá, West and Stanovich, 1999). The syllogisms task provided a measure of syllogistic reasoning and belief bias (the extent to which participants are influenced by the believability of conclusions rather than their logical validity) and was included to extend the breadth of the investigation into more contextualised reasoning. We compared high-intensity and low-intensity mathematics students' changes in performance on the two tasks over the two years of study.

We predicted that reasoning skills on both tasks would be improved to a greater extent in the high-intensity group than in the low-intensity group. If this were the case, it could be due to the higher level of mathematics studied, the greater intensity of mathematics study, or the presumably higher level of motivation to study mathematics given the choice to study it at the high intensity level. On the conditional task, due to a more formal style of deductive geometry in the Cypriot mathematics curriculum, we predicted that the change would be best characterised as an increase in the material interpretation of the conditional, as opposed to the defective interpretation found by Attridge and Inglis (2013). On the syllogisms task, we predicted that syllogisms scores would increase, while belief bias would decrease. A key issue is whether the low-intensity group would improve on any of the measures at all. If they did, it might suggest that electing to study mathematics is not essential for students to benefit from improved reasoning skills.

Method

Design

The study followed a longitudinal quasi-experimental design. Participants were recruited after they had chosen whether to study low-intensity or high-intensity mathematics, and were tested at the beginning of their penultimate academic year (at age 15 or 16), at the end of their penultimate year and again at the end of their final year. They completed the same tasks in each session.

Participants

Participants were 188 students (82 male, 106 female), aged 15-18 at Time 1 (M=16.02, SD=0.42) recruited from a single school. The large-sized school (726 students) was located in an urban area in a large town and had a mixed catchment area containing families of high, medium and low socioeconomic statuses. Seventy-four students were taking high-intensity mathematics and 114 were taking low-intensity mathematics.

Mathematics syllabus

Approximately 40% of students opt for high-intensity mathematics in the participants' school, with the rest taking low-intensity mathematics. The low-intensity mathematics syllabus involves three 45-minute sessions per week in the first year and two in the second year. The high-intensity syllabus involves seven 45-minutes sessions per week in the first year and six in the second year. Of particular interest is the large Euclidean geometry component of the high-intensity syllabus. The school gave copies of Argyropoulos, Vlamos, Katsoulis, Markatis, and Sideris's (2001)

textbook to all high-intensity students, and teachers were also encouraged to use other books they considered appropriate, such as Leptou and Loizou (1991), to meet the curriculum requirements. Argyropoulos et al.'s textbook was the subject of a detailed analysis by Mouzakitis (2006). He described the presentation as following a "traditional Euclidean deductive style, in which the definitions and the enunciations of theorems are not motivated but imposed upon the reader" (p. 15), and suggested that the textbook "emphasises tasks that require a formal proof of some mathematical fact" (p. 17). Formal proof is also present in the non-geometric sections of the high-intensity curriculum. For instance, the 2014 Pancyprian examinations contained a question that asked candidates to define and explain the Mean Value Theorem.

Measures

All measures and instructions were translated from English to Greek by two independent translators, then compared and merged into a third version.

Abstract Conditional Inference. Participants completed the abstract Conditional Inference Task (Evans, Clibbens & Rood's, 1995). The full task, including instructions, is presented in Appendix A. Participants are asked to assess the validity of 32 inferences of four types: modus ponens (MP: if *p* then *q*, *p*, therefore *q*), denial of the antecedent (DA: if *p* then *q*, not *p*, therefore not *q*), affirmation of the consequent (AC: if *p* then *q*, *q* therefore *p*) and modus tollens (MT: if *p* then *q*, not *q*, therefore not *p*). The structure of the inference types is summarised in Table 1. In half of the items, negations in the minor premises were explicit (e.g. 'if the letter is A then the number is 4, the number is not 4, therefore the letter is not A') and in half the negations were implicit (e.g. 'if the letter is A then the number is 4, the number is 8,

therefore the letter is not A'). The lexical content of the rules were generated randomly and the order of the problems was randomised for each participant.

From participants' responses we calculated two interpretation indices for each participant at each time point: a normative conditional index (NCI) and a defective conditional index (DCI). Each participant received a score on each index to represent the number of items they answered consistently with that model of the conditional (Attridge & Inglis, 2013). A person who responded completely normatively would endorse all MP and MT inferences and no DA or AC inferences, while a person conforming perfectly to the defective conditional would endorse all MP inferences and no DA, AC, or MT inferences. The normative and defective responses to each item are also presented in Appendix A.

Thematic Syllogisms. The Belief Bias Syllogisms task (Sá, West & Stanovich, 1999) was used as a measure of the ability to reason independently of prior beliefs. The original 24 items were split into two equivalent 12-item forms (Attridge, 2013) to minimise time demands and to allow us to give participants opposite forms at consecutive testing points. The order of forms was counterbalanced. Each form consisted of 12 contextual syllogisms. Four were congruent – the believability of the conclusion agreed with the logical validity of the syllogism (i.e. either the conclusion was believable and the syllogism valid, or it was unbelievable and the syllogism invalid) – four were incongruent (believable-invalid, unbelievable-valid), and four were neutral (the conclusions had no belief value), see Figure 2. Participants decided whether each syllogism was logically valid or not after being instructed to ignore their prior beliefs. Two measures were taken: a total score out of 12, indicating the extent to which participants responded normatively, and a Belief Bias Index (BBI). The BBI was calculated for each participant by subtracting the number of incongruent items

answered correctly from the number of congruent items answered correctly (Sá, West & Stanovich, 1999). The resulting score indicates the degree to which a person's answers are swayed by believability or validity. The BBI could range from -4 to +4 with positive scores indicating some degree of belief bias.

Raven's Advanced Progressive Matrices (RAPM). To control for differences between groups in general intelligence, we included an 18 item subset of RAPM with a 15 minute time limit (Sá, West & Stanovich, 1999).

Cognitive Reflection Test (CRT). Participants completed the CRT (Figure 3), which prompts intuitive but incorrect responses that participants must inhibit in order to think through the correct responses. Toplak, West and Stanovich (2011) found the CRT to be a better predictor of rational responding to reasoning tasks than cognitive ability, executive functions, or the Actively Openminded Thinking scale. The CRT was included to control for between-groups differences in thinking disposition. The questions were randomly intermixed with three simple mathematical word problems of a similar length from the Woodcock-Johnson III Applied Problems subtest. This was intended to reduce recall of the 'trick' nature of the CRT questions at the second and third tests.

Mathematics Manipulation Check. A 15-item mathematics test was included as a manipulation check (i.e. to see whether the high-intensity group learnt, as one would expect, more mathematics than the low-intensity group, Attridge and Inglis, 2013). Twelve items were taken from the Woodcock-Johnson III Calculation subtest and three items were the most difficult items on the Loughborough University diagnostic test for incoming mathematics undergraduates in 2008 and 2009. The questions were selected to prevent floor and ceiling effects and were presented in a set order that was intended to be progressive.

Procedure

Participants took part in class under examination-style conditions. The tasks were presented in a single paper booklet. Since the RAPM task had a time limit it was always completed first. The order of the subsequent tasks was counterbalanced between-participants following a Latin Square design and participants were instructed to work at their own pace until they had completed all tasks or until the class came to an end. The sessions lasted approximately 40 minutes.

Results

Data cleaning

Of the 188 participants who took part at Time 1 (74 in the high-intensity group and 114 in the low-intensity group), 184 returned at Time 2 and 180 returned at Time 3. However, not all participants completed all tasks at each time point: 107 (44 high-intensity and 63 low-intensity) completed the conditional inference task at all three time points and were included in its analysis, and 124 (50 high-intensity and 74 low-intensity) completed the syllogisms task at all three time points and were included in its analysis. Those who attended but did not complete tasks either missed them out or left them incomplete.

To determine whether these incomplete responses were a threat to validity, we analysed the characteristics of the participants who completed each task at all three time points compared to those who did not. There were no differences between those who completed the conditional inference task at all three sessions and those who did not on: Time 1 Raven's scores, $F < 1$, Time 1 CRT scores, $F < 1$, or Time 1 NCI scores, $F < 1$. Furthermore, there was no significant interaction between Group and

Completion (a factor which stated whether participants completed the conditional inference task or not) for Time 1 Ravens scores, $F(1,204) = 2.11$, $p = .184$, Time 1 CRT scores, $F < 1$, or Time 1 NCI scores, $F(1,146) = 2.28$, $p = .134$.

There were no differences between those who completed the belief bias syllogisms task at all three sessions and those who did not on Time 1 Raven's scores, $F(1,204) = 1.35$, $p = .246$, Time 1 CRT scores, $F < 1$, Time 1 Syllogisms scores, $F(1,162) = 1.13$, $p = .290$, or Time 1 BBI scores, $F < 1$. Furthermore, there was no interaction between Group and Completion for Time 1 Ravens scores, $F < 1$, Time 1 CRT scores, $F < 1$, Time 1 Syllogisms scores, $F < 1$, or Time 1 BBI scores, $F < 1$.

Overall we found no evidence that there was systematic bias in the participants who did or did not complete the tasks, i.e. those students who missed out tasks at Time 2 or 3 were typical of the overall sample.

Covariates

Time 1 scores on Ravens and CRT were significantly positively correlated, $r(148) = .287$, $p < .001$, and the high-intensity students outperformed the low-intensity students on the Ravens, $t(186) = 5.24$, $p < .001$, and CRT, $t(146) = 2.58$, $p = .001$, measures. To investigate changes in each group over time, Ravens and CRT scores were entered into separate 2 (Group: high-intensity, low-intensity) × 3 (Time: 1, 2, 3) ANOVAs (see Table 2 for descriptive statistics).

For Ravens scores, this revealed significant main effects of Time, $F(2,326) = 27.05$, $p < .001$, $n_p^2 = .142$, and Group, $F(1,163) = 41.72$, $p < .001$, $n_p^2 = .204$, but no significant interaction, $F(2,326) = 1.26$, $p = .286$, $n_p^2 = .004$. Overall, scores were higher in the high-intensity group (M = 8.22, SD = 2.55) than in the low-intensity group (M = 5.61, SD = 2.55) and post hoc tests with Bonferroni correction showed

that scores increased between Time 1 (M = 6.03, SD = 2.14) and Time 2 (M = 7.27, SD = 2.36, $p < .001$), and between Time 1 and 3 (M = 7.44, SD = 2.64, $p < .001$) but not between Time 2 and 3 ($p = 1.00$).

For CRT scores, there were significant main effects of Time $F(2,202) = 18.06$, $p < .001$, $n_p^2 = .152$, and Group, $F(1,101) = 11.42$, $p = .001$, $n_p^2 = .102$, but no significant interaction, $F(2,202) = 2.36$, $p = .097$, $n_p^2 = .023$. Overall, scores were higher in the high-intensity group (M = .78, SD = .70) than in the low-intensity group (M = .29, SD = .70) and post hoc tests with Bonferroni correction showed that scores increased between Time 1 (M = .33, SD = .63) and Time 2 (M = .60, SD = .86, $p < .001$), and between Time 1 and 3 (M= .68, SD = .90, $p < .001$) but not between Time 2 and 3 ($p = .565$).

In sum, for both Ravens and CRT scores, the high-intensity group scored higher than the low-intensity group, but there were no significant differences between groups in the extent of improvement over the two years of the study.

Ravens scores at Time 1 were not correlated with change in NCI, DCI, syllogisms scores, or BBI scores between Time 1 and Time 3 (all *p*s > .27), so are not used as covariates in the analyses below. Similarly, CRT scores at Time 1 were not correlated with change in NCI, DCI, syllogisms scores, or BBI scores between Time 1 and Time 3 (all *p*s > .18), so are also not used as covariates in the analyses below[1].

Manipulation check

To investigate whether, as one would expect, the high-intensity students learnt more mathematics than the low-intensity students, mathematics scores were entered into a 2 (Group: high-intensity, low-intensity) × 3 (Time: session 1, session 2, session 3) ANOVA. This showed significant main effects of Time, $F(2,248) = 64.78$, $p <$

.001, n_p^2 = .343, and Group, $F(1,124) = 105.03$, $p < .001$, n_p^2 = .459, and an interaction between Time and Group, $F(2,248) = 40.96$ $p < .001$, n_p^2 = .248 (see Figure 4). The high-intensity mathematics students scored higher than the low-intensity students on the mathematics test at Time 1, $t(124) = 4.52$, $p < .001$, $d = .78$, and scores increased over time in the high-intensity, $F(2,106) = 51.11$, $p < .001$, n_p^2 = .491, and to a lesser extent in the low-intensity group, $F(2,142) = 6.26$, $p = .021$, n_p^2 = .081. Post hoc tests with Bonferroni correction indicated that, in the high-intensity group, scores increased between Time 1 (M = 4.56, SD = 2.63) and Time 2 (M = 6.74, SD = 2.92, $p < .001$), and between Time 2 and Time 3 (M = 8.33, SD = 3.19, $p < .001$). In the low-intensity group, scores increased between Time 1 (M = 2.92, SD = 1.38) and Time 2 (M = 3.46, SD = 1.32, $p = .005$), and between Time 1 and Time 3 (M = 3.32, SD = 1.26, $p = .018$), but not between Time 2 and Time 3 ($p = 1.00$). Means and standard deviations are shown in Table 2.

Overall this analysis suggests that both groups improved on the mathematics test across the two years of the study, with the high-intensity group improving to a greater extent than the low-intensity group, as expected.

Changes in conditional reasoning behaviour

At Time 1, there were no differences between groups in conformity to either the NCI or the DCI (both *p*s > .65). We first conducted a 2 (Index: NCI, DCI) × 2 (Group: high-intensity, low-intensity) × 3 (Time: Session 1, Session 2, Session 3) ANOVA. There was a significant Time × Group interaction, $F(2,210) = 3.953$, $p = .021$, n_p^2 = .036, suggesting that the two groups developed their conditional reasoning behaviour differently. There was no significant Index × Time × Group interaction, $F < 1$, suggesting that the divergence in development between groups was similar for

both NCI and DCI. To explore this in more detail we entered the two indices into separate 2 (Group: high-intensity, low-intensity) × 3 (Time: Session 1, Session 2, Session 3) ANOVAs.

Scores on the NCI, which indexed the extent to which students responded in accordance with the normative material model of the conditional, are shown in Table 2 and Figure 4. On the 2 × 3 ANOVA there was a significant main effect of Time, $F(2,210) = 3.22$, $p = .042$, $n_p^2 = .030$, and an interaction between Time and Group, $F(2,210) = 4.98$, $p = .008$, $n_p^2 = .045$. This interaction was due to a significant main effect of Time within the high-intensity group, $F(2,86) = 6.93$, $p = .002$, $n_p^2 = .139$, but not within the low-intensity group, $F < 1$. This was further investigated with paired sample *t*-tests within each group. The high-intensity group's NCI scores marginally increased between Time 1 and Time 2, $t(43) = 1.89$, $p = .066$, $d = .16$, and between Time 2 and Time 3, $t(43) = 1.97$, $p = .055$, $d = .65$, and significantly increased between Time 1 and Time 3, $t(43) = 3.49$, $p = .001$, $d = .69$. The low-intensity group's scores did not change between any of the time points (all *p*s > .67, *d*s < .24). Although our three-way ANOVA found no significant Index × Time × Group interaction, when we conducted a Time × Group ANOVA on DCI scores, we found no main effects or interactions (all *p*s > .21, *d*s < .43), see Figure 4. Overall, our findings indicate that, although the two groups did not differ in their reasoning behaviour at Time 1, the high-intensity group reasoned more in line with the normative model over time, while the low-intensity group did not change.

Changes in syllogistic reasoning

Total syllogisms scores and BBI scores were entered into separate 2 (Group: high-intensity, low-intensity) × 3 (Time: Session 1, Session 2, Session 3) ANOVAs

(see Table 2 for means and standard deviations). For total syllogisms scores, there was a marginal effect of Group, $F(1, 122) = 3.63$, $p = .059$, $n_p^2 = .029$, where the high-intensity group scored higher (M = 7.05, SD = 1.22) than the low-intensity group (M = 6.62, SD = 1.22). There was no main effect of Time, $F < 1$, and no interaction between Time and Group, $F < 1$.

For BBI scores, there was no main effect of Group, $F < 1$, but there was a main effect of Time, $F(2,238) = 4.52$, $p = .012$, $n_p^2 = .037$. Post hoc tests with Bonferroni corrections indicated that scores did not change between Time 1 (M = 1.55, SD = 1.63) and Time 2 (M = 1.59, SD = 1.56), $p = 1.00$, but that they decreased significantly between Time 2 and Time 3 (M = 1.14, SD = 1.51), $p = .013$, and decreased marginally between Time 1 and Time 3, $p = .062$. There was also an interaction between Time and Group, $F(2,238) = 3.89$, $p = .022$, $n_p^2 = .031$, see Figure 4. This was due to a significant effect of Time within the high-intensity group, $F(2,140) = 6.17$, $p = .003$, $n_p^2 = .112$, but not the low-intensity group, $F(2,140) = 1.49$, $p = .228$, $n_p^2 = .021$. In the high-intensity group, BBI scores did not change between Time 1 (M = 1.84, SD = 1.70) and Time 2 (M = 1.58, SD = 1.44, $p = 1.00$), but they significantly decreased between Time 2 and Time 3 (M = .98, SD = 1.27, $p = .022$), and between Time 1 and Time 3, $p = .007$.

In summary, the high-intensity group scored slightly higher on the syllogisms task than the low-intensity group, but neither group improved over time. However, the high-intensity group did improve on a measure of belief bias, while the low-intensity group did not. In other words, across the two years of the study the high-intensity group became less influenced by the believability of the conclusions of the syllogisms they were asked to evaluate, whereas the low-intensity group did not change.

Discussion

We investigated the development of abstract conditional reasoning skills and contextual syllogistic reasoning skills in 16- to 18-year old students studying high and low-intensity mathematics in a Cypriot school. While the high-intensity mathematics group improved on a measure of normative conditional reasoning, and decreased in belief bias, the low-intensity mathematics group did not. There were no differences between groups in improvement on the intelligence or thinking disposition measures.

We see three potential explanations for the developmental differences we found between the groups. First, it could be due to the high-intensity group studying mathematics at a higher level than the low-intensity group. Second, it could be due to the greater number of hours that the high-intensity group spend studying mathematics. Third, it could be because the low-intensity group chose the low-intensity option because they were less interested and/or successful in mathematics than the group who chose high-intensity mathematics. Our results do not allow us to distinguish between these possibilities, and it is hard to imagine a curriculum where these factors could be distinguished. Testing these explanations would require an experiment in which students are randomly assigned to a high or low level of mathematics study and more or fewer hours of study per week, with their attitudes towards mathematics measured at the start. What our results do allow us to conclude is that in a system where some level of mathematics study is compulsory, not all students will benefit from improved reasoning skills.

Our findings differ from those of Attridge and Inglis (2013) in terms of the nature of the students' development. We found that the Cypriot high-intensity students' change in conditional reasoning behaviour was better described by a change

in the direction of the normative model, whereas in England the change was in the direction of the sophisticated but non-normative defective conditional (Attridge & Inglis, 2013). This may be due to the difference in curricula between the two countries – in England there is no explicit logic component to the mathematics A level curriculum, while in Cyprus there is a substantial deductive geometry component at the equivalent stage. As noted in the methods section, the high-intensity students used a textbook that presented Euclidean geometry in the traditional definition-theorem-proof deductive style. This may account for the more normative change in conditional reasoning behaviour, and also the reduction in belief bias found in the Cypriot students.

Our interpretation – that the difference in direction of development between the Cypriot high-intensity mathematics students and Attridge and Inglis' (2013) mathematics students is due to the differences between the Cypriot and English curricula – is supported by the findings in both studies that there were no differences between groups at the start of the courses. This makes it unlikely that the differences between studies in the direction of development observed were due to the previous educational or cultural experiences of the students.

One potential implication of these results relates to the suggestion that in England and Wales students should continue studying mathematics until the age of 18. There are many reasons put forward in favour of such a policy change, and one of these is the suggestion that studying mathematics develops students' abilities to reason logically (e.g. Smith, 2004; Vorderman, 2011). Our findings suggest that if it were made compulsory for English and Welsh students to study mathematics at the 16 to 18 stage, then their reasoning skills may not improve in the same manner as students who have chosen to study mathematics at this stage. This suggestion rests on

our finding that the Cypriot low-intensity students did not develop on any reasoning measure across the two years of the study, along with the assumption that the low-intensity mathematics students are in some sense equivalent to English non-mathematics A level students (students who would opt out of mathematics in an optional system would presumably chose the low intensity option in a compulsory system), while the Cypriot high-intensity mathematics students, who did improve on the conditional reasoning measure, are in a sense equivalent to English mathematics A level students. However, other factors to consider are the level and intensity of mathematics studied. The low intensity students in Cyprus studied mathematics at a lower level and for fewer hours per week than the high intensity group. Perhaps if these students had been required to study the same mathematics as the high intensity group, they would have improved on the reasoning tasks. The implication is that if students are compelled to study a low level of mathematics post-16, then while their mathematics skills may improve their reasoning skills may not.

A limitation to this study is that it was conducted in only one school, restricting the generalizability of the findings. It would be useful for future research to examine the development of reasoning skills in a more representative sample of schools in Cyprus, and other educational jurisdictions. This would also allow an examination of factors other than intensity that might influence the results, such as class size. Furthermore, a notable proportion of our sample did not fully complete all measures at all three time points. Although we found no evidence that there was any bias to the drop out, it is nevertheless undesirable and may have been biased on factors that we did not measure.

There is now a growing body of evidence suggesting that studying mathematics is associated with improved abstract conditional reasoning skills

(Lehman & Nisbett, 1990; Attridge & Inglis, 2013; the current data). Furthermore it seems that different types of mathematics influence conditional reasoning skills in different ways. It would be fruitful for future research to investigate more systematically which areas of mathematics are associated with what type of change in reasoning behaviour (i.e. material conditional, defective conditional, belief bias, and more real-world reasoning).

Proponents of the TFD believe that studying mathematics improves one's reasoning skills in a useful way. But what relation do the tasks used in this and similar studies have to real-word reasoning activities (cf. Reid & Inglis, 2005)? Although it may seem that the skills required to successfully tackle the tasks used in research studies (i.e. formal, abstract, paper-based) are far removed from real-world thinking, several researchers have argued that this is not the case. Stanovich (2004), for instance, suggested that decontextualised reasoning of the type required to successfully complete the reasoning tasks used here is disproportionately important in post-industrial societies. He argued that "the argument that the laboratory tasks and tests are not like 'real life' is becoming less and less true. 'Life', in fact, is becoming more like the tests!" (p. 124). He supported his position with multiple examples of situations where abstract reasoning skills and following formalized rules are essential to succeed in modern life. For example, choosing a mortgage or insurance policy, determining whether you are eligible to claim benefits, arguing that a company has not fulfilled their contractual obligations and graduating from university.

While it may be the case that we are not often required to deduce conclusions from abstract conditional statements, it may well be the case that the improvement found on such tasks is reflective of a more general improvement in decontextualised logical reasoning skills. Future research would benefit from investigating the effect of

studying mathematics on tasks that more closely resemble real world reasoning contexts, such as those included on the Adult Decision-Making Competence task battery, performance on which has been shown to correlate with a measure of real life decision-making outcomes (Bruine de Bruine, Parker & Fischhoff, 2007).

To conclude, we have shown that studying high (but not low) intensity mathematics at the 16 to 18 stage of Cypriot education is associated with an improvement in conditional reasoning skills and a reduction in belief-based responding. This is a positive finding for proponents of the TFD and for mathematics educators in general.

Endnotes

1. When including RAPM and CRT as covariates, the main effect of Time on NCI scores loses significance, as does the marginal effect of group on BBT scores, but the results are otherwise the same.

References

Argyropoulos, E., Vlamos, P., Katsoulis, G., Markatis, S. & Sideris, P. (2001). *Euclidean Geometry General Lyceum A and B.* Organization Textbook Publishing (OEDB), Athens. (Publication Iota 2010).

Attridge, N. (2013). *Advanced mathematics and deductive reasoning skills: testing the Theory of Formal Discipline.* Unpublished doctoral thesis, Loughborough University, UK.

Attridge, N., & Inglis, M. (2013). Advanced Mathematical Study and the Development of Conditional Reasoning Skills. *PLoS ONE 8*(7): e69399.

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology, 92*, 938-956.

Durand-Guerrier, V. (2003). Which notion of implication is the right one? From logical considerations to a didactic perspective. *Educational Studies in Mathematics, 53,* 5-34.

Evans, J. St. B. T., Clibbens, J. & Rood, B. (1995). Bias in conditional inference: implications for mental models and mental logic. *The Quarterly Journal of Experimental Psychology, 48A,* 644-670.

Evans, J. St. B. T., Handley, S. J., Neilens, H. & Over, D. E. (2007). Thinking about conditionals: a study of individual differences. *Memory & Cognition, 35,* 1772-1784.

Hodgen, J., Pepper, D., Sturman, L., & Ruddock, G. (2010). *Is the UK an outlier? An international comparison of upper secondary mathematics education.* London: Nuffield Foundation.

Hodgen, J., Marks, R., & Pepper, D. (2013). *Towards Universal Participation in*

Post-16 Mathematics: Lessons from High-performing Countries. London: Nuffield Foundation.

Houston, K. (2009). *How to Think Like a Mathematician*. Cambridge: Cambridge University Press.

Hoyles, C. & Küchemann, D. (2002). Students' Understanding of Logical Implication. *Educational Studies in Mathematics, 51,* 193-223.

Inglis, M. & Simpson, A. (2009a). Conditional inference and advanced mathematical study: Further evidence. *Educational Studies in Mathematics, 72,* 185-198.

Inglis, M. & Simpson, A. (2009b). The defective and material conditionals in mathematics: does it matter? In Tzekaki, M., Kaldrimidou, M. & Sakonidis, C. (Eds.). *Proceedings of the 33rd Conference of the International Group for the Psychology of Mathematics Education*, Vol. 3, pp. 225-232. Thessaloniki, Greece.

Lehman, D. R. & Nisbett, R. E. (1990). A longitudunal study of the effects of undergraduate training on reasoning. *Developmental Psychology, 26,* 952-960.

Leptou, E. & Loizou, N. (1991). *Geometry Lessons Lyceum A and B.* The Curriculum Development Unit, Nicosia.

Locke, J. (1971/1706). *Conduct of the Understanding*. New York: Burt Franklin.

Mouzakitis, A. (2006). A comparative analysis of Italian and Greek Euclidean geometry textbooks: A case study. *Philosophy of Mathematics Education Journal, 19.*

Oakley, C. O. (1949). Mathematics. *The American Mathematical Monthly, 56,* 19.

Reid, D. & Inglis, M. (2005). Talking about logic. *For the Learning of Mathematics, 25*(2), 24-25.

Sá, W. C., West, R. F. & Stanovich, K. E. (1999). The domain specificity and

generality of belief bias: Searching for a generalisable critical thinking skill. *Journal of Educational Psychology, 91,* 497-510.

Smith, A. (2004). *Making mathematics count: The report of Professor Adrian Smith's inquiry into post-14 mathematics education.* London: The Stationery Office.

Stanic, G. M. A. (1986). The growing crisis in mathematics education in the early twentieth century. *Journal for Research in Mathematics Education, 17,* 190-205.

Stanic, G. M. A., & Kilpatrick, J. (1992). Mathematics curriculum reform in the United States: A historical perspective. *International Journal of Educational Research, 17,* 407-417.

Stanovich, K. (2004). *The robot's rebellion.* Chicago: The University of Chicago Press, Ltd.

Toplak, M. E., West, R. F. & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition, 39,* 1275-1289.

Vorderman, C. (2011). *A world-class mathematics education for all our young people*. Retrieved from http://www.tsm-resources.com/pdf/VordermanMathsReport.pdf

Walport, M. (2010). *Science and mathematics secondary education for the 21st century: Report of the Science and Learning Expert Group*. London: Crown.

Appendix A

The conditional inference task

This section is concerned with people's ability to reason logically with sentences in various forms. You will be presented with a total of 32 problems on the following pages. In each case you are given two statements together with a conclusion which may or may not follow from these statements.

Your task in each case is to decide whether or not the conclusion *necessarily* follows from the statements. A conclusion is *necessary* if it must be true, given that the statements are true.

Each problem concerns an imaginary letter-number pair and contains an initial statement or rule which determines which letters may be paired with which numbers. An example of a rule of similar form to those used would be:

If the letter is B then the number is not 7.

In each case you must assume that the rule holds and then combine it with the information given in the second statement. This will concern either the letter or the number of an imaginary pair, for example:

The letter is Y.
The number is not 4.

If the information concerns the letter the conclusion will concern the number and vice-versa.

A full problem looks something like:

If the letter is X then the number is 1.
The letter is X.
Conclusion: The number is 1.
☐Yes
☐No

If you think the conclusion necessarily follows then please tick the YES box, otherwise tick the NO box. Please work through the problems in order and make sure you do not miss any. Do not return to a problem once you have finished and moved on to another.

Item	Inference	Normative	Defective
If the letter is A then the number is 3. The letter is A. The number is 3.	MP	Yes	Yes
If the letter is T then the number is 5. The letter is not T. The number is not 5.	DA	No	No
If the letter is F then the number is 8. The number is 8. The letter is F.	AC	No	No
4. If the letter is D then the number is 4. The number is not 4. The letter is not D.	MT	Yes	No
If the letter is G then the number is not 6. The letter is G. The number is not 6.	MP	Yes	Yes
If the letter is R then the number is not 1. The letter is not R. The number is 1.	DA	No	No
If the letter is K then the number is not 3. The number is not 3. The letter is K.	AC	No	No
If the letter is U then the number is not 9. The number is 9. The letter is not U.	MT	Yes	No
If the letter is not B then the number is 5. The letter is not B. The number is 5.	MP	Yes	Yes
If the letter is not S then the number is 6. The letter is S. The number is not 6.	DA	No	No

If the letter is not V then the number is 8. The number is 8. The letter is not V.	AC	No	No
If the letter is not H then the number is 1. The number is not 1. The letter is H.	MT	Yes	No
If the letter is not F then the number is not 3. The letter is not F. The number is not 3.	MP	Yes	Yes
If the letter is not L then the number is not 9. The letter is L. The number is 9.	DA	No	No
If the letter is not J then the number is not 8. The number is not 8. The letter is not J.	AC	No	No
If the letter is not V then the number is not 7. The number is 7. The letter is V.	MT	Yes	No
If the letter is D then the number is 2. The letter is D. The number is 2.	MP	Yes	Yes
If the letter is Q then the number is 1. The letter is K. The number is not 1.	DA	No	No
If the letter is M then the number is 4. The number is 4. The letter is M.	AC	No	No
If the letter is V then the number is 5. The number is 2. The letter is not V.	MT	Yes	No

If the letter is S then the number is not 8. The letter is S. The number is not 8.	MP	Yes	Yes
If the letter is B then the number is not 3. The letter is H. The number is 3.	DA	No	No
If the letter is J then the number is not 2. The number is 7. The letter is J.	AC	No	No
If the letter is U then the number is not 7. The number is 7. The letter is not U.	MT	Yes	No
If the letter is not E then the number is 2. The letter is R. The number is 2.	MP	Yes	Yes
If the letter is not A then the number is 6. The letter is A. The number is not 6.	DA	No	No
If the letter is not C then the number is 9. The number is 9. Conclusion: The letter is not C.	AC	No	No
If the letter is not N then the number is 3. The number is 5. The letter is N.	MT	Yes	No
If the letter is not A then the number is not 1. The letter is N. The number is not 1.	MP	Yes	Yes
If the letter is not C then the number is not 2. The letter is C. The number is 2.	DA	No	No

If the letter is not W then the number is not 8. The number is 3. The letter is not W.	AC	No	No
If the letter is not K then the number is not 1. The number is 1. The letter is K.	MT	Yes	No

Table 1. The four inferences (modus ponens, denial of the antecedent, affirmation of the consequent and modus tollens) with and without negated premises (Prem) and conclusions (Con).

	MP		DA		AC		MT	
	Prem	Con	Prem	Con	Prem	Con	Prem	Con
if *p* then *q*	*p*	*q*	not-*p*	not-*q*	*q*	*p*	not-*q*	not-*p*
if *p* then not-*q*	*p*	not-*q*	not-*p*	*q*	not-*q*	*p*	*q*	not-*p*
if not-*p* then *q*	not-*p*	*q*	*p*	not-*q*	*q*	not-*p*	not-*q*	*p*
if not-*p* then not-*q*	not-*p*	not-*q*	*p*	*q*	not-*q*	not-*p*	*q*	*p*

Table 2. Mean scores on each task in each group at the three time points, with standard deviations in parentheses.

Task	Group	Time 1	Time 2	Time 3
RAPM	High	6.81 (2.42)	8.11 (3.20)	8.02 (2.95)
	Low	5.09 (3.01)	5.84 (3.10)	6.43 (3.42)
CRT	High	.50 (.81)	.86 (.96)	.97 (.97)
	Low	.16 (.45)	.33 (.75)	.39 (.80)
Mathematics	High	4.56 (2.63)	6.74 (2.92)	8.33 (3.19)
	Low	2.92 (1.38)	3.46 (1.32)	3.32 (1.27)
NCI	High	15.57 (2.37)	16.43 (2.40)	17.32 (2.73)
	Low	16.06 (2.03)	15.94 (2.69)	15.87 (2.01)
DCI	High	13.43 (3.63)	13.48 (4.04)	14.45 (4.36)
	Low	13.71 (4.42)	12.63 (4.79)	12.98 (4.39)
Syllogisms	High	7.04 (1.32)	7.08 (1.59)	7.02 (1.88)
	Low	6.62 (1.69)	6.55 (1.74)	6.69 (1.34)
BBI	High	1.84 (1.70)	1.58 (1.44)	.98 (1.27)
	Low	1.25 (1.53)	1.59 (1.61)	1.30 (1.62)

Figure 1. Example items from the Conditional Inference task, a) Modus Tollens and b) Denial of the Antecedent.

If the letter is S then the number is 6	If the letter is M then the number is 4
The number is not 6	The letter is not M
Conclusion: The letter is not S	Conclusion: The number is not 4
a) Modus Tollens	b) Denial of the antecedent

Figure 2. Examples of each item type from the Belief Bias Syllogisms task.

Believable valid:	Believable invalid:
All fish can swim	All living things need water
Tuna are fish	Roses need water
Tuna can swim	Roses are living things
Unbelievable valid:	Unbelievable invalid:
All things with four legs are dangerous	All guns are dangerous
Poodles are not dangerous	Rattlesnakes are dangeours
Poodles do not have four legs	Rattlesnakes are guns
Neutral valid:	Neutral invalid:
All ramadions taste delicious	All lapitars wear clothes
Gumthorps taste delicious	Podips wear clothes
Gumthorps taste delicious	Podips are lapitars

Figure 3. The three-item Cognitive Reflection Task.

1. A bat and a ball costs €1.10 in total. The bat costs €1 more than the ball. How much does the ball cost?
2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?
3. In a lake there is a patch of lily pads. Every day the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?

Intuitive answers: Q1 = 10 cents, Q2 = 100 minutes, Q3 = 24 days.

Correct answers: Q1 = 5 cents, Q2 = 5 minutes, Q3 = 47 days.

Figure 4. Changes in scores on the Conditional Inference. Belief Bias, and mathematics measures in each group over time. Error bars represent ±1 standard error of the mean.