Plan

1. Does studying advanced mathematics develop general reasoning skills?
2. Short break: have a go at the question on your sheet!
3. Using comparative judgement to improve mathematics teaching and learning.
4. A demonstration of the NoMoreMarking system.

Does studying advanced mathematics develop general reasoning skills?

Matthew Inglis
Royal Society Worshipful Company of Actuaries Research Fellow
Mathematics Education Centre, Loughborough University

Loughborough University

Plan

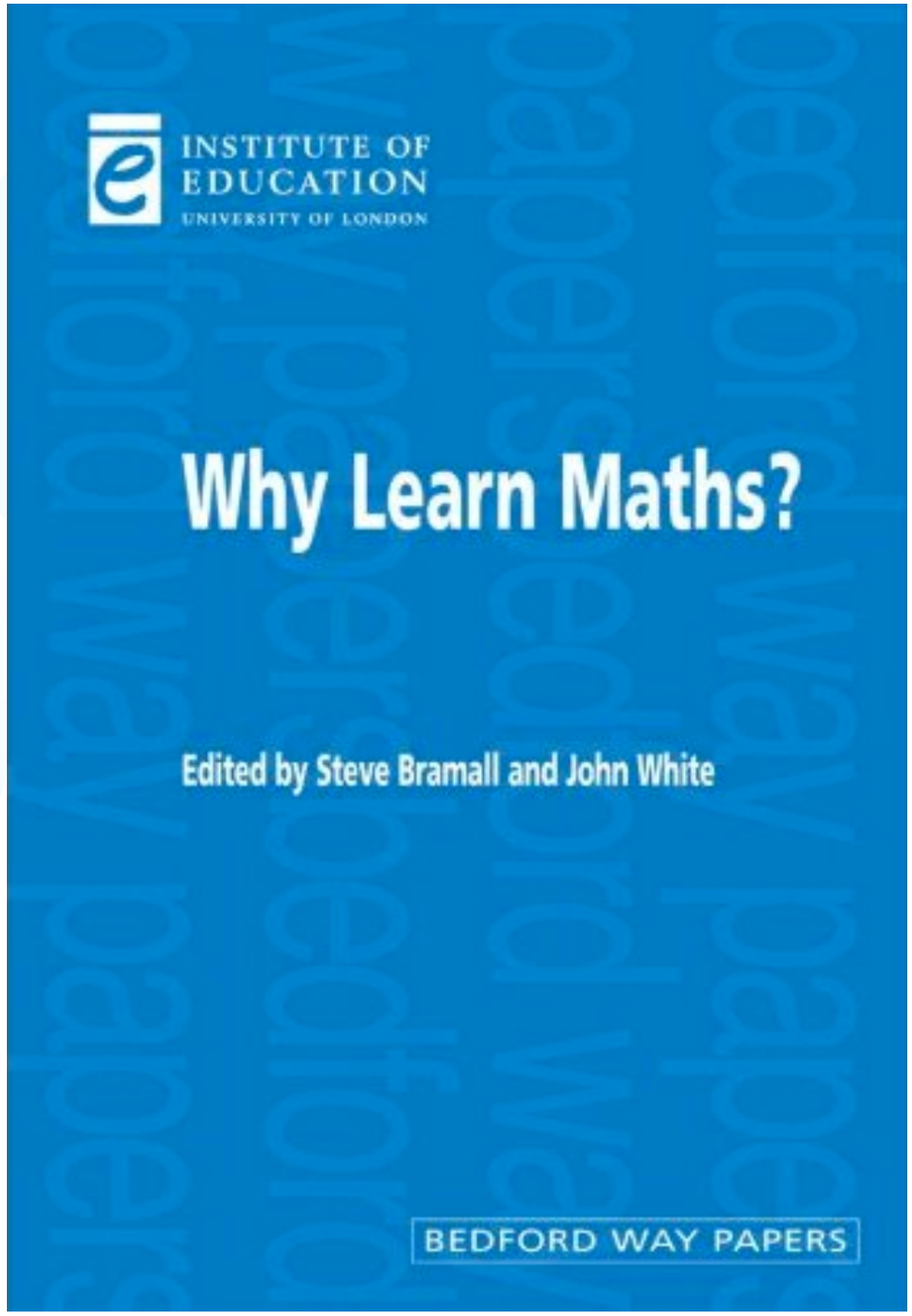
- Why should people study mathematics?
  - The Plato/Vorderman Hypothesis: *Theory of Formal Discipline.*
  - Reasons to doubt the value of mathematics.
- Do mathematicians reason differently to non-mathematicians?
- Is this developmental?

Why Study Mathematics?

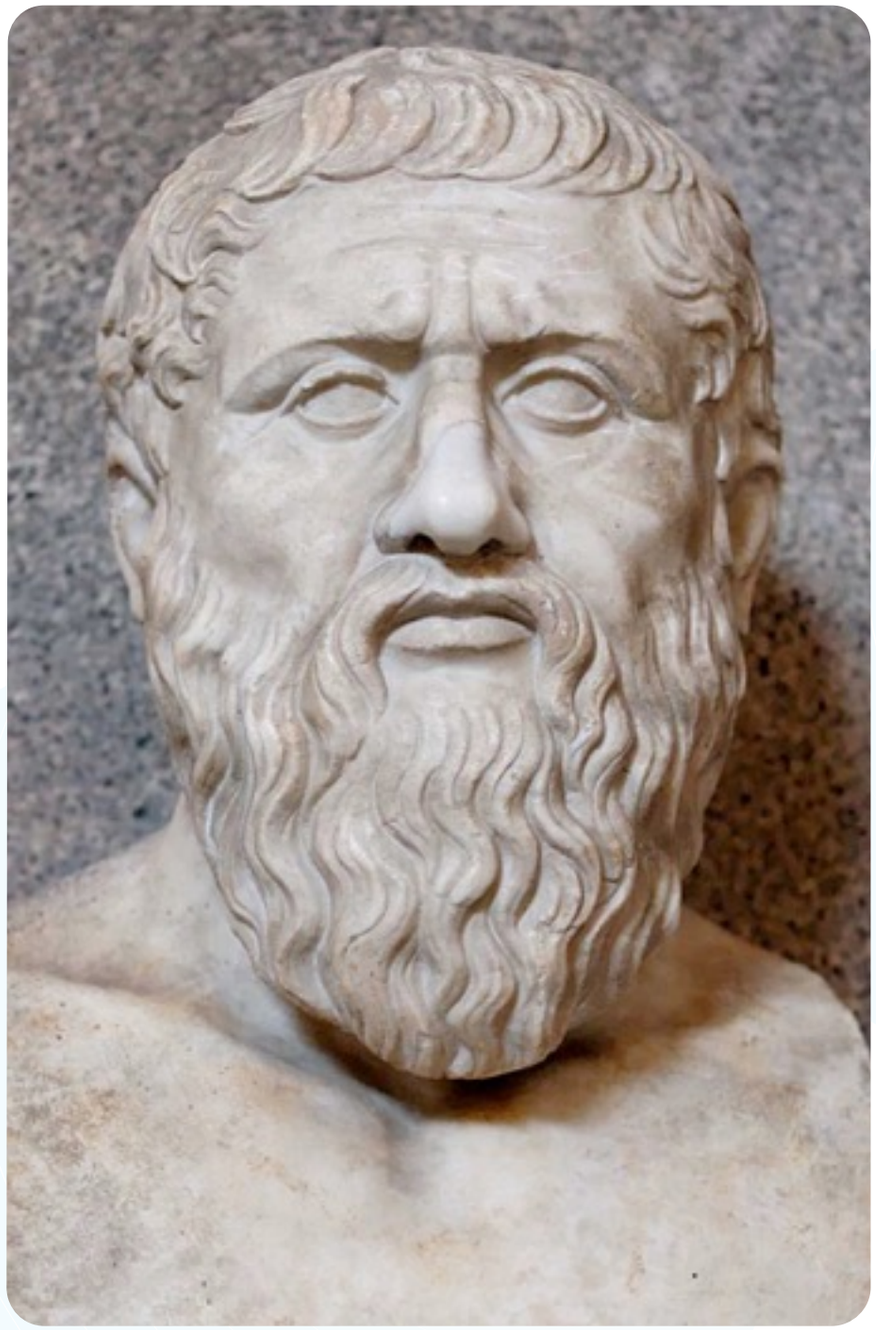
Mathematics has a privileged place on the school curriculum. Why? Two traditional reasons:

1. It's useful in real life
2. It teaches you to think

Focus of talk:
The Theory of Formal Discipline



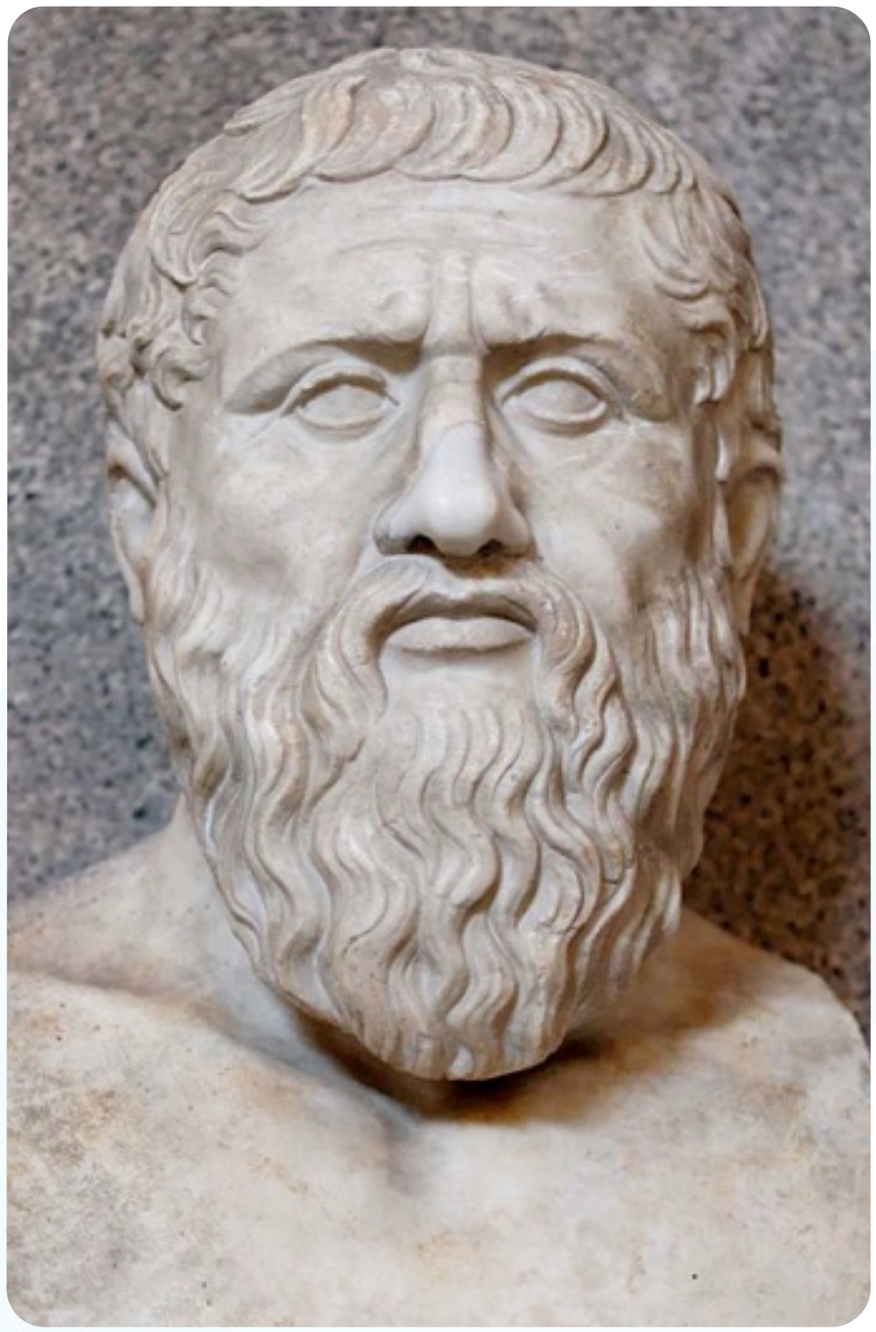
Why Study Mathematics?



Plato (400 BC):

“Those who have a natural talent for calculation are generally quick at every other kind of knowledge; and even the dull, if they have had an arithmetical training... become much quicker than they would otherwise have been.”

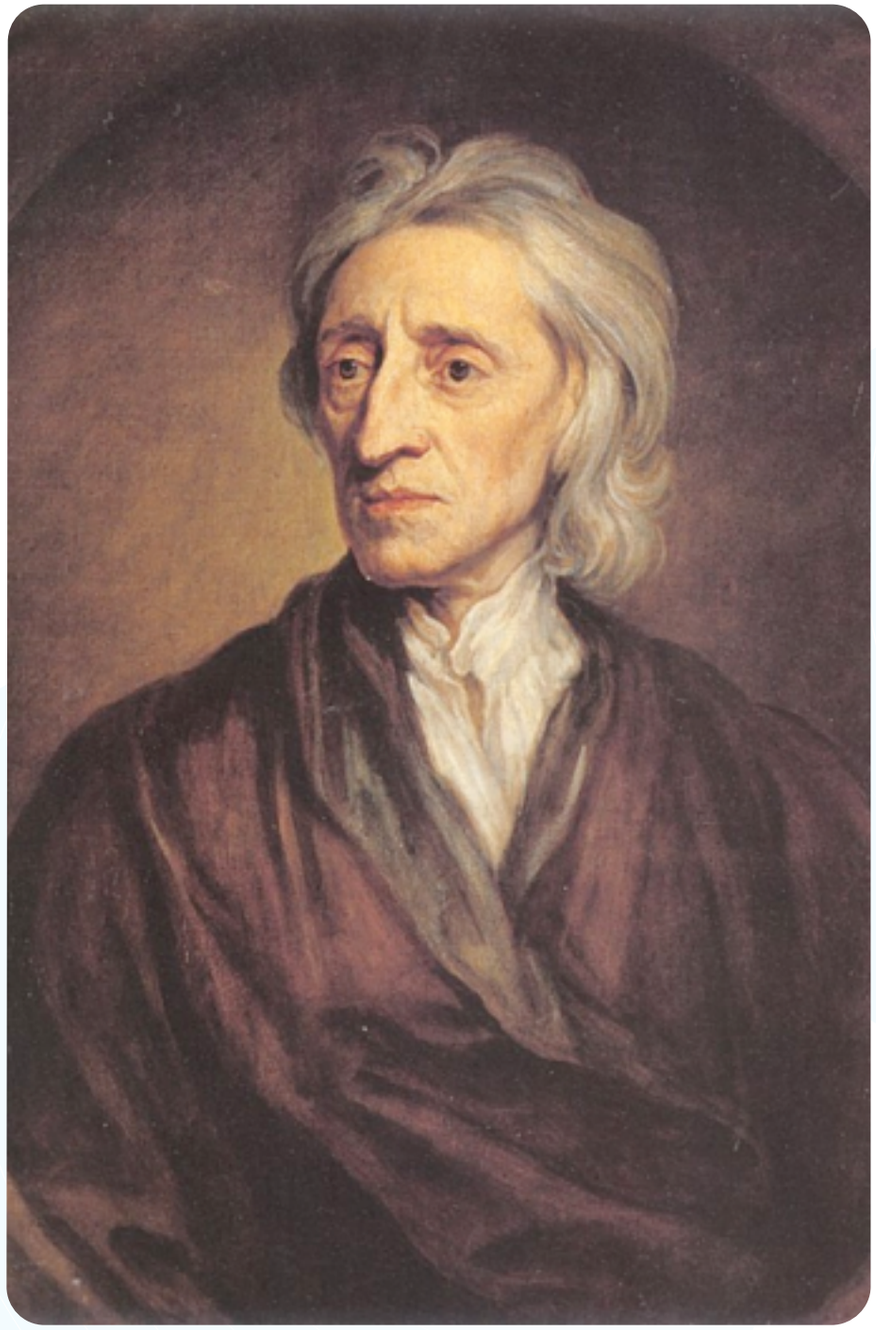
Why Study Mathematics?



Plato (400 BC):

“We must endeavour to persuade those who are to be the principal men of our state to go and learn arithmetic”

Why Study Mathematics?



John Locke (1706):

Mathematics ought to be taught to “all those who have time and opportunity, not so much to make them mathematicians as to make them reasonable creatures”

Why Study Mathematics?

Isaac Watts (1752)

“If we pursue mathematical Speculations, they will inure us to attend closely to any Subject, to seek and gain clear Ideas, to distinguish Truth from Falsehood, to judge justly, and to argue strongly”

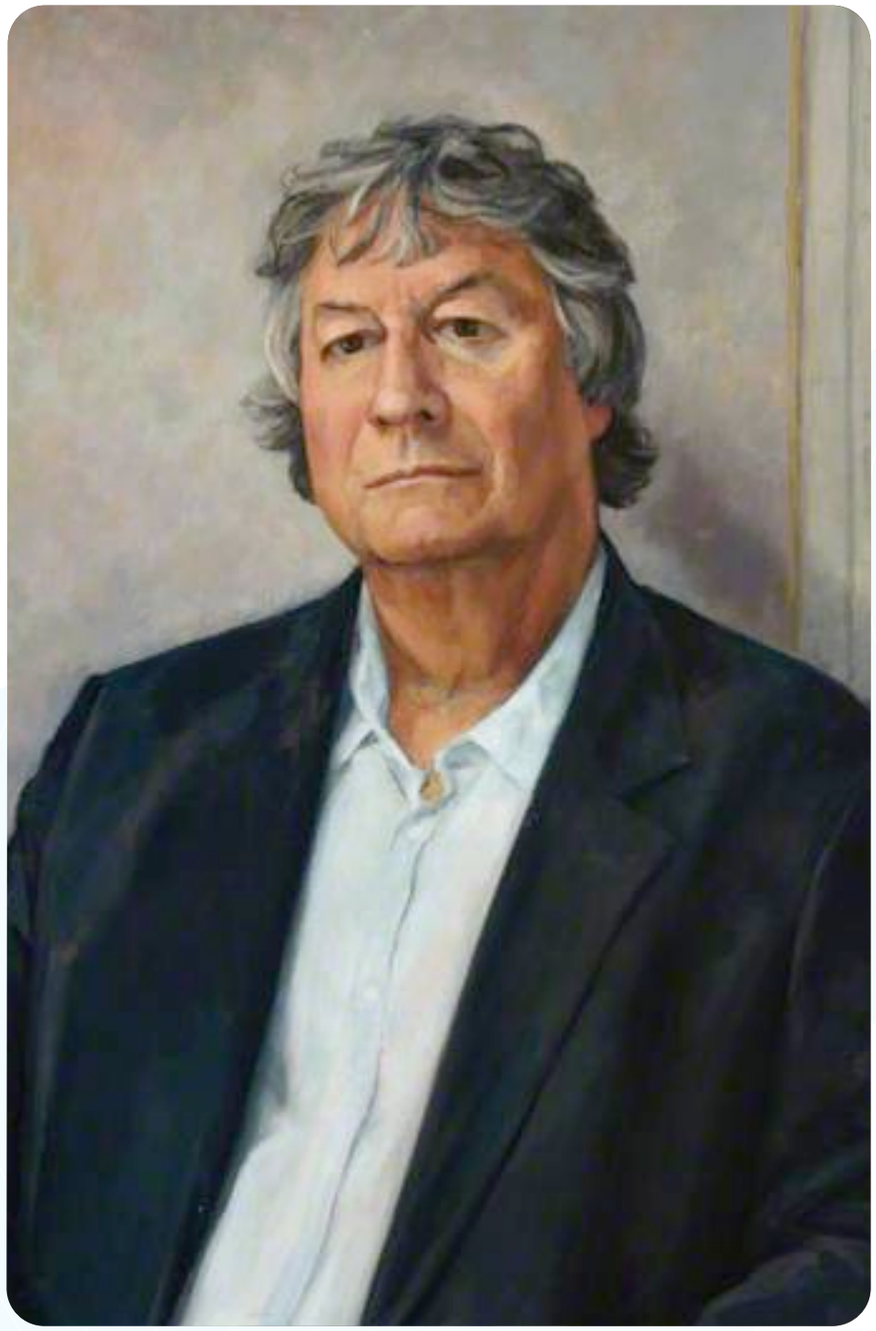
Theory of Formal Discipline

Features of the Theory of Formal Discipline:

- Studying mathematics develops general reasoning abilities, which apply to non-mathematical areas of life;
- This link is causal.

Not just of historical interest.

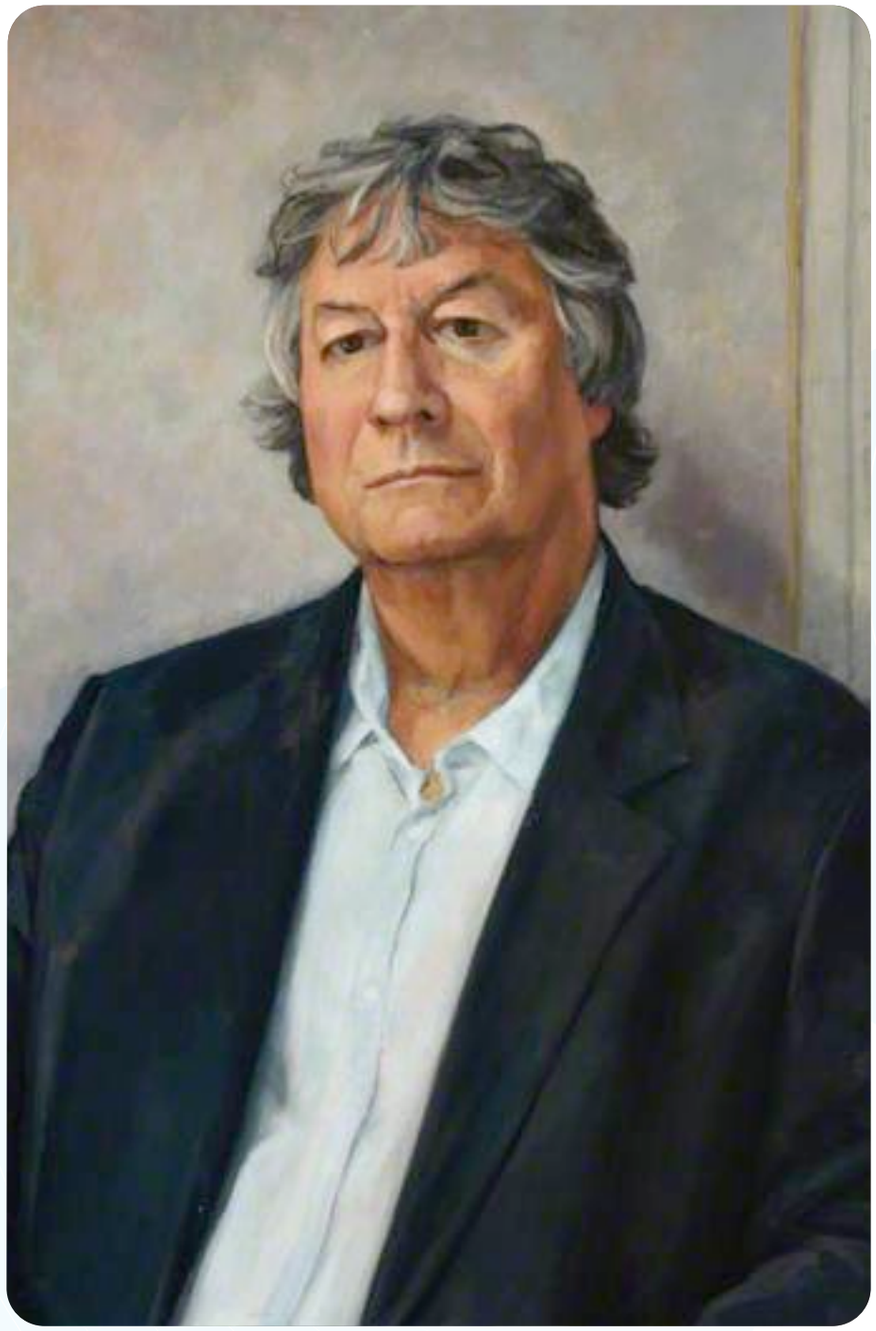
Why Study Mathematics?



Professor Adrian Smith (Smith Report, 2004):

“Mathematical training disciplines the mind, develops logical and critical reasoning, and develops analytical and problem-solving skills to a high degree.”

Why Study Mathematics?



The Smith Report recommended tuition fee rebates for mathematics students, and higher salaries for mathematics teachers.

Why Study Mathematics?

Vorderman Report commissioned by the Conservative Party:

“Mathematics is not only a language and a subject in itself, but it is also critical in fostering logical and rigorous thinking”

Obvious Question

- Mathematicians are **incredibly** good at arguing for the importance of their subject.
  [Compare to psychology: “Psychology, law and media studies: the ‘scandalous’ routes to A-grade success”, *The Independent*, August 2003].

- But notice that none of these advocates offered **any** scientific evidence at all.

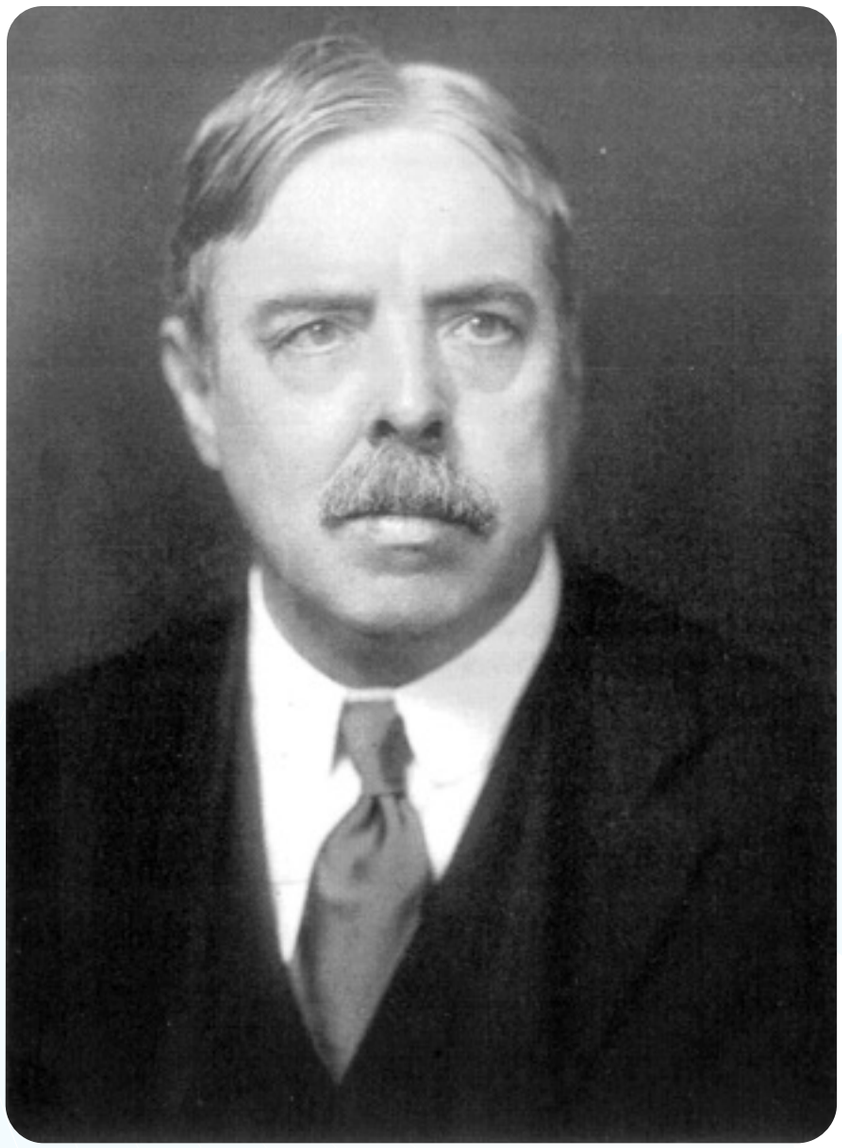
- So is the Theory of Formal Discipline correct?

- It could be that those who choose to study mathematics are already better at reasoning: the *filtering hypothesis*.

Obvious Question

- Does studying mathematics cause the development of general reasoning skills?
- In fact (limited) empirical evidence does exist.

Thorndike & Woodworth



Edward Thorndike
(1874 - 1949)

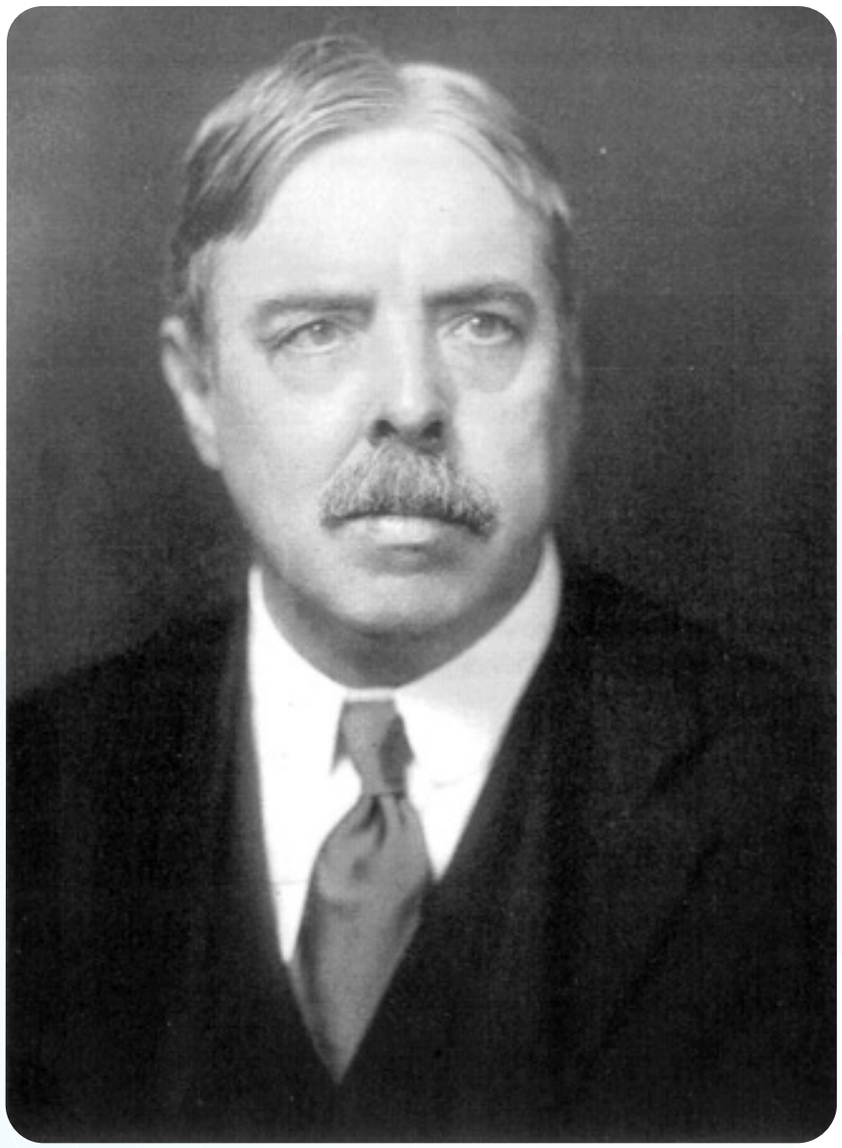
THE INFLUENCE OF IMPROVEMENT IN ONE MENTAL FUNCTION UPON THE EFFICIENCY OF OTHER FUNCTIONS. (I.)

BY DR. E. L. THORNDIKE,
Teachers College, New York,

AND DR. R. S. WOODWORTH,
New York University Medical School.

This is the first of a number of articles reporting an inductive study of the facts suggested by the title. It will comprise a general statement of the results and of the methods of obtaining them, and a detailed account of one type of experiment.

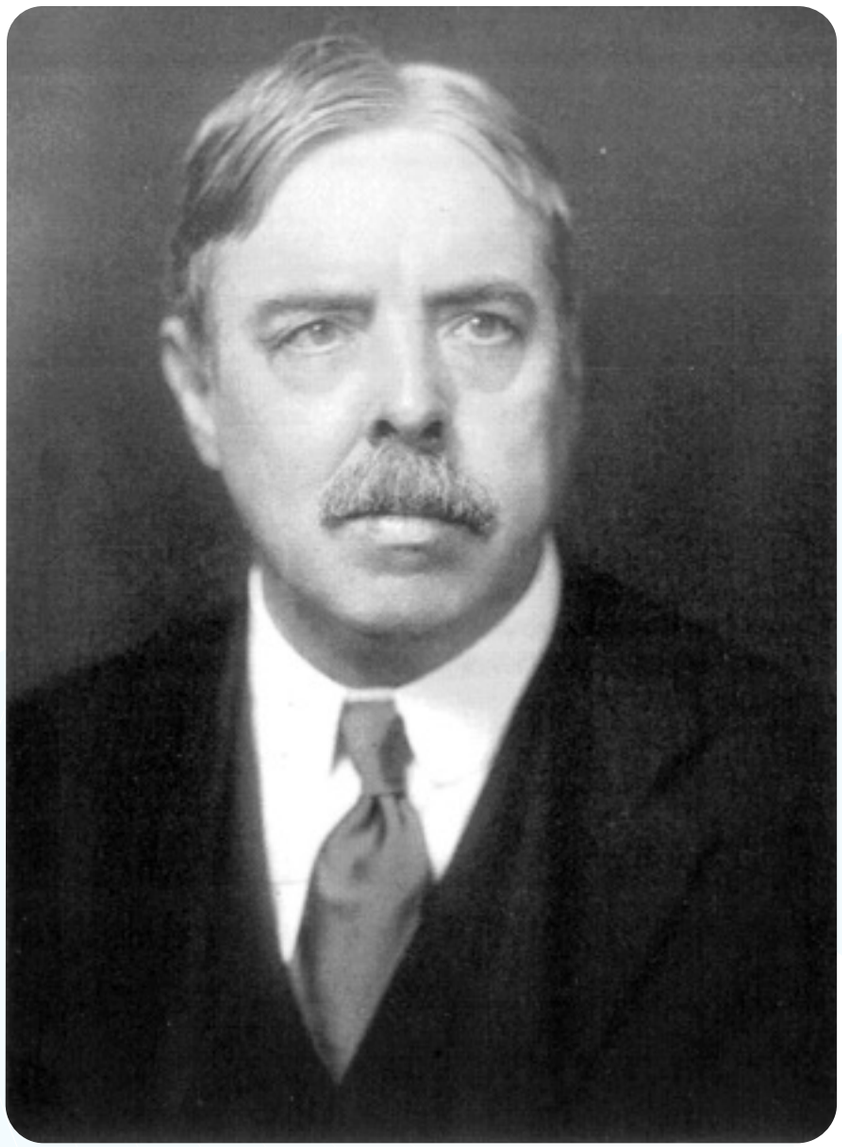
Thorndike & Woodworth



Edward Thorndike
(1874 - 1949)

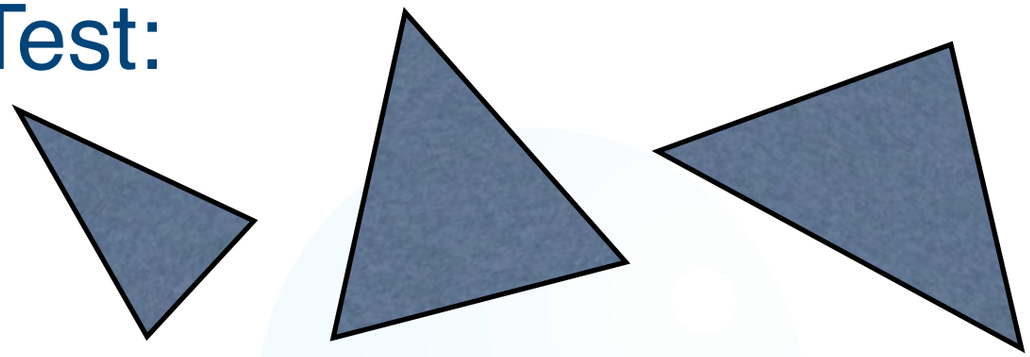
Edward Thorndike investigated the extent to which training on mental function X improves the closely related mental function Y.

Thorndike & Woodworth

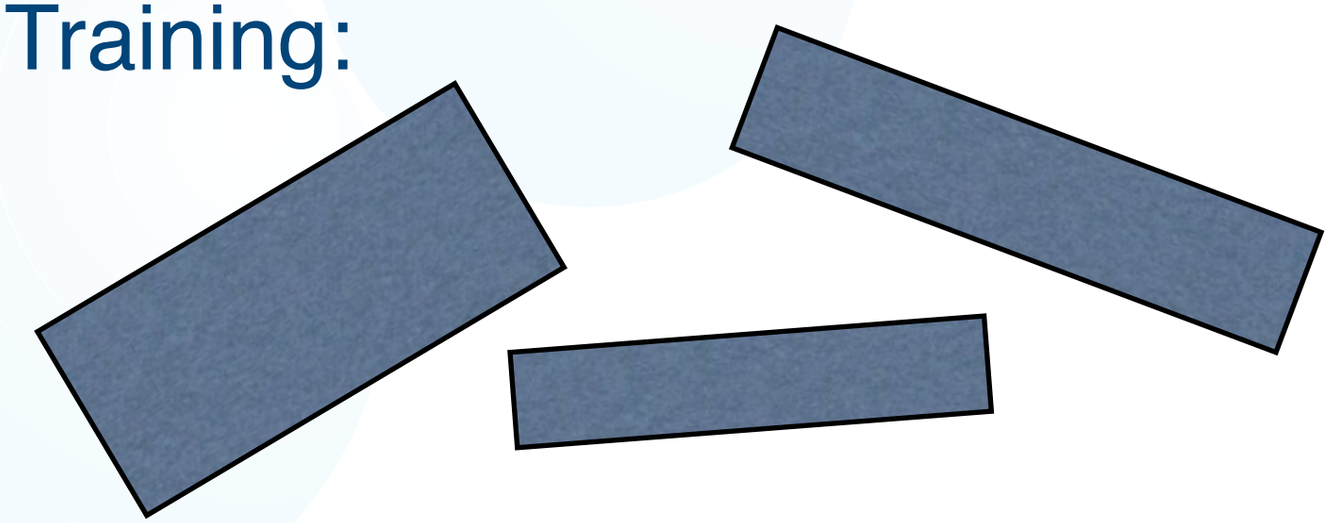


Edward Thorndike
(1874 - 1949)

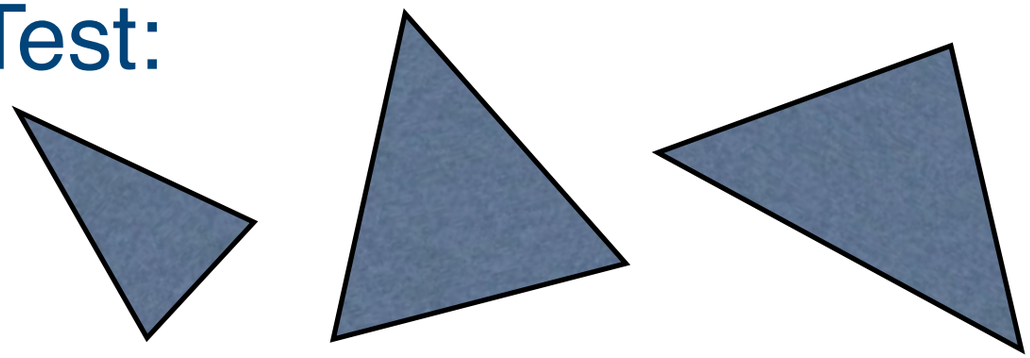
Test:



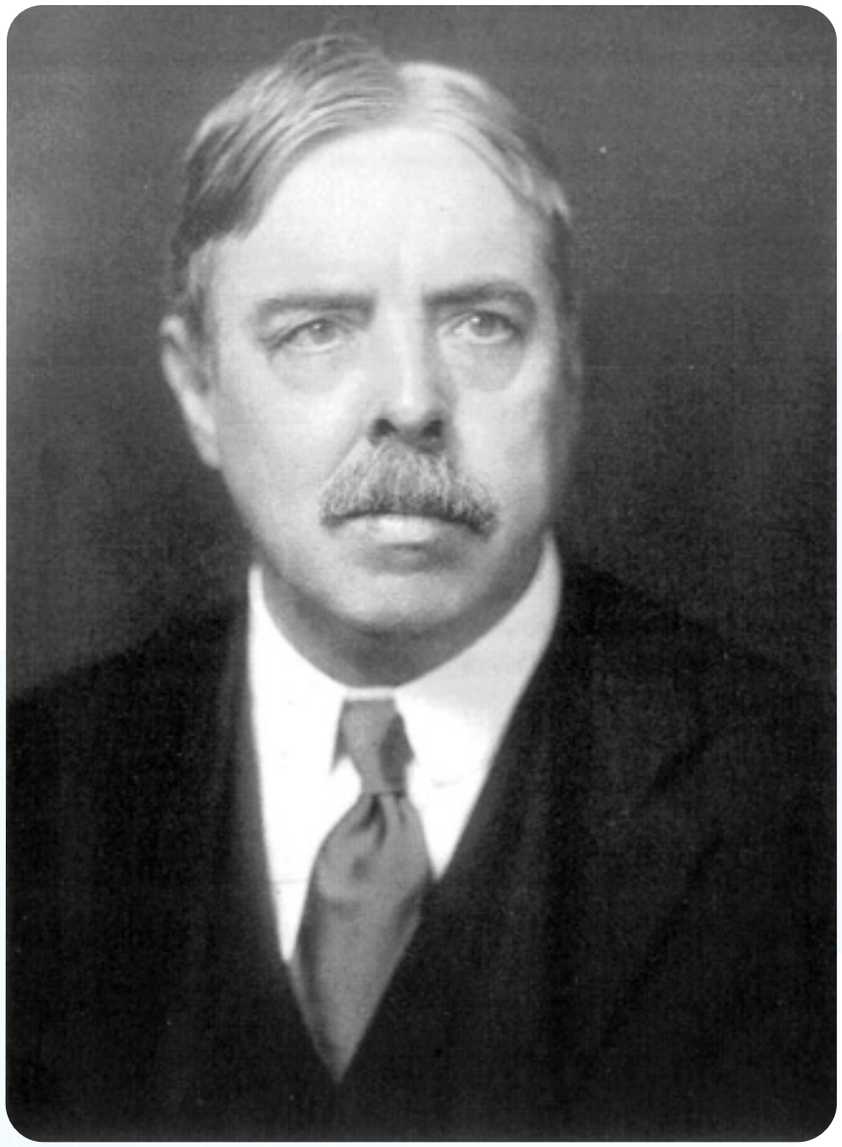
Training:



Test:



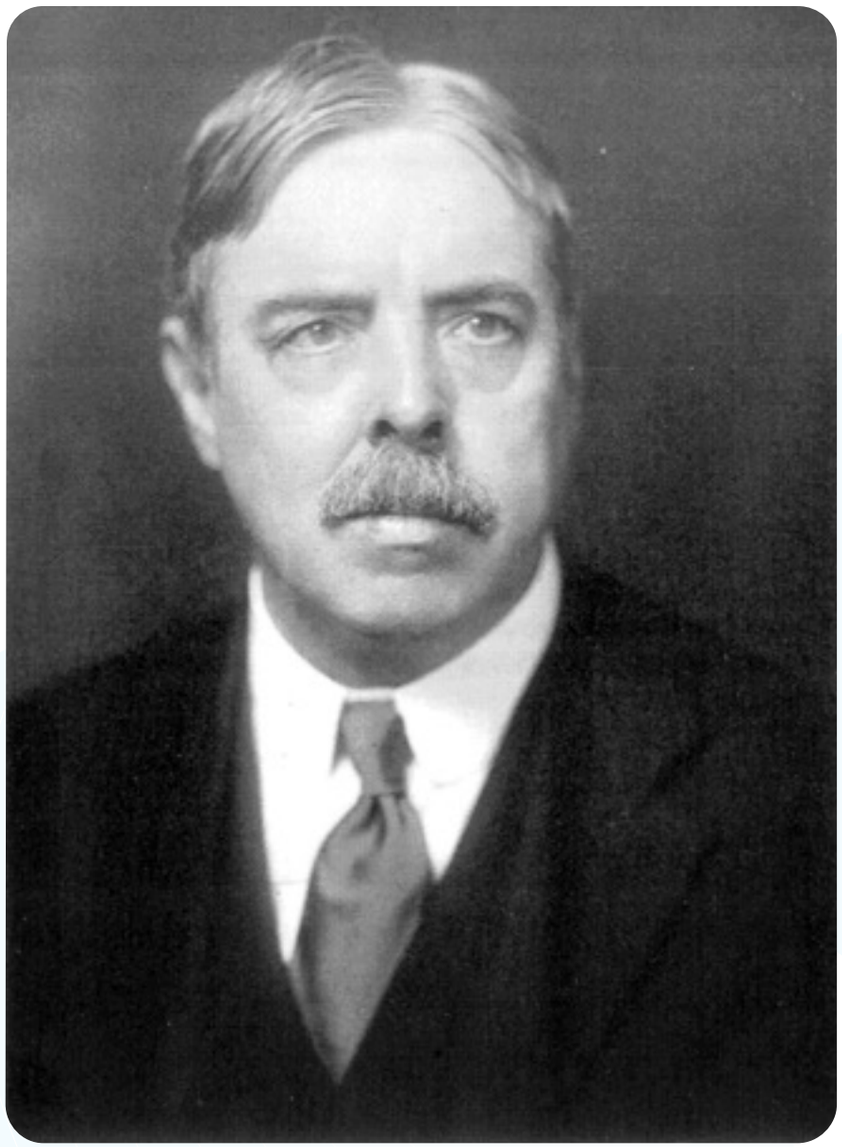
Thorndike & Woodworth



Edward Thorndike
(1874 - 1949)

“Improvement in any single mental function rarely brings about equal improvement in any other function, no matter how similar, for the working of every mental function-group is conditioned by the nature of the data in each particular case.”

Thorndike



Edward Thorndike
(1874 - 1949)

What about formal schooling?

THE JOURNAL OF
EDUCATIONAL PSYCHOLOGY

Volume XV January, 1924 Number 1

MENTAL DISCIPLINE IN HIGH SCHOOL STUDIES[1]

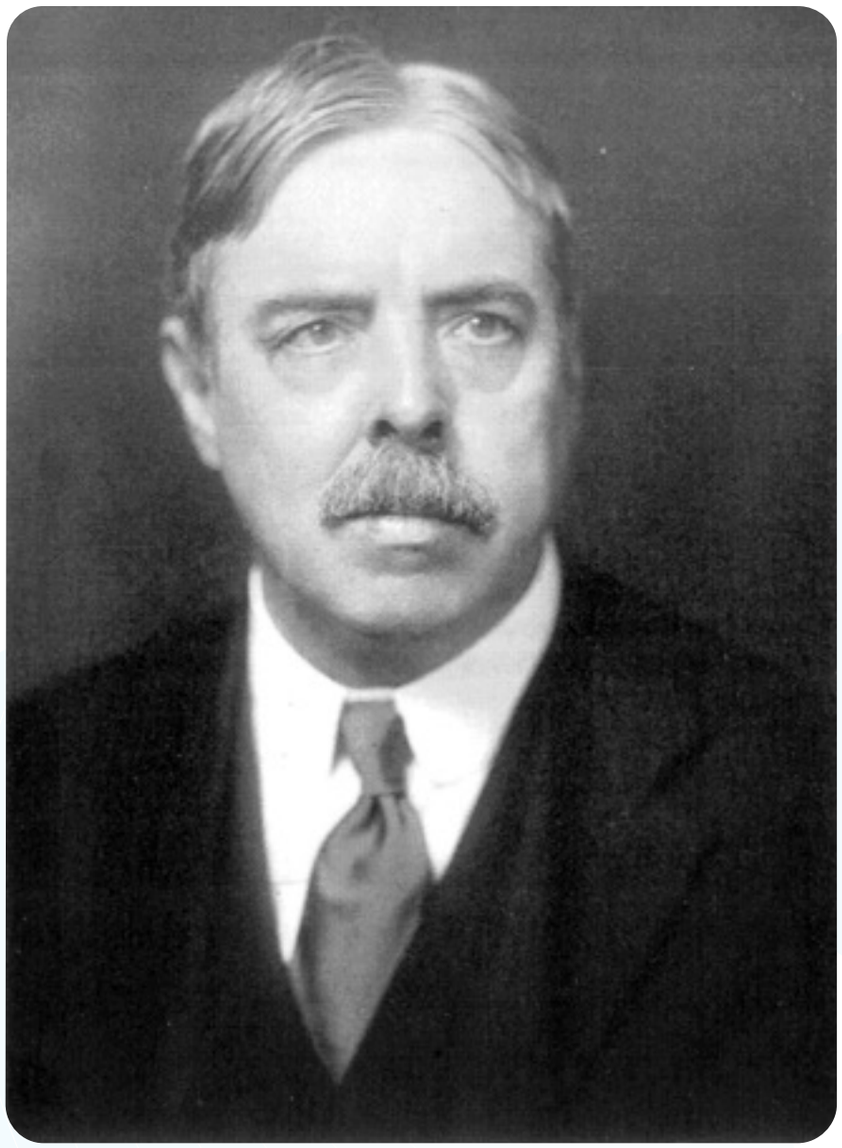
E. L. THORNDIKE

With the aid of the staff of The Institute of Educational Research,

Teachers College, Columbia University

The experiment to be reported consisted of an examination in May, 1922, and a reexamination in May, 1923, of 8564 pupils who, in May, 1922, were in grades IX, X and XI. The two examinations were alternative forms of a composite of tests of "general intelligence" that are in common use, plus certain ones added in order to have measures with spatial as well as verbal and numerical content. This composite examination is that described in Vol. V, No. 4 of the *Journal of Educational Research*, April, 1922. Each pupil who took both examinations recorded the subjects which he studied during the school year Sept. 22, 1922 to June 23, 1923; and the gains made in the test were put into relation with the subjects studied. For example, we compare the gains for the pupils who studied English, history, geometry and Latin during the year with the gains for the pupils who studied English, history, geometry and shop-work. If other factors

Thorndike



Edward Thorndike
(1874 - 1949)

Selected Findings:

Subject	"Regression Coefficient"
French	+ 0.48
Bookkeeping	+ 0.25
Arithmetic	+ 0.13
Geometry	+ 0.13
Algebra	+ 0.12
Drawing	– 0.01
Economics	– 0.50
Sewing	– 0.66

Critique of Thorndike

Lev Vygotsky
(1896 - 1934)

Vygotsky suggested that Thorndike's "general intelligence" measure wasn't sensitive enough to measure developmental changes in reasoning skills.

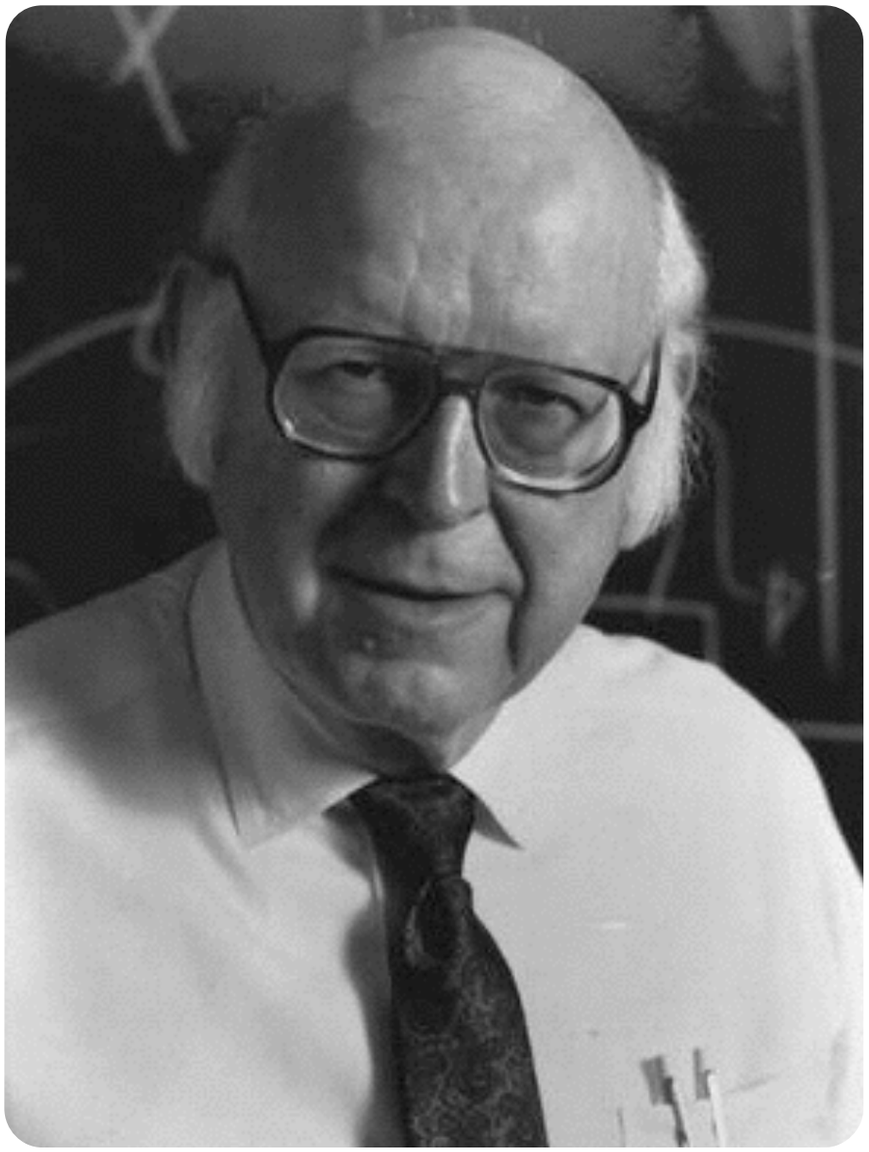
Piaget

Jean Piaget
(1896 - 1980)

Piaget argued that domain-independent thinking skills did exist, but that they couldn't be taught.

You just have to wait until the child is ready to enter the "stage of formal operations". You can do nothing at all to help.

The Cognitive Revolution



Alan Newell

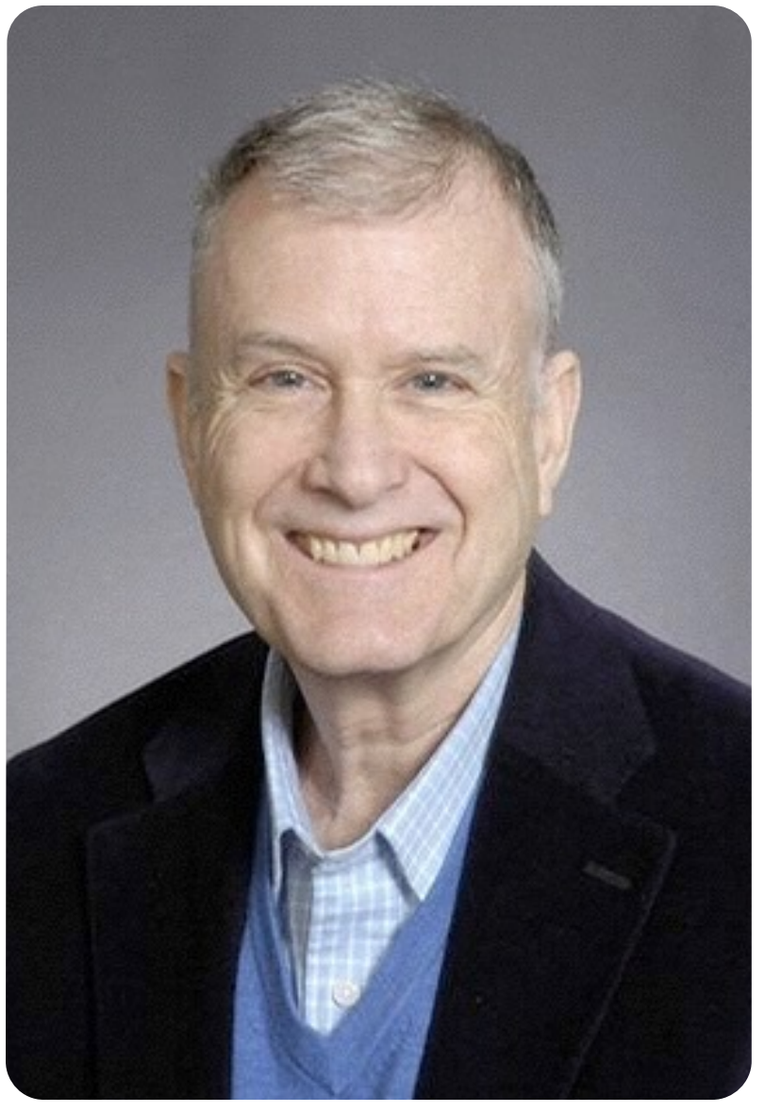
Following the cognitive revolution, most cognitive scientists rejected Piaget's claims.

Newell wrote:
"The modern position is that learned problem-solving skills are, in general, idiosyncratic to the task."

Bad news for Plato/Vorderman: mathematics cannot develop domain-general skills, as they don't exist!

Newell, A. (1980). One last word. In D. Tuma and F. Reif (Eds.) *Problem Solving and Education*, Hillsdale, NJ: Erlbaum.

Studying Psychology Improves Thinking

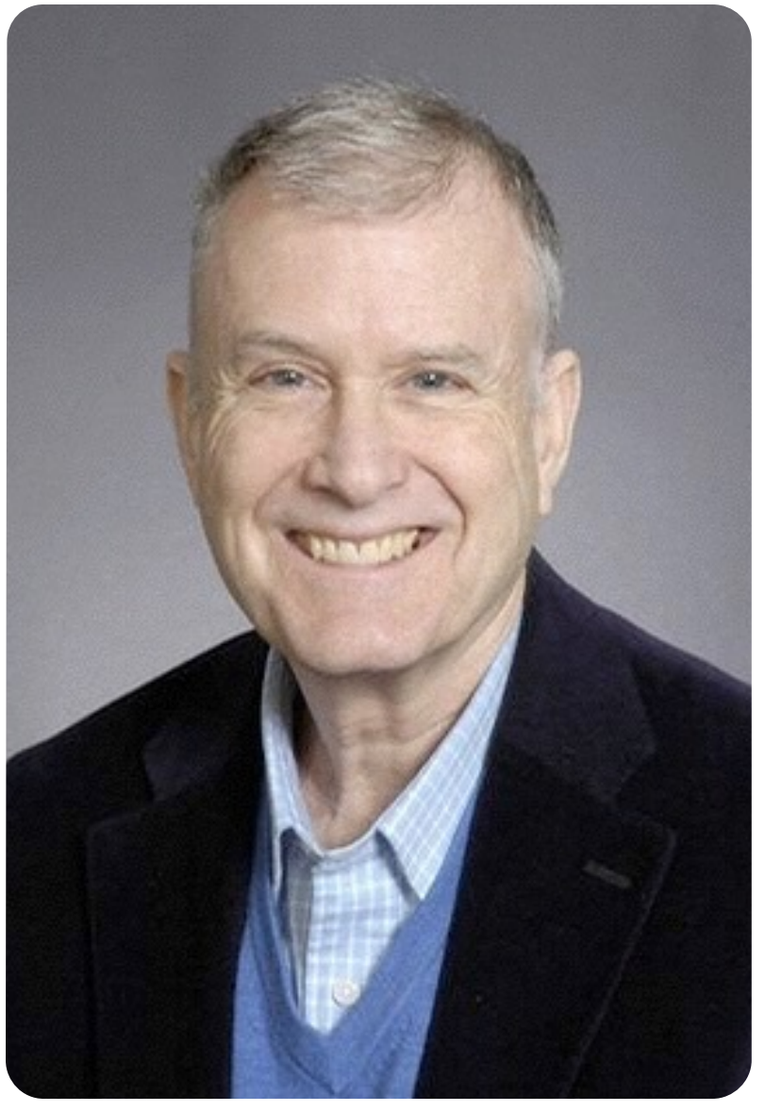


Richard Nisbett

However, more recently Richard Nisbett has found that **some** domain-independent thinking skills **do** exist, and that these can be taught.

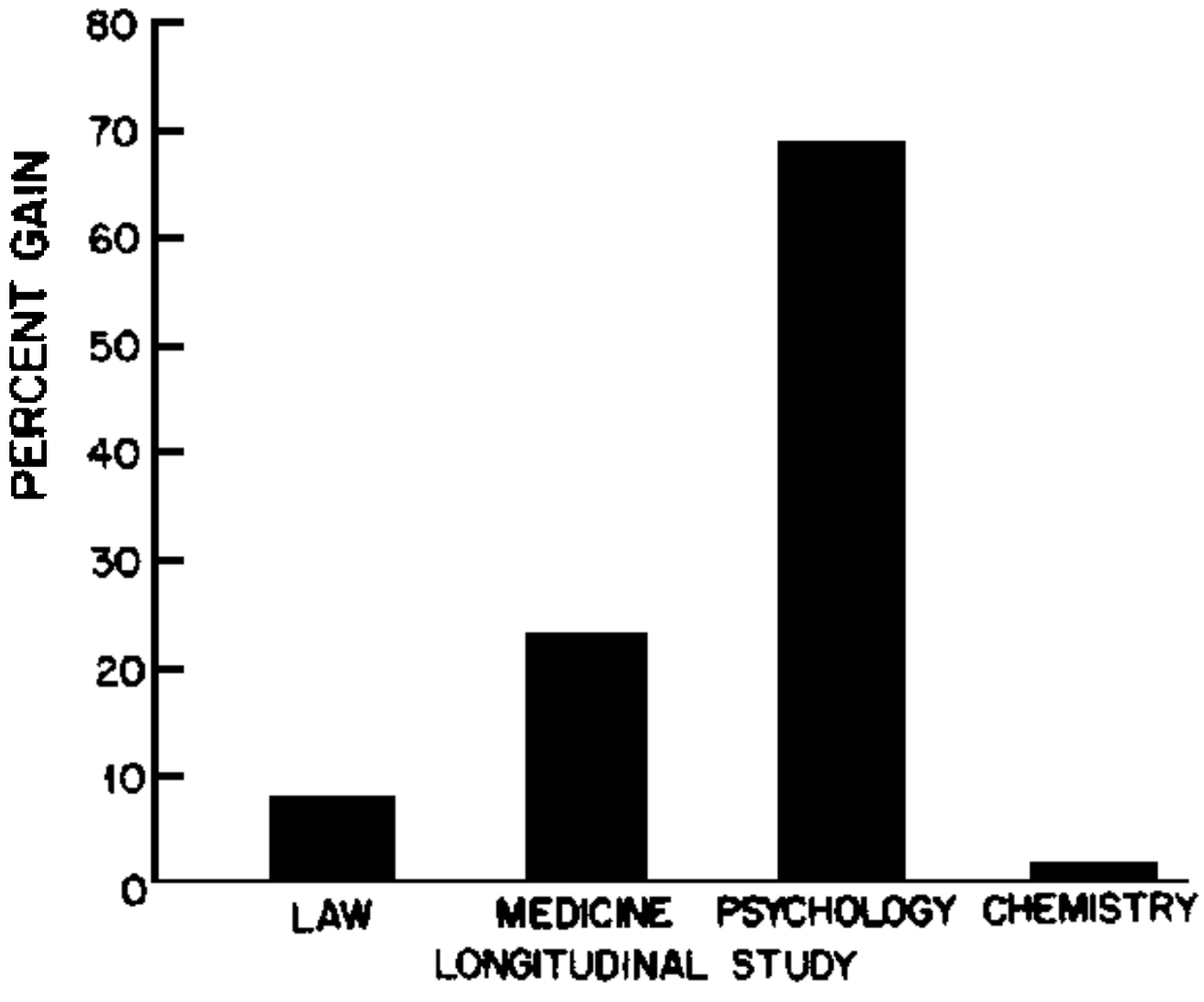
In particular, he has shown that studying psychology makes you better at “statistical and methodological reasoning”. Not so for law or chemistry.

Studying Psychology Improves Thinking

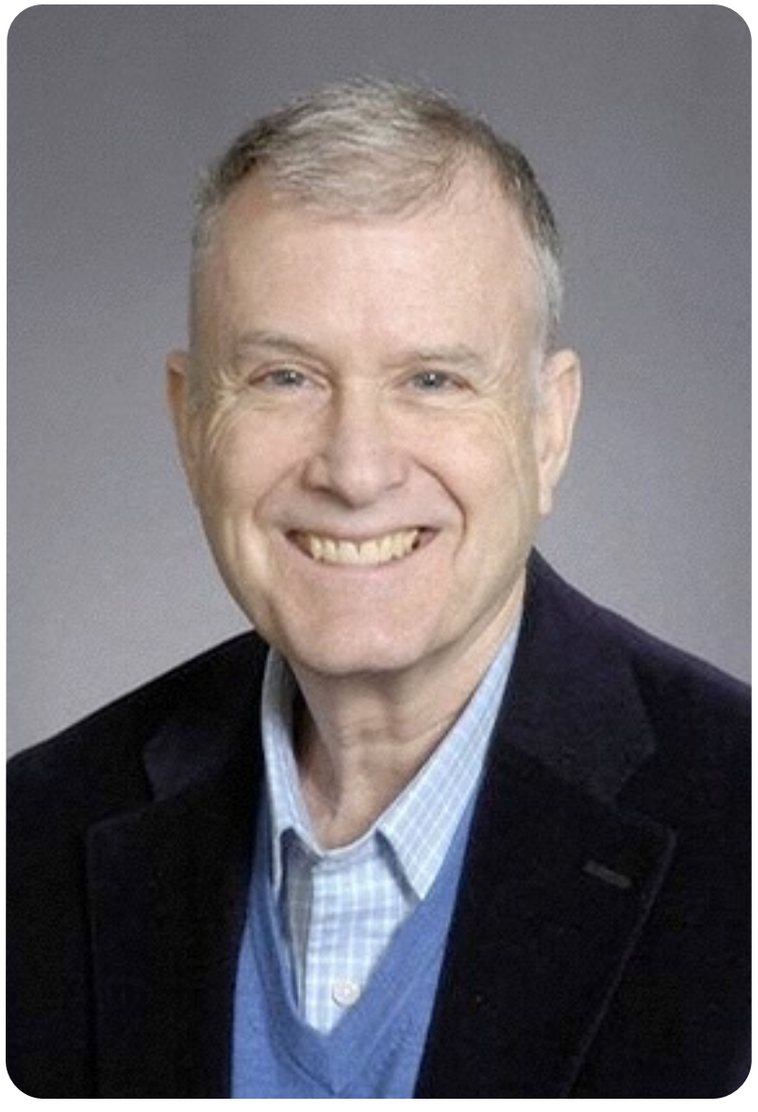


Richard Nisbett

Changes in “Statistical and Methodological Reasoning” across three years of graduate school in Michigan

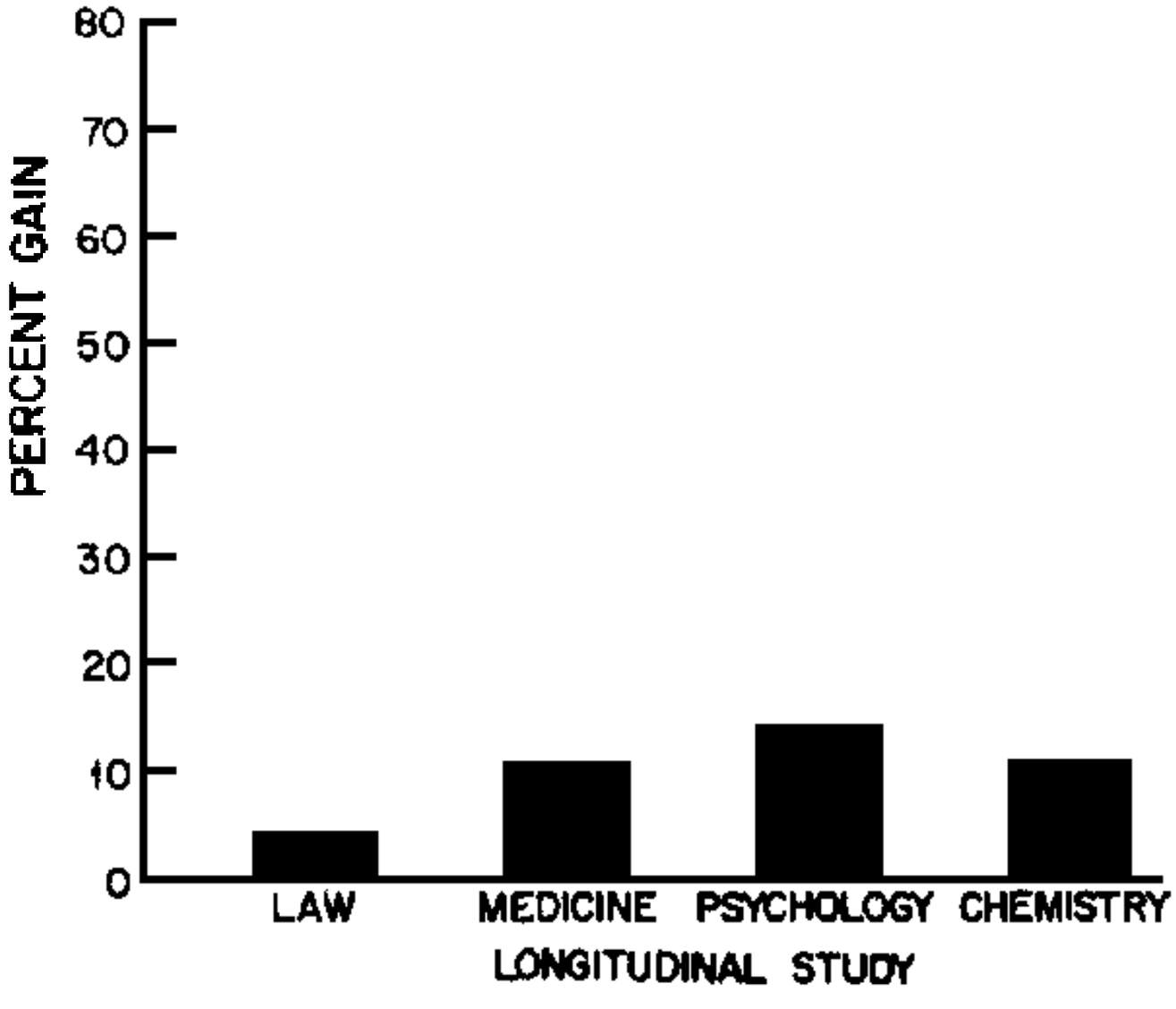


Not the Case for Deductive Logic



Richard Nisbett

Changes in “Verbal Reasoning” across three years of graduate school in Michigan



Not the Case for Deductive Logic

Patricia Cheng even showed that studying a full course in formal logic doesn't improve one's abilities to tackle logic tasks.

(there may be methodological issues with this… see Attridge, Aberdein & Inglis, in press)

COGNITIVE PSYCHOLOGY 18, 293–328 (1986)

Pragmatic versus Syntactic Approaches to Training Deductive Reasoning

PATRICIA W. CHENG

Carnegie–Mellon University

KEITH J. HOLYOAK

University of Michigan

AND

RICHARD E. NISBETT AND LINDSAY M. OLIVER

University of Michigan

Two views have dominated theories of deductive reasoning. One is the view that people reason using syntactic, domain-independent rules of logic, and the other is the view that people use domain-specific knowledge. In contrast with both of these views, we present evidence that people often reason using a type of

Vol 465|10 June 2010|doi:10.1038/nature09042

nature

LETTERS

Putting brain training to the test

Adrian M. Owen[1], Adam Hampshire[1], Jessica A. Grahn[1], Robert Stenton[2], Said Dajani[2], Alistair S. Burns[3], Robert J. Howard[2] & Clive G. Ballard[2]

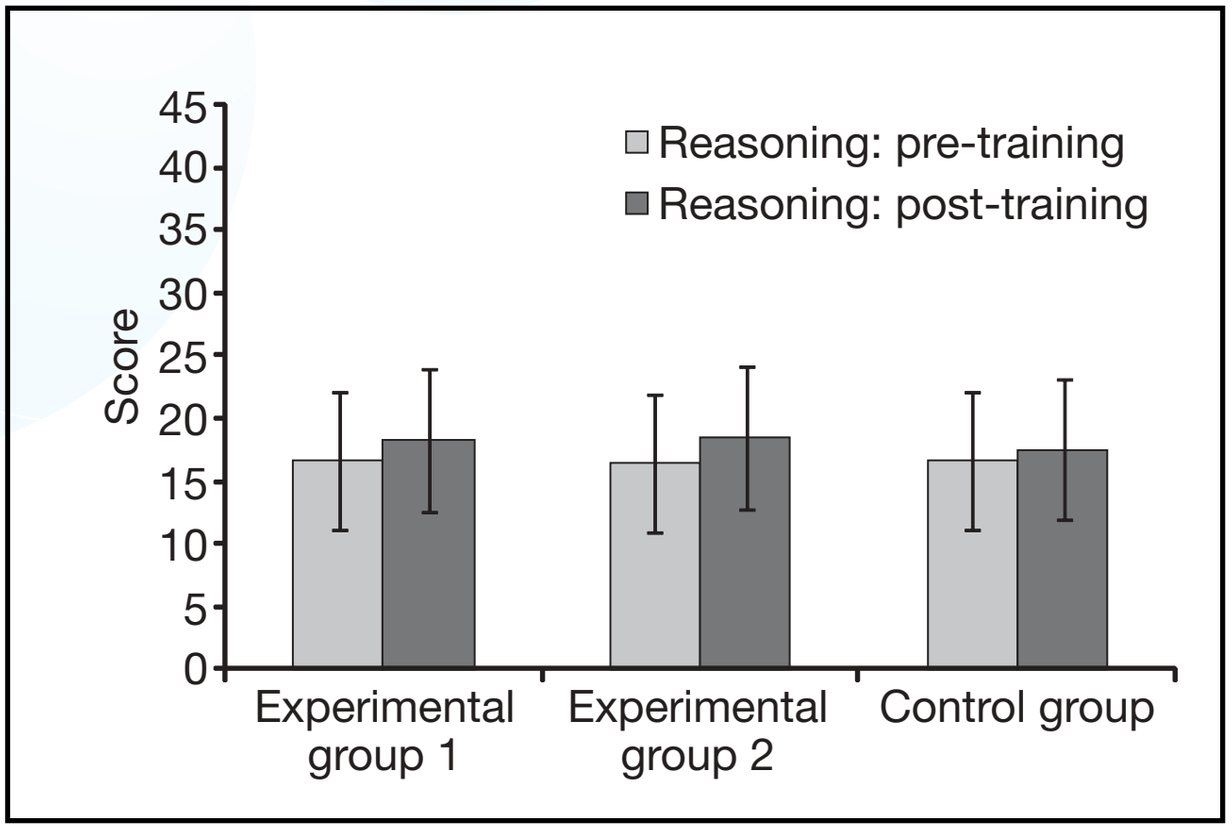
'Brain training', or the goal of improved cognitive function through the regular use of computerized tests, is a multimillion-pound industry[1], yet in our view scientific evidence to support its efficacy is lacking. Modest effects have been reported in some studies of older individuals[2,3] and preschool children[4], and video-game players outperform non-players on some tests of

broader range of cognitive functions was trained using tests of short-term memory, attention, visuospatial processing and mathematics similar to those commonly found in commercially available brain-training devices. The difficulty of the training tasks increased as the participants improved to continuously challenge their cognitive performance and maximize any benefits of training. The control group

Collaboration with BBC's "Bang Goes the Theory"

N = 11,430

Used 'brain training' for six weeks.



Background Summary

1. Overwhelming view among mathematicians and policy-makers is that studying mathematics **causally** develops general reasoning skills.
2. Overwhelming view among psychologists is that it does not (or, if you're Nisbett, that it does not develop *logical* reasoning skills, but might develop other non-logical reasoning skills).
3. Very little direct empirical evidence either way.

Background Summary

- This situation is a bit of a mess.
- Clearly unsatisfactory that important educational policy decisions are being made on anecdotal evidence.
- Main goal of the Fellowship, funded by the Worshipful Company of Actuaries via the Royal Society, was to provide some compelling evidence either way.

Research Strategy

1. How can we measure reasoning performance?
2. Do mathematicians “reason differently” to non-mathematicians?
3. Are such differences developmental?
4. Does the curriculum matter?

How can we measure reasoning performance?

Measuring Reasoning

- What reasoning skills do TFD proponents think studying mathematics develops?
- When asked, people say things like “logic, critical thinking, problem solving…”
- But I wanted to pin them down to making specific predictions.
- First I conducted a literature review to identify tasks that seem to be related to the kinds of skills Plato and Vorderman talk about.

Measuring Reasoning

I interviewed a series of “stakeholders” to ask them their views:

- Presidents of learned societies;
- MPs associated with education;
- Mathematicians involved in influencing curriculum development;

I showed them a series of reasoning tasks and asked them to predict the extent to which studying mathematics would help.

I insisted they made specific predictions (1-5 scale).

Measuring Reasoning

Task	Median
Argument Evaluation Task	4
Belief Bias Syllogism Task	5
Cognitive Reflection Task	4
Conditional Inference Task	5
Evaluation of Arguments	3.5
Interpretation of Arguments	4
Recognition of Assumptions	4
Estimation	4.5
Insight Problem Solving	2
Statistical Reasoning	4
Wason THOG Task (disjunctive reasoning)	4
Wason Selection Task (conditional reasoning)	5
Ravens' Matrices (intelligence)	4

Measuring Reasoning

Task	Median
Argument Evaluation Task	4
Belief Bias Syllogism Task	5
Cognitive Reflection Task	4
Conditional Inference Task	5
Evaluation of Arguments	3.5
Interpretation of Arguments	4
Recognition of Assumptions	4
Estimation	4.5
Insight Problem Solving	2
Statistical Reasoning	4
Wason THOG Task (disjunctive reasoning)	4
Wason Selection Task (conditional reasoning)	5
Ravens' Matrices (intelligence)	4

Conditional Inference Task

This problem concerns an imaginary letter-number pair. Your task is to decide whether or not the conclusion *necessarily* follows from the rule and the premise.

Rule: If the letter is not T then the number is 6.

Premise: The number is not 6.

Conclusion: The letter is T.

◯ YES (it follows) ◯ NO (no, it does not follow)

Denial of the Antecedent,
Affirmation of the Consequent

Modus Ponens,
Modus Tollens

Table 1 The four conditional types and four inference types used in the study

Conditional	MP		DA		AC		MT	
	Pr	Con	Pr	Con	Pr	Con	Pr	Con
if p then q	p	q	$\neg p$	$\neg q$	q	p	$\neg q$	$\neg p$
if p then $\neg q$	p	$\neg q$	$\neg p$	q	$\neg q$	p	q	$\neg p$
if $\neg p$ then q	$\neg p$	q	p	$\neg q$	q	$\neg p$	$\neg q$	p
if $\neg p$ then $\neg q$	$\neg p$	$\neg q$	p	q	$\neg q$	$\neg p$	q	p
Inference-type	Affirmative		Denial		Affirmative		Denial	
Validity	Valid		Invalid		Invalid		Valid	

If you are a good lecturer then you will get good student feedback.

Suppose I get good student feedback.

Does this mean I am a good lecturer?

Absolutely not, I might just be good at telling jokes, or setting easy examinations.

nal Inference Task

ary letter-number pair. Your task is to decide whether or not the
m the rule and the premise.

e number is 6.

it does not follow)

Denial of the Antecedent,
Affirmation of the Consequent

Modus Ponens,
Modus Tollens

Table 1 The four conditional types and four inference types used in the study

Conditional	MP		DA		AC		MT	
	Pr	Con	Pr	Con	Pr	Con	Pr	Con
if p then q	p	q	$\neg p$	$\neg q$	q	p	$\neg q$	$\neg p$
if p then $\neg q$	p	$\neg q$	$\neg p$	q	$\neg q$	p	q	$\neg p$
if $\neg p$ then q	$\neg p$	q	p	$\neg q$	q	$\neg p$	$\neg q$	p
if $\neg p$ then $\neg q$	$\neg p$	$\neg q$	p	q	$\neg q$	$\neg p$	q	p
Inference-type	Affirmative		Denial		Affirmative		Denial	
Validity	Valid		Invalid		Invalid		Valid	

Conditional Inference Task

Four "typical" ways of interpreting an "if *p* then *q*" statement:

1. Material conditional (*q* or not-*p*)
2. Defective conditional (irrelevant unless *p*)
3. Biconditional (*p* if and only if *q*)
4. Conjunctive conditional (*p* and *q*)

Normative model, as taught in logic courses

Unfortunate terminology (from Peter Wason): not a stupid way of thinking at all.

Material v Defective

- The difference between the material and defective conditionals is about the MT inference.
- ‘if *p* then *q*’ interpreted materially allows you to conclude not-*p* from not-*q*.
- ‘if *p* then *q*’ interpreted defectively does not allow this (as there is no *p*, the conditional is irrelevant, so the only premise you have is not-*q*).

  (Although: it is possible to draw MT if you have a defective conditional and sufficient Working Memory capacity to construct a mini contradiction proof: evidence suggests few people in this category).

Material v Defective

- The differenc rial and defective con e MT inference.
- ‘if *p* then *q*’ in allows you to conclude not-
- ‘if *p* then *q*’ in does not allow this (as there nal is irrelevant, so the only pr ot-*q*).

(Although: it is poss defective conditional and sufficient Worki ct a mini contradiction proof: evidence sug ory).

Defective Conditional:

“If good lecturer then good student feedback” only adds information if we know I’m a good lecturer.

In the case where I’m not, the conditional adds *no extra information*.

Material Conditional:

“Bad feedback” and “if good lecturer then good feedback” allows us to directly conclude “not good lecturer”

Conditional Inference Task

The conditional you adopt influences the validity of the four inferences:

Conditional	MP	DA	AC	MT
Material	Valid	Invalid	Invalid	Valid
Defective	Valid	Invalid	Invalid	Invalid*
Biconditional	Valid	Valid	Valid	Valid
Conjunctive	Valid	Invalid	Valid	Invalid

Conditional Inference Task

The conditional you adopt influences the validity of the four inferences:

Conditional	MP	DA	AC	MT
Material	Vali			lid
Defective	Val			lid*
Biconditional	Valid	Valid	Valid	Valid
Conjunctive	Valid	Invalid	Valid	Invalid

By looking at which inferences are endorsed, you can work out which interpretation the person adopts

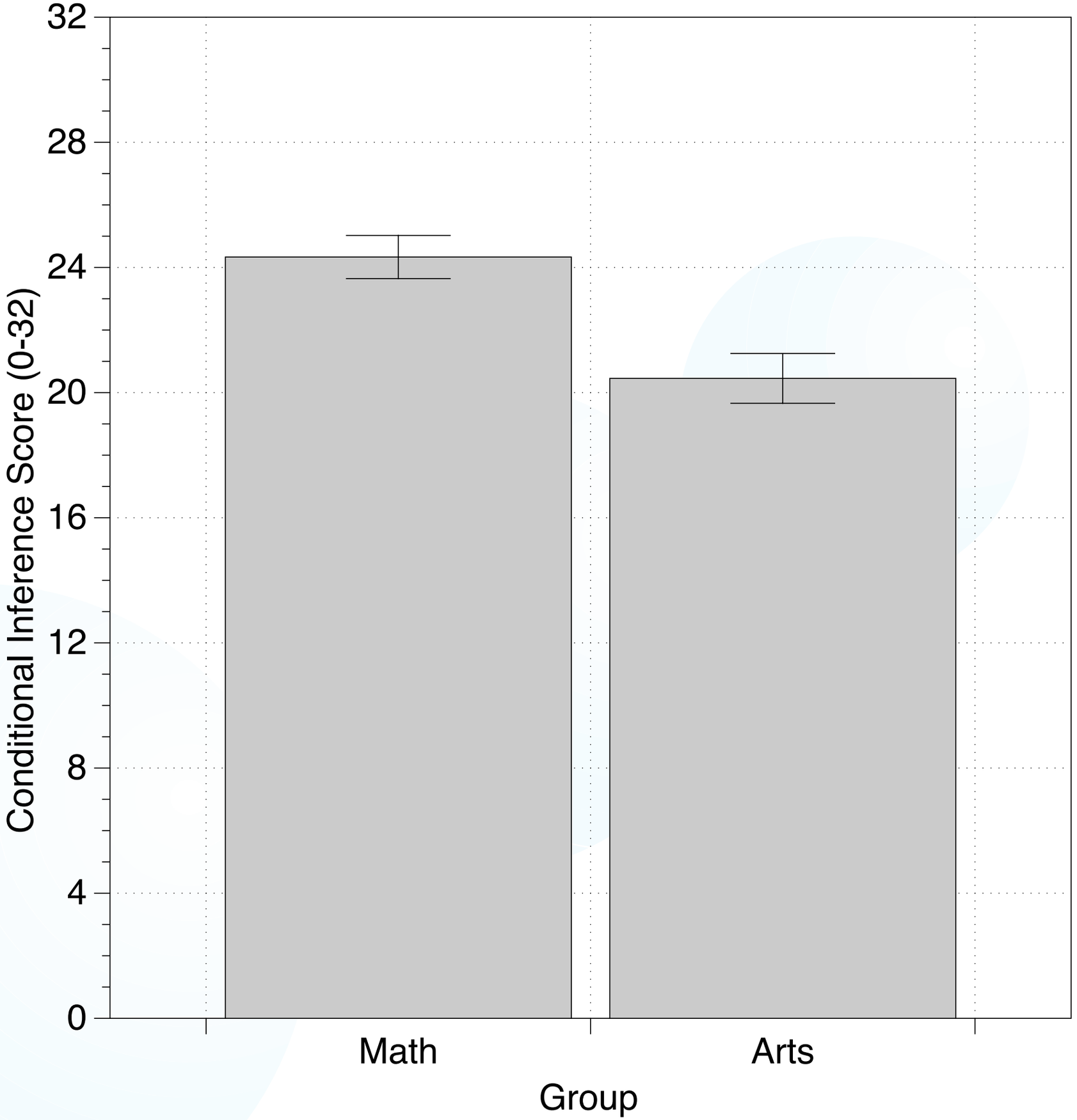
Research Strategy

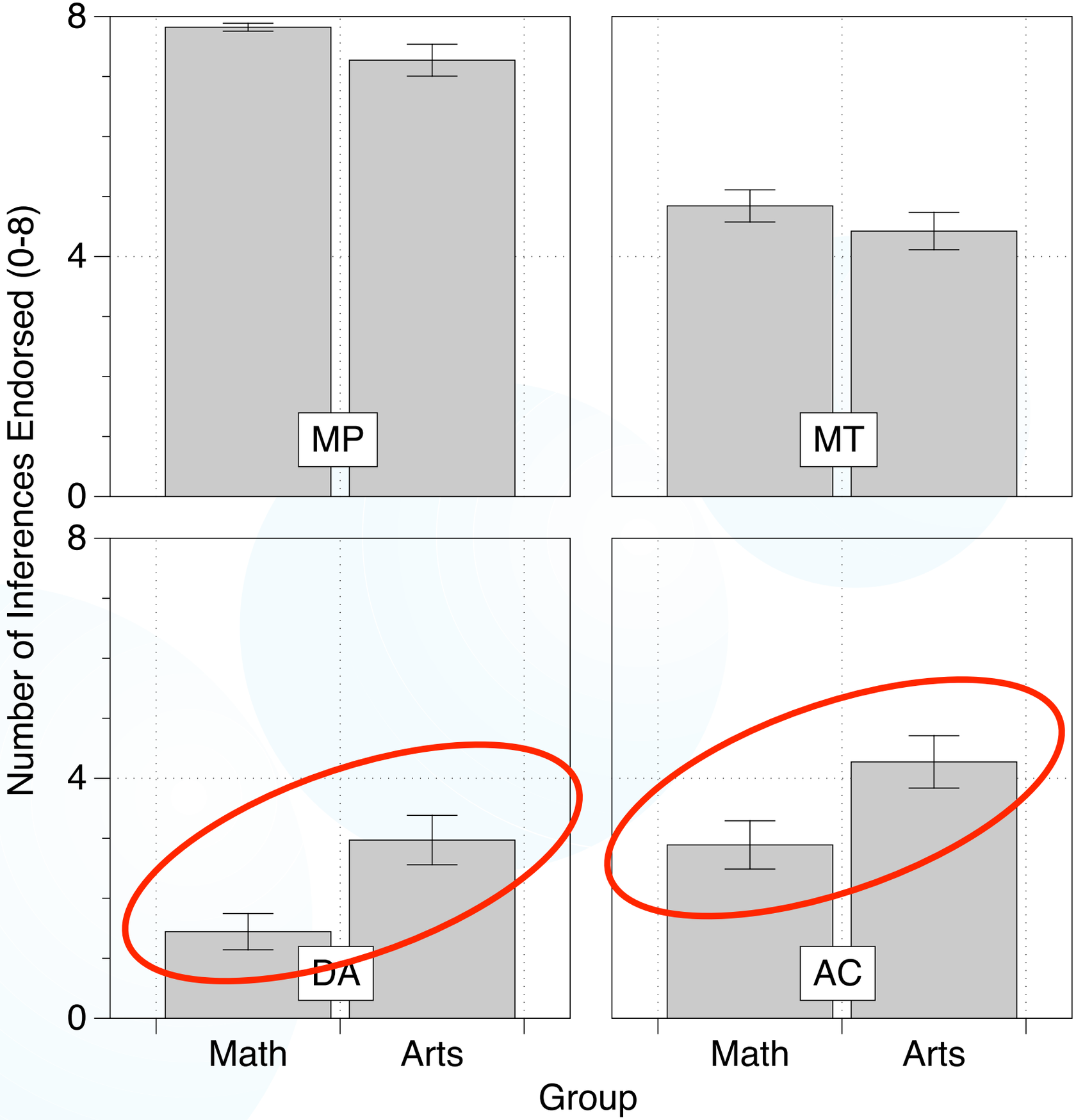
1. ~~How can we measure reasoning performance?~~
2. Do mathematicians “reason differently” to non-mathematicians?
3. Are such differences developmental?
4. Does the curriculum matter?

Study 1

Study 1

- Cross-sectional comparison of first year mathematics undergraduates ($N = 44$) and first year arts undergraduates ($N = 33$) at "highly rated" UK university (high IQ sample);
- Took place in Week 1 of u/g study (no lectures yet);
- Groups matched for IQ (AH5 test);
- Used Evans's Abstract Conditional Inference Task (Evans et al., 1996);
- Thirty two item test of abstract conditional inference.


Conditional Inference Score (0-32)
32
28
24
20
16
12
8
4
0
Math
Arts
Group


Number of Inferences Endorsed (0-8)
8
4
0
MP
MT
DA
AC
Math
Arts
Math
Arts
Group

Summary

- Maths students show an advantage on the conditional inference task prior to **any** undergraduate study;
- Not the result of differences in intelligence (groups were matched on AH5 scores);
- Advantage was uneven: came from advantage at rejecting DA and AC inferences, not from increased acceptance of MP or MT (move from biconditional to material/defective?).
- (Sort of) Consistent with predictions of Plato/ Vorderman. But is it developmental?

Research Strategy

1. ~~How can we measure reasoning performance?~~
2. ~~Do mathematicians "reason differently" to non-mathematicians?~~
3. Are such differences developmental?
4. Does the curriculum matter?

Study 2

Study 2

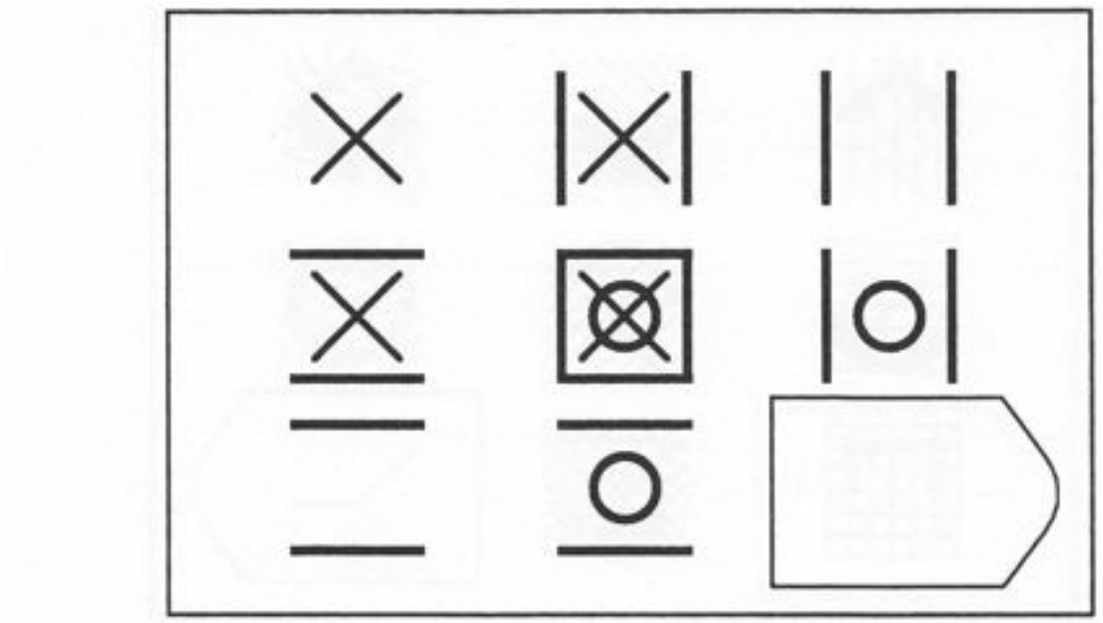
- Were the differences in Study 1 the result of **filtering** or **development**?
- Can't be filtering on intelligence (unless AH5 is a poor measure), so maybe on thinking disposition?
- Longitudinal quasi-experimental design, tracking students across AS level mathematics and AS level English literature.
- Two test points: start and end of year of study.

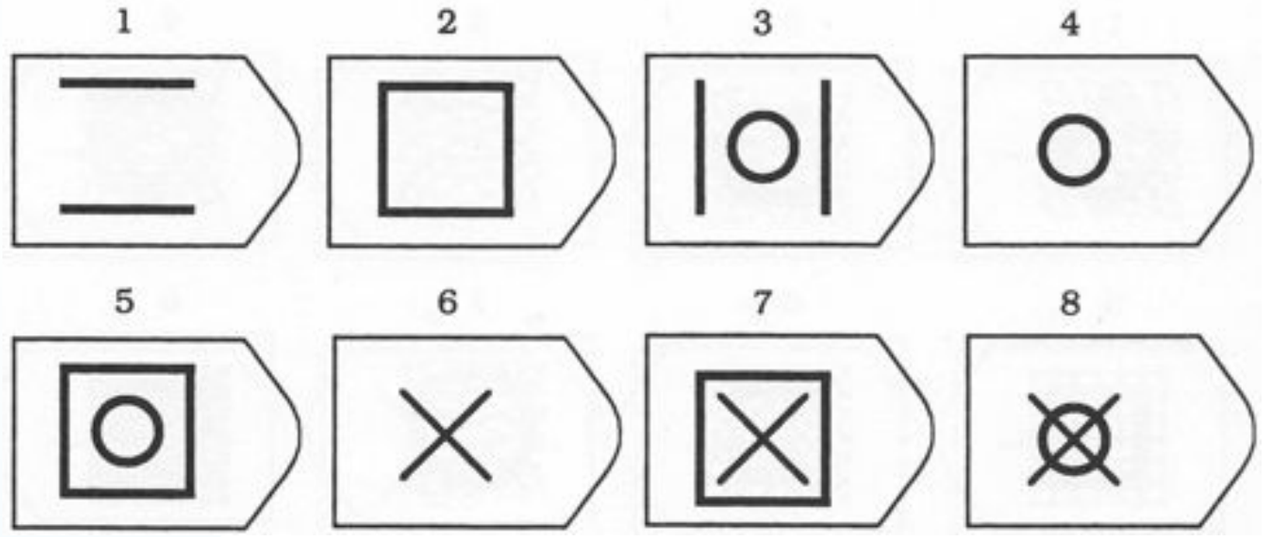
Study 2

Covariates:

- Raven’s Intelligence Test;
- Frederick’s Cognitive Reflection Test (measure of thinking disposition).

Raven's IQ Measure





Cognitive Reflection Test

(1) A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? _____ cents

Study 2

Manipulation Check:

- Maths Test

When expressing $\dfrac{x}{(x+1)^2(x^2+2)}$ in partial fractions, the appropriate form is

(a) $\dfrac{A}{x+1}+\dfrac{Bx+C}{x^2+2}$

(b) $\dfrac{A}{x+1}+\dfrac{B}{x^2+2}$

(c) $\dfrac{A}{(x+1)^2}+\dfrac{B}{x+1}+\dfrac{C}{x^2+2}$

(d) $\dfrac{A}{(x+1)^2}+\dfrac{B}{x+1}+\dfrac{Cx+D}{x^2+2}$

Study 2

Dependent Measure:

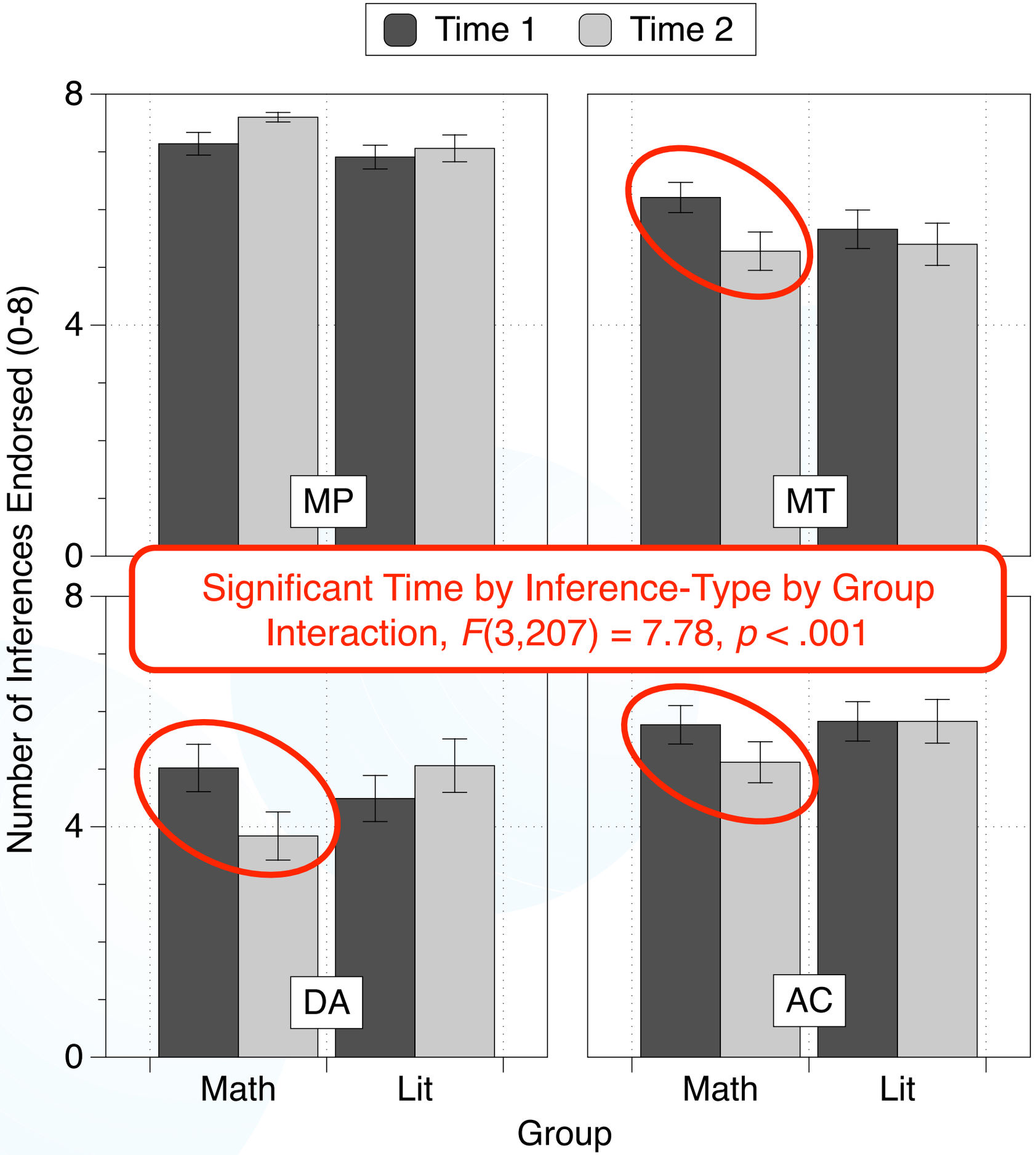
- Evans's Conditional Inference Task

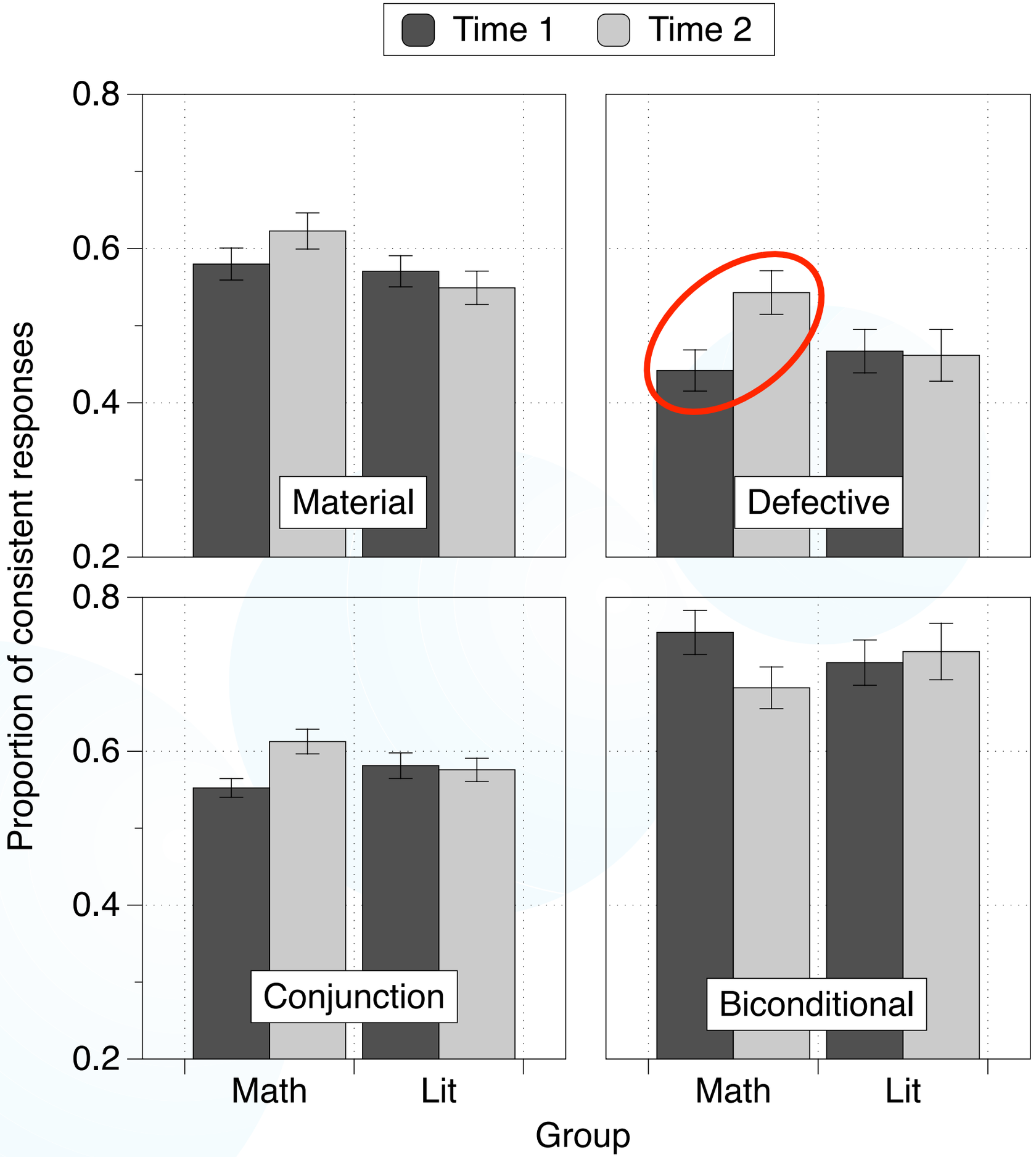
If the letter is U then the number is not 9.
The number is 9.
Conclusion: The letter is not U.

◯ YES

◯ NO

Study 2 Results


Time 1
Time 2
8
4
0
Number of Inferences Endorsed (0-8)
MP
MT
DA
AC
Math
Lit
Math
Lit
Group
Significant Time by Inference-Type by Group Interaction, F(3,207) = 7.78, p < .001



All 2 x 2 interactions significant, including with covariates

Causes?

If studying A Level mathematics is associated with a development towards the defective conditional interpretation, is this due to domain general changes (intelligence or thinking disposition), or domain specific experience (mathematical study)?

Ran a regression including change scores.

R^2	Predictors	Beta
.713**	Initial Defective Conditional Index	0.745**
	Initial RAPM (intelligence)	0.065
	Initial CRT (thinking disposition)	0.116
	Prior academic attainment	-0.006
	RAPM (intelligence) change	0.143
	CRT (thinking disposition) change	0.088
	Group (0 = lit, 1 = maths)	0.195*
	RAPM change x Group	0.023
	CRT change x Group	-0.091

Causes?

Apparently not due to general changes in intelligence or thinking disposition, but rather specific to mathematical study.

Obvious question: Were they simply taught how to solve such tasks during their A Level studies?

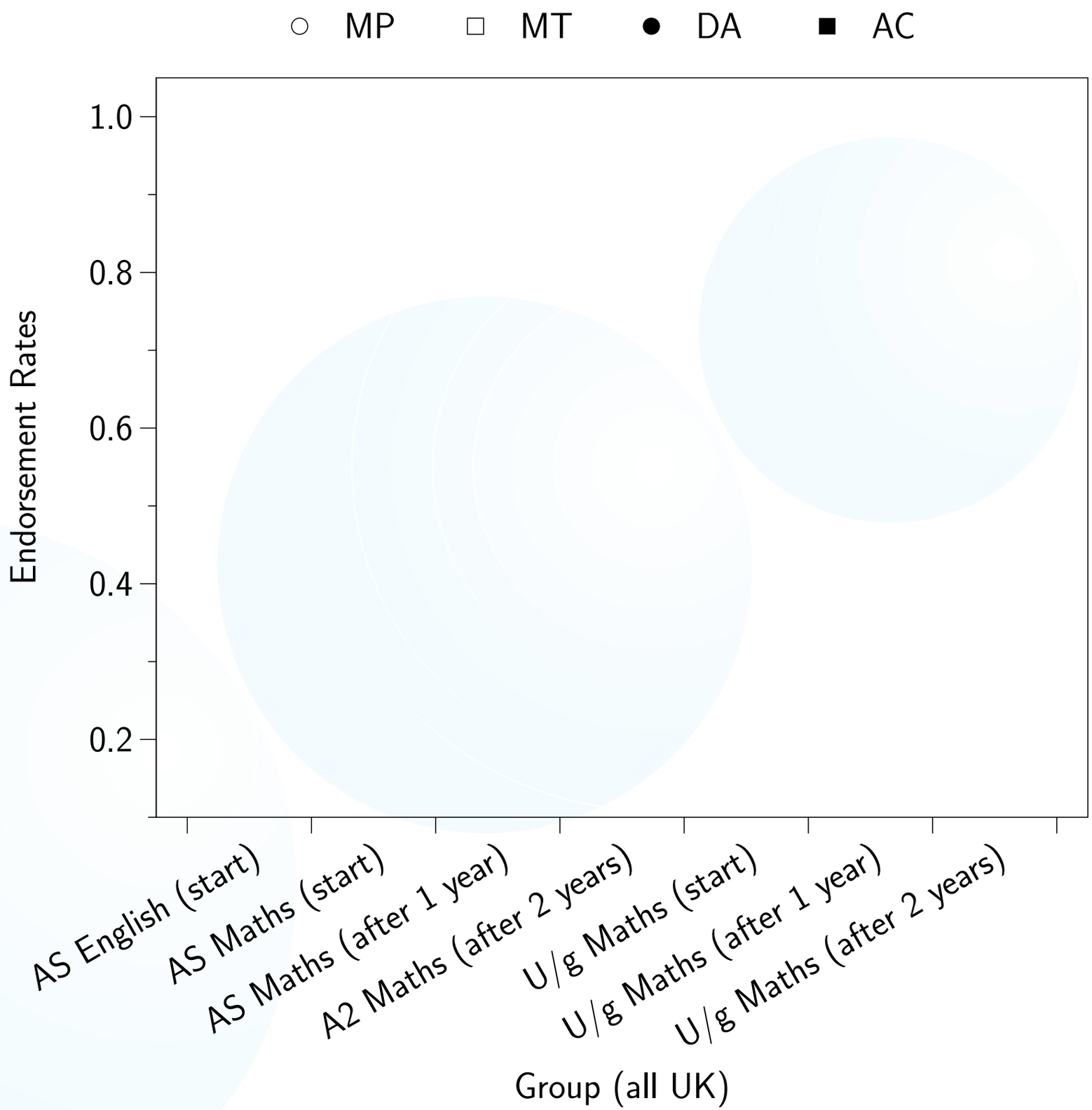
No. Two sources of evidence:

1. Not uniform “improvement” across all inference types.
2. Conditional inference is not on the syllabus, and is not examined: of 929 A Level mathematics examination questions set between 2009 and 2011, only one contained an explicit “if...then” sentence, and there were no mentions of “modus ponens”, “modus tollens” or “conditional”.

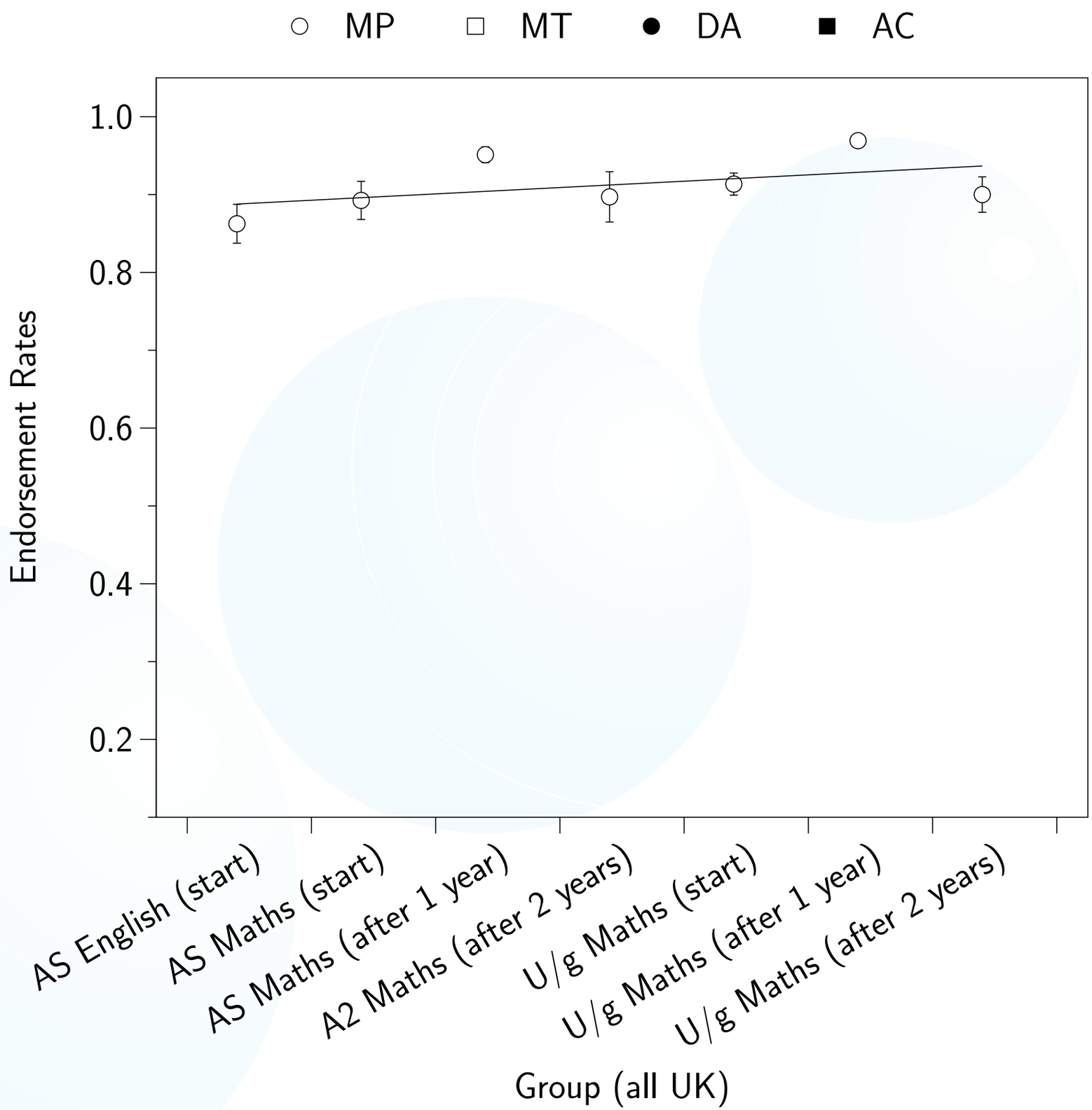
Summary

- There is an association between post-compulsory mathematical study and the development of conditional reasoning skills.
- But this appears to be towards a defective conditional interpretation rather than the normatively correct material conditional.
- **You can think about this as being increased scepticism of deductions: does studying mathematics make you better at spotting flaws in arguments?**
- Not caused by development in intelligence or thinking disposition, or by explicit curriculum content.

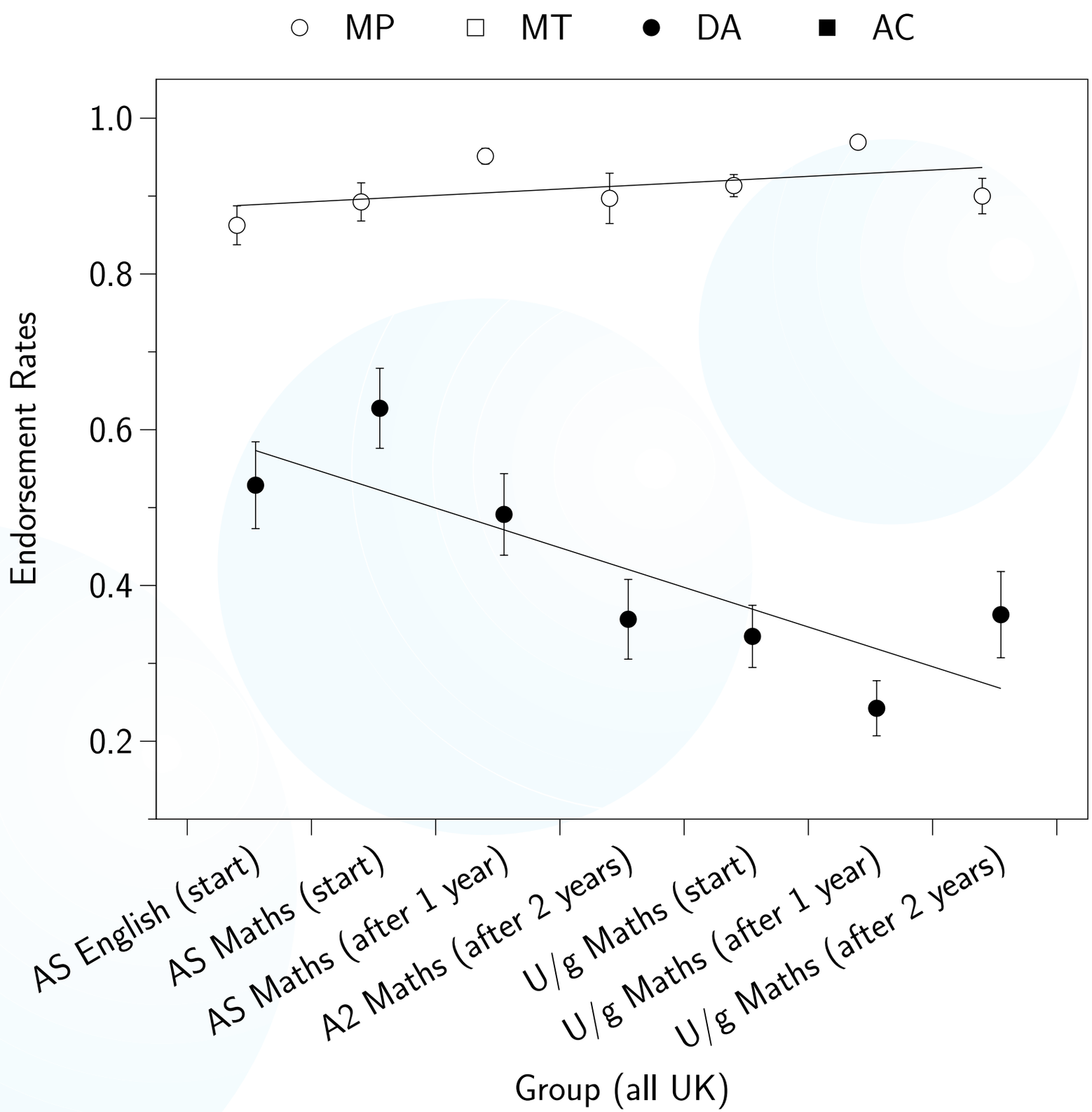
Summary of Lots of Similar Studies



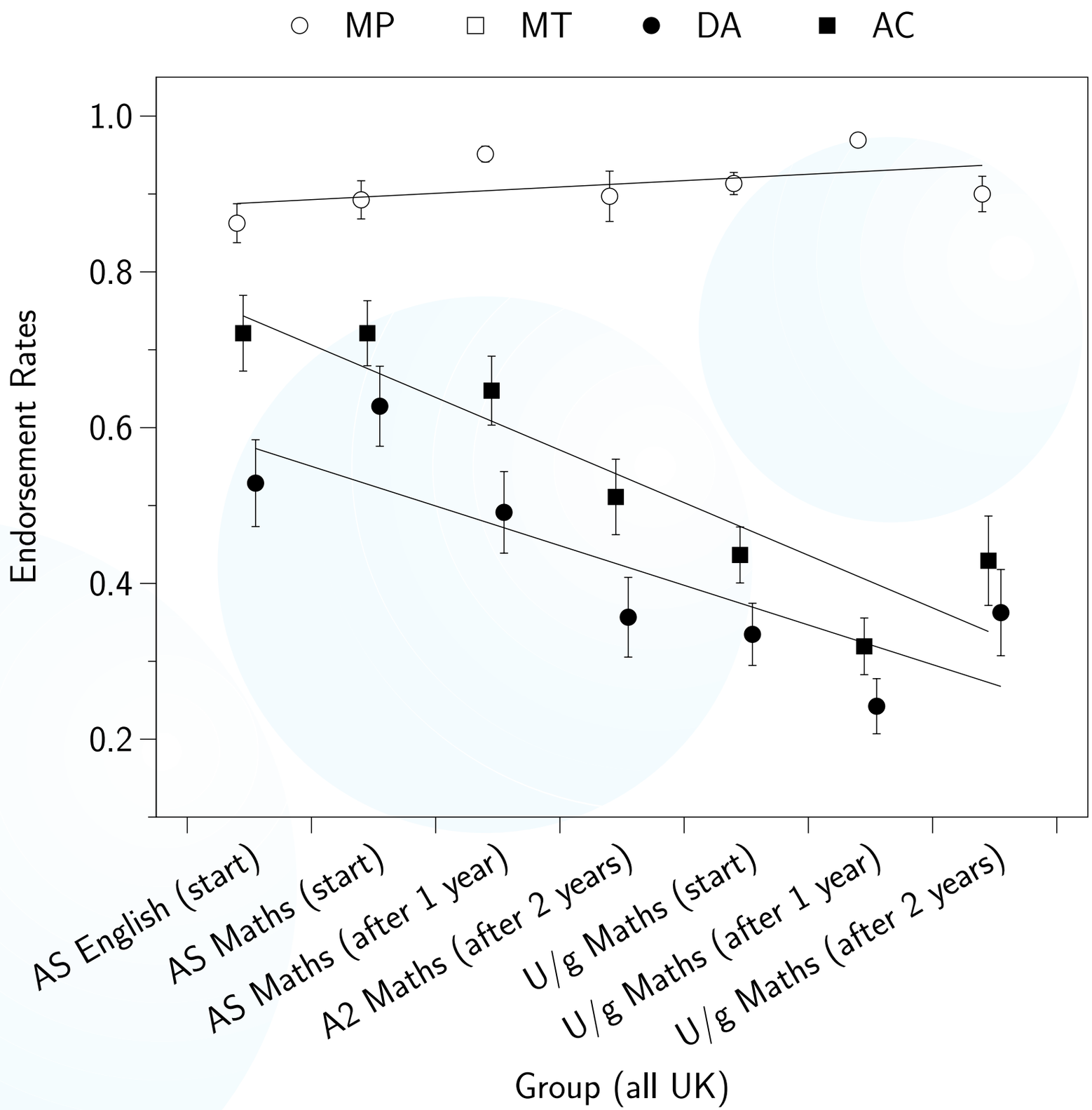
Summary of Lots of similar studies



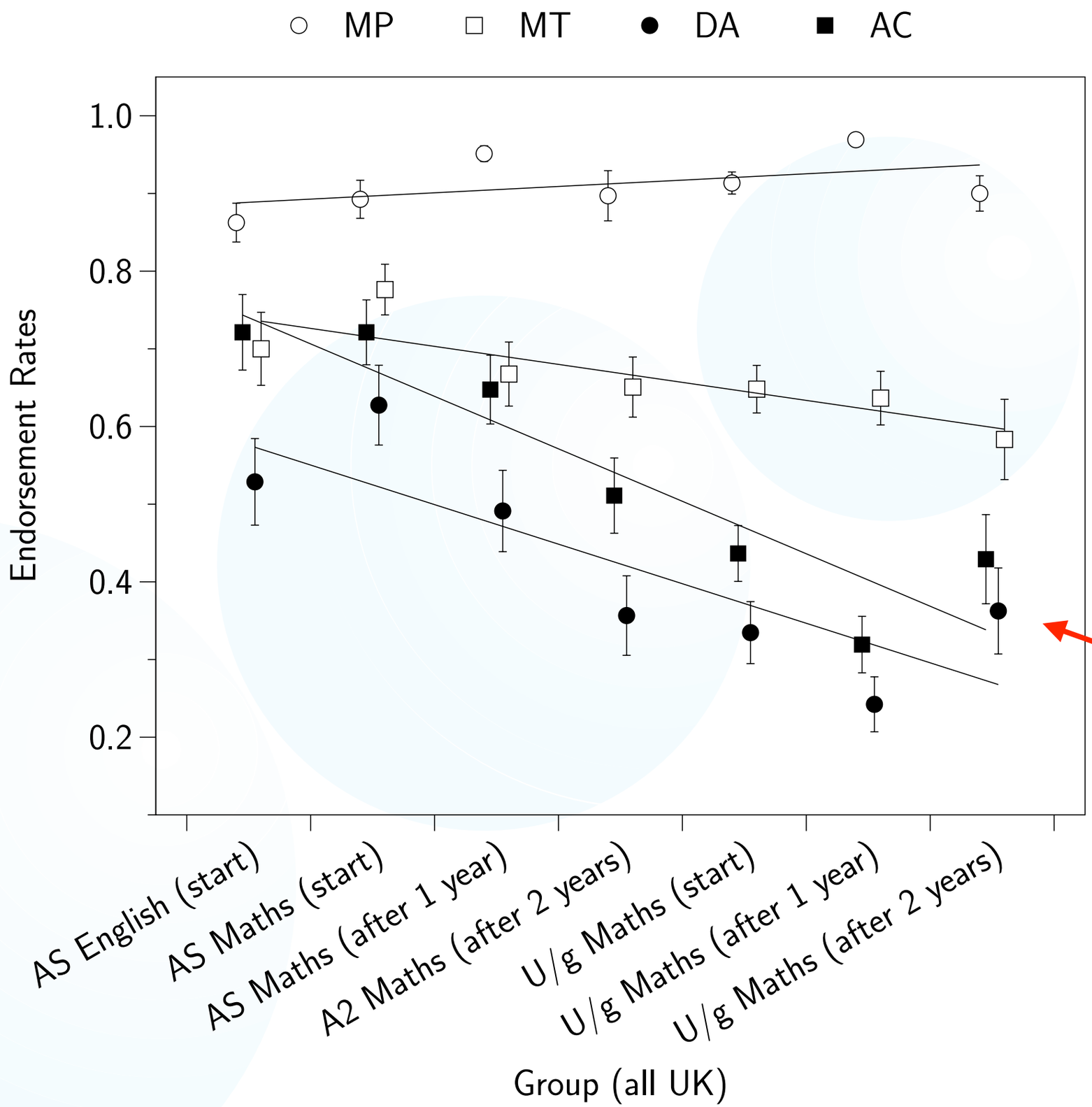
Summary of Lots of similar studies



Summary of Lots of similar studies



Summary of Lots of similar studies



spotting flaws

Research Strategy

1. ~~How can we measure reasoning performance?~~
2. ~~Do mathematicians "reason differently" to non-mathematicians?~~
3. ~~Are such differences developmental?~~
4. Does the curriculum matter?

Cypriot Comparison

- To investigate the curriculum question, I needed to look at the same issues in a different context.
- Repeated this study in Cyprus.
- Were able to run the study over two years.
- Cypriots can study “high intensity” or “low intensity” mathematics from 16-18.
- In this sense it is a more typical country than England (Hodgen et al., 2011).

Αν πιστεύετε ότι το συμπέρασμα συνεπάγεται αναγκαία παρακαλώ βάλτε νι (ν) στο κουτί που λέει ΝΑΙ, διαφορετικά βάλτε νι (ν) στο κουτί που λέει ΟΧΙ. Μην επιστρέψετε σε κάποιο πρόβλημα εάν το έχετε τελειώσει και προχωρήσατε ήδη στο επόμενο πρόβλημα. Απαντήστε σε όλες τις ερωτήσεις.

---

1. Αν το γράμμα είναι το Π τότε ο αριθμός δεν είναι το 2.
   Ο αριθμός είναι το 7.
   Συμπέρασμα: Το γράμμα είναι το Π.

   ◯ ΝΑΙ

   ◯ ΟΧΙ

23

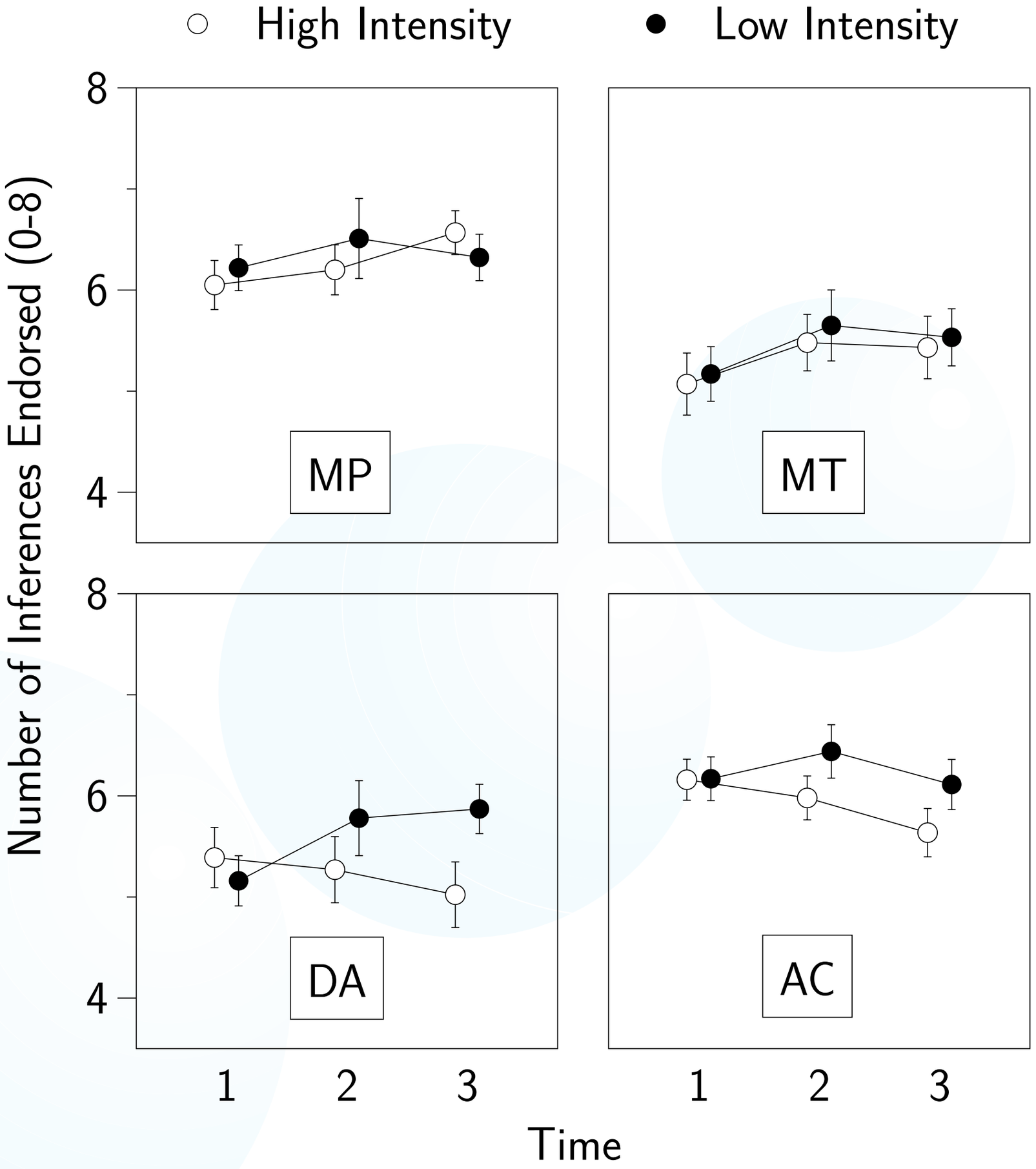
---

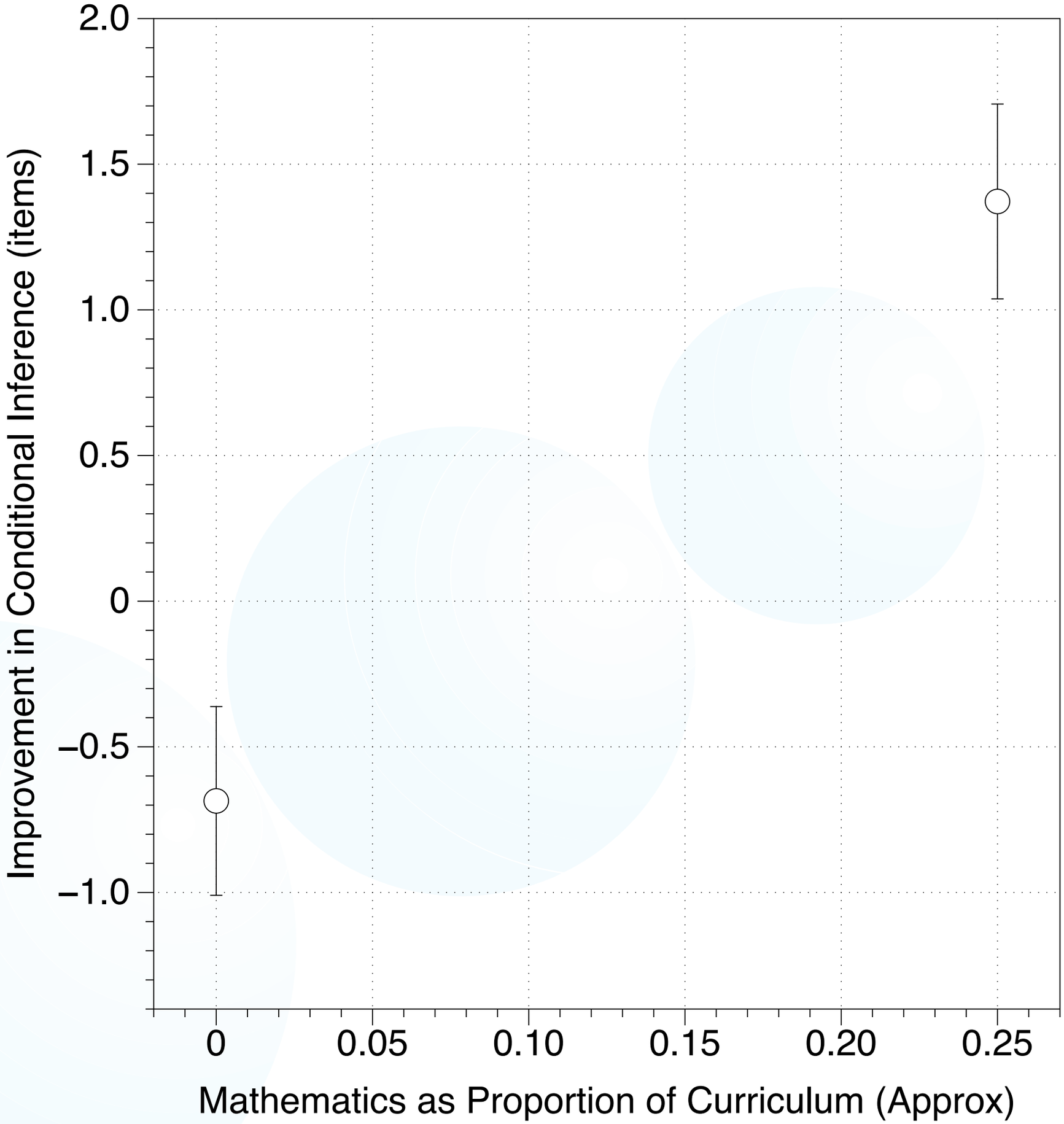
2. Αν το γράμμα δεν είναι το Α τότε ο αριθμός δεν είναι το 1.
   Το γράμμα είναι το Ν.
   Συμπέρασμα: Ο αριθμός δεν είναι το 1.

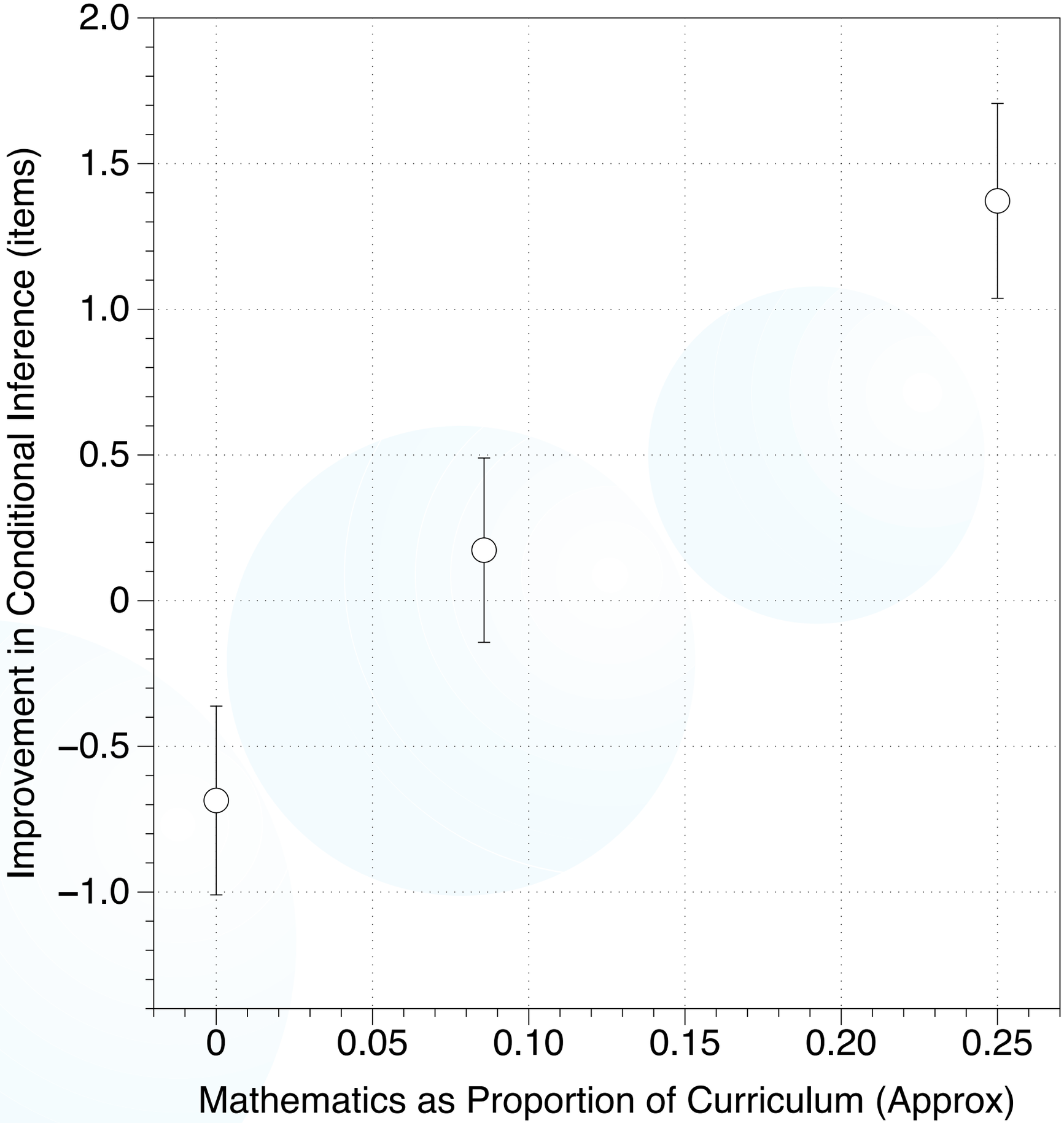
   ◯ ΝΑΙ

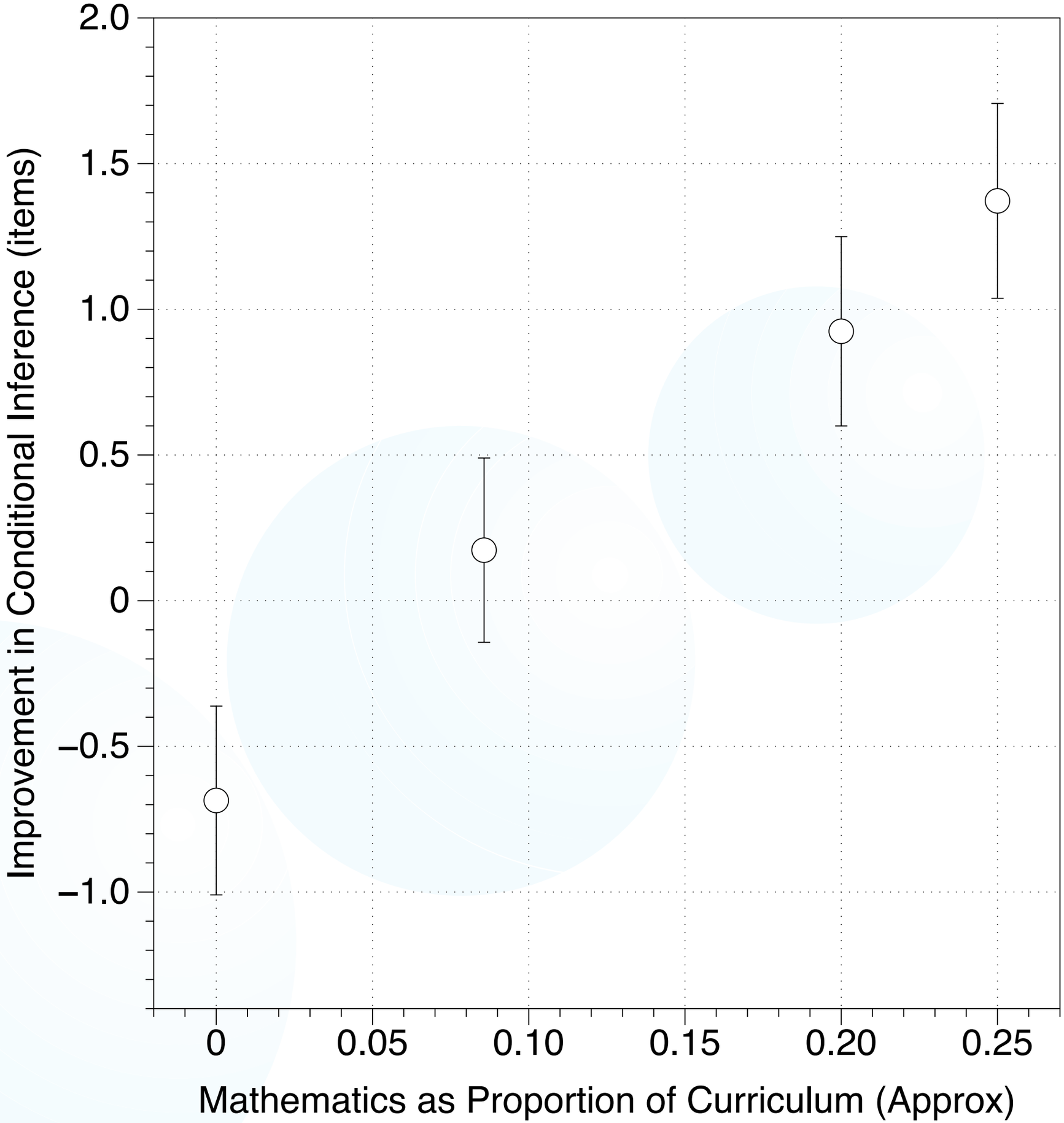
   ◯ ΟΧΙ

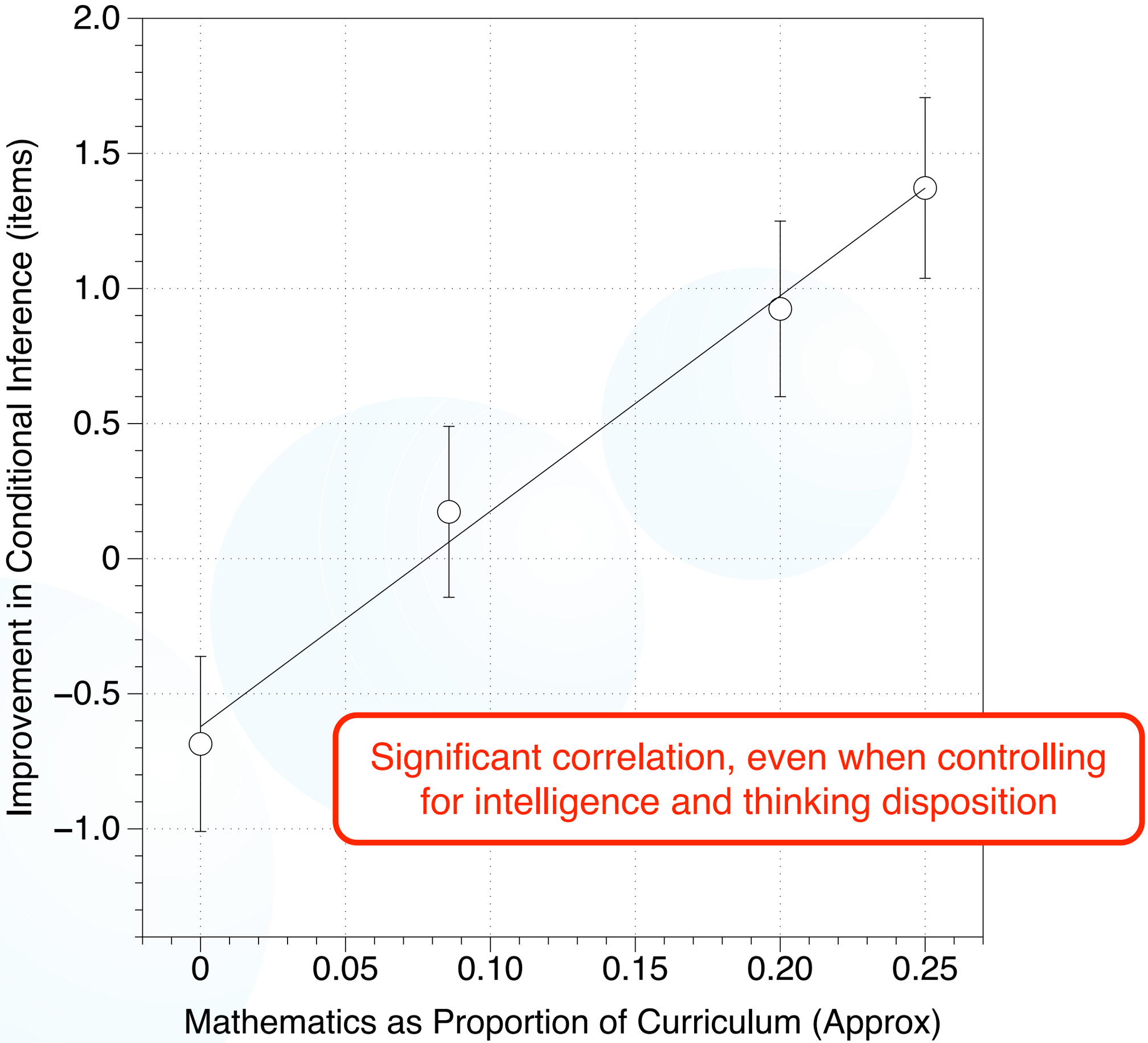
29


High Intensity
Low Intensity
Number of Inferences Endorsed (0-8)
8
6
4
MP
MT
DA
AC
1
2
3
Time


Improvement in Conditional Inference (items)
2.0
1.5
1.0
0.5
0
−0.5
−1.0
0
0.05
0.10
0.15
0.20
0.25
Mathematics as Proportion of Curriculum (Approx)


Improvement in Conditional Inference (items)
2.0
1.5
1.0
0.5
0
−0.5
−1.0
0
0.05
0.10
0.15
0.20
0.25
Mathematics as Proportion of Curriculum (Approx)


Improvement in Conditional Inference (items)
2.0
1.5
1.0
0.5
0
−0.5
−1.0
0
0.05
0.10
0.15
0.20
0.25
Mathematics as Proportion of Curriculum (Approx)

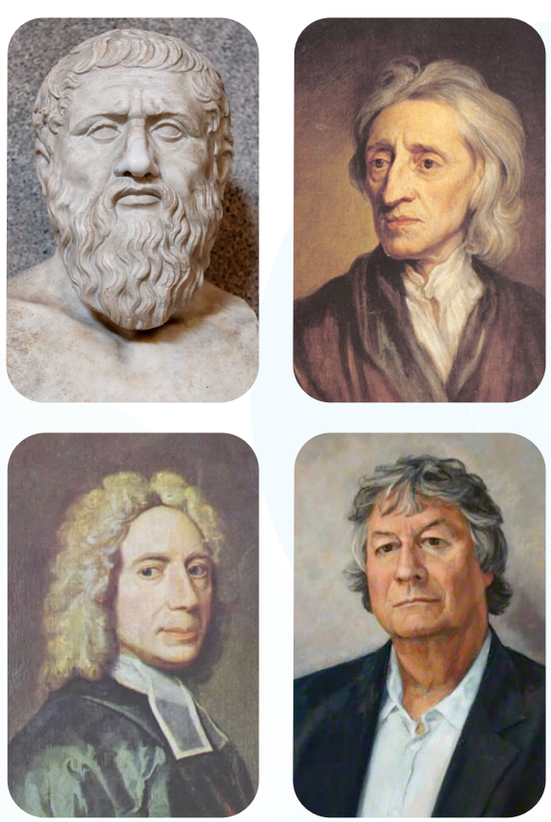

2.0
1.5
1.0
0.5
0
−0.5
−1.0
Improvement in Conditional Inference (items)
0
0.05
0.10
0.15
0.20
0.25
Mathematics as Proportion of Curriculum (Approx)
Significant correlation, even when controlling for intelligence and thinking disposition

Summary

- It seems that studying mathematics may be associated with the development of a defective conditional, at least for abstract “if *p* then *q*” statements, and the reduced influence of the biconditional.
- Good news for Plato/Vorderman: inconsistent with Thorndike, Piaget, Newell etc.

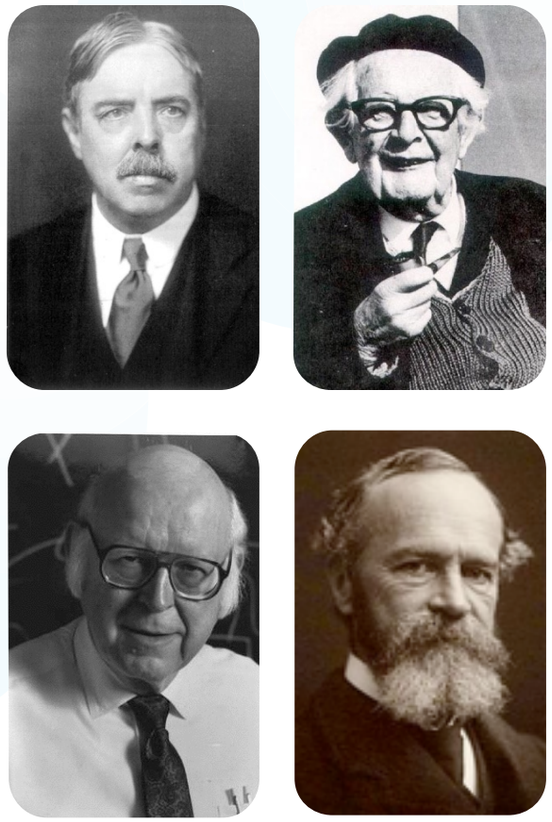
Summary

There is an fundamental (but under-debated) disagreement between people who claim that studying mathematics develops reasoning skills, and those who don't.

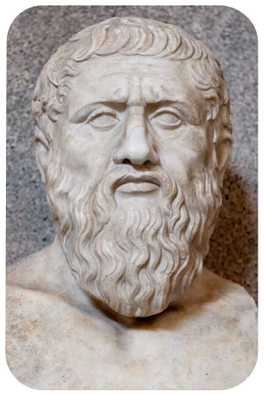


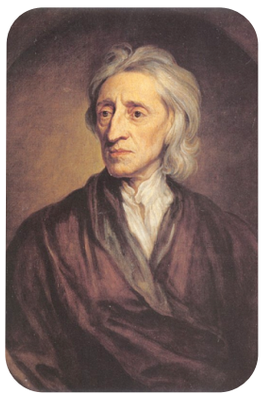
Plato, John Locke,
Isaac Watts, Adrian Smith

V



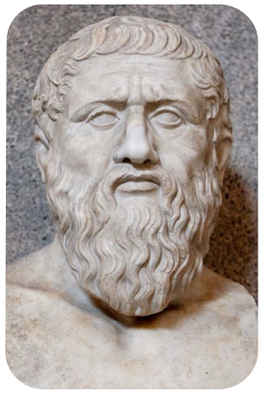
Edward Thorndike, Jean Piaget,
Alan Newell, William James

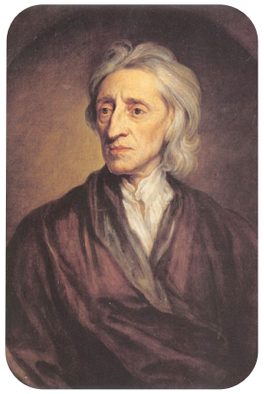


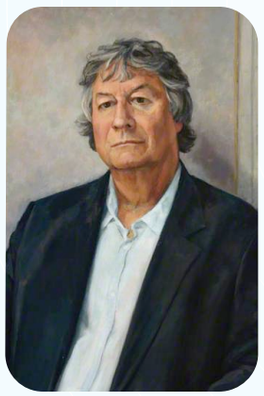


Summary

- These data are consistent with the suggestion that mathematics **is** associated with the development of conditional reasoning skills.
- Using modern psychology of reasoning measures allows for a more sensitive design than Thorndike's (1924) study.
- However: the development appears not to be towards the normative model of the conditional, but towards the defective conditional.
- Can conceptualise this as a tendency to be more sceptical of deductions than the general population.







Summary

Was Plato right?

I think so: but it’s a bit more nuanced than he thought.

Acknowledgements

Collaborators

Nina Attridge
(Loughborough)

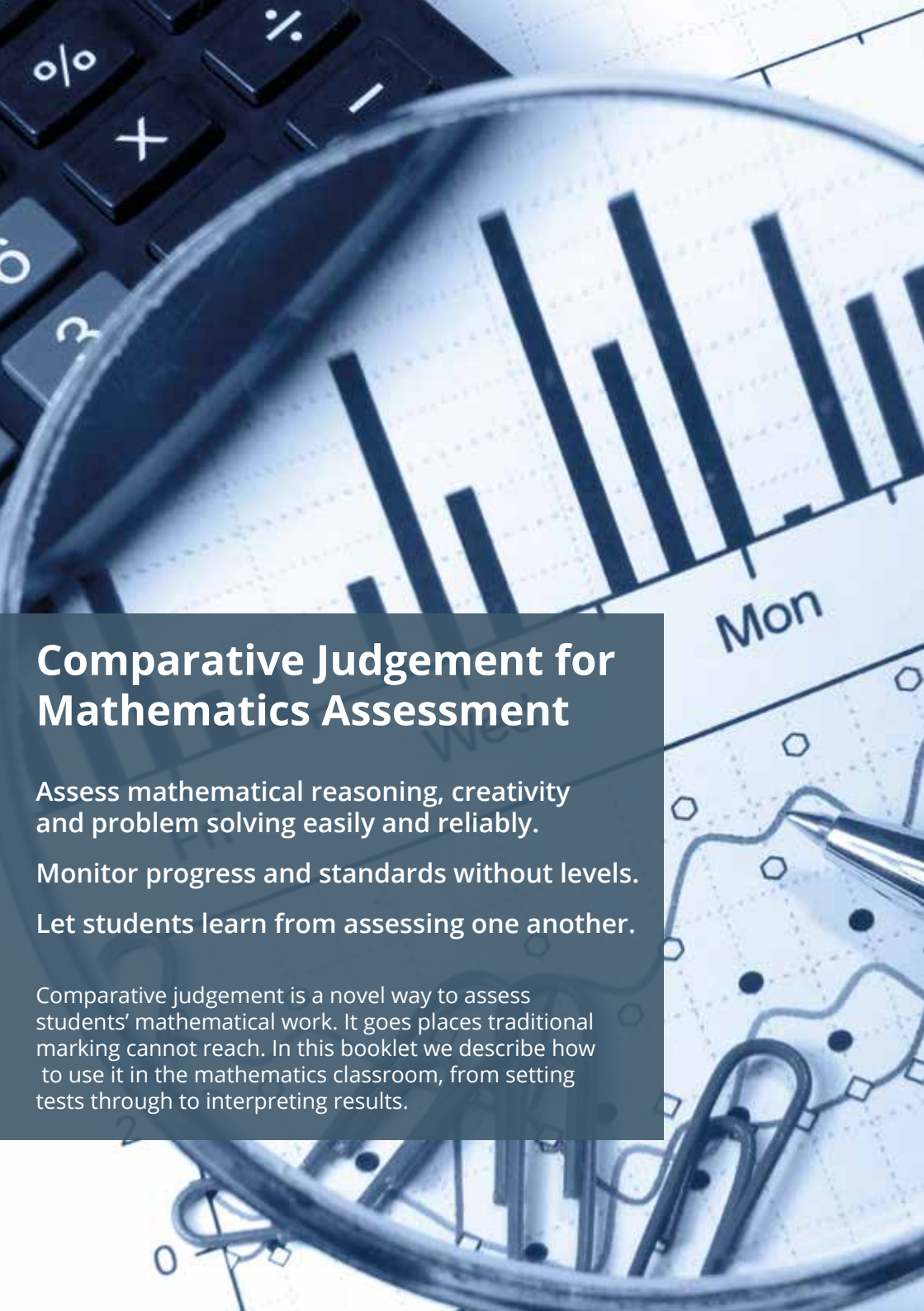
Adrian Simpson
(Durham)

Derrick Watson
(Warwick)

Funding

Worshipful Company
of Actuaries

THE ROYAL SOCIETY



Comparative Judgement for Mathematics Assessment

Assess mathematical reasoning, creativity and problem solving easily and reliably.

Monitor progress and standards without levels.

Let students learn from assessing one another.

Comparative judgement is a novel way to assess students' mathematical work. It goes places traditional marking cannot reach. In this booklet we describe how to use it in the mathematics classroom, from setting tests through to interpreting results.

Comparative Judgement

What is comparative judgement?

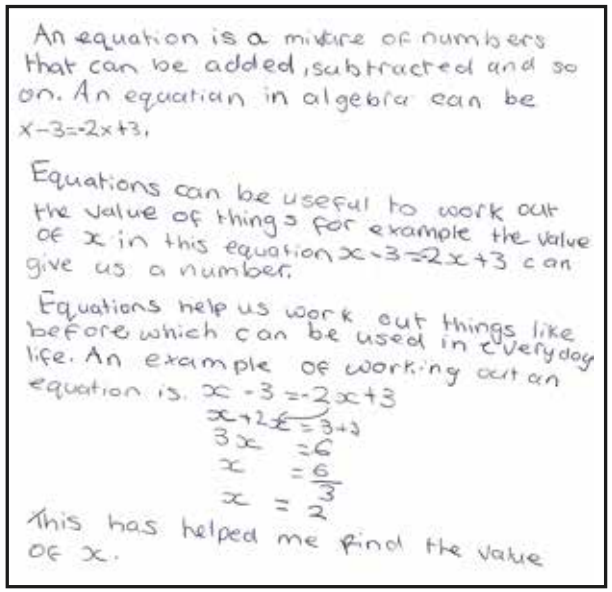
Comparative judgement is a way to assess open-ended and creative mathematical work. It involves no mark schemes and no marking. Instead two pieces of student work are presented on a screen and the assessor is asked to decide which is "better". The decision may be based on a specific objective, such as "the better understanding of fractions", or may be general, such as "the better mathematician". This is a binary decision. There is no need to decide how much better one piece of work is than the other.

When many such pairings are shown to many assessors the decision data can be statistically modelled to generate a score for each student. The statistical modelling also produces quality control measures, such as checking the consistency of the assessors. Research has shown the comparative judgement approach produces reliable and valid outcomes for assessing the open-ended mathematical work of primary, secondary and even undergraduate students.

Why comparative judgement?

There are three main reasons comparative judgement might sometimes be more appropriate then traditional marking.

Open-ended work.
Consider the test question "What is an equation? Give examples of how equations can be useful", followed by a blank page for students to provide their answer. How would you go about writing a mark scheme for such a question? If you did, how similar would it be to a mark scheme written by someone else? And how consistent would markers be when applying such a mark scheme to the students' answers? (Two student responses to this question are shown in Figures 1a and 1b) Research, and perhaps your intuition too, suggests there would be very little consistency in the mark schemes and the marking.Comparative judgement is instead based on direct judgements of what the students have written. Therefore it enables open-ended test questions to be assessed easily and consistently.


An equation is a mixture of numbers that can be added, subtracted and so on. An equation in algebra can be x-3=-2x+3.

Equations can be useful to work out the value of things for example the value of x in this equation x-3=-2x+3 can give us a number.

Equations help us work out things like before which can be used in everyday life. An example of working out an equation is. x-3=-2x+3
x+2x=3+3
3x=6
x=6/3
x=2

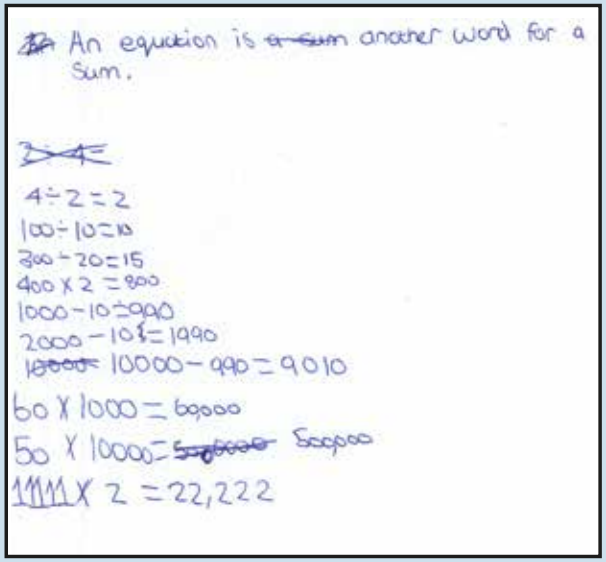
This has helped me find the value of x.

Figure 1a

Progress and standards.
Comparative judgement can be applied directly to different test questions. For example, if one student sat an algebra question and another sat a geometry question, then we can still compare their answers and decide who is "the better mathematician". This sounds a little strange at first, and it is indeed a more difficult decision than when comparing answers to the same test, but research has repeatedly shown it works. This feature of comparative judgement means it is ideal for monitoring progress and standards, particularly post levels. For example, we might include some tests from earlier in the year to see how students have progressed. Alternatively, we might include tests from another class or school to gauge our students' achievement compared to others.

Peer assessment.
Open-ended test questions are valued because they assess mathematical reasoning, problem solving and creativity. Research suggests peer discussion and critiquing the work of others helps students to develop these skills. One way to do this is to let students assess their peers' work using comparative judgement. Students can reflect on and discuss what features make one answer better than another when making a decision. We have found that secondary students find comparatively judging one another's mathematical work engaging and valuable.

Follow-up lessons can be used to discuss how students made their judgements, reflect on the qualities of the "best" answers (which should be anonymised, see Page 4), and consider how similar questions can be better answered in the future.


An equation is ~~a sum~~ another word for a sum.

~~3=4~~

4÷2=2
100÷10=10
300-20=15
400x2=800
1000-10=990
2000-10=1990
~~10000~~ 10000-990=9010
60x1000=60000
50x10000=~~500000~~ 500000
1111x2=22,222

Figure 1b

Setting comparative judgement tests

You can use comparative judgement for free in the classroom by visiting nomoremarking.com. How to use the website is described in the rest of this booklet. The first step is designing and setting a test for your students.

Comparative judgement can be used to assess many different kinds of tests. However, it is best suited to open-ended questions that encourage a wide range of unpredictable answers from students. We have found that a short test question followed by a single blank page for the answer is ideal. Examples of questions we have used are seen in Figure 2.

> Why do we need negative numbers? Give examples of how negative numbers can be useful.
>
> Write down these fractions in order of size from smallest to largest. Underneath, describe and explain your method for doing this.
>
> 3/4 3/8 2/5 8/10 1/4 1/25 1/8
>
> Give examples of powers and roots in mathematics. Explain the relationship between powers and roots.
>
> Explain what a derivative is to someone who hasn't encountered it before. Use diagrams, examples and writing to include everything you know about derivatives.

Figure 2

These are only examples. We expect teachers will explore various styles of test question. Other kinds of work can also be assessed, such as projects.

You might even want to try more innovative forms of mathematics assessment. For example, the nomoremarking website also supports audio and video files. We are very keen to hear from teachers who create their own ways of using nomoremarking!

Anonymisation

We recommend that tests are anonymised when they are uploaded to the nomoremarking website. Anonymisation reduces assessment bias that can arise from knowledge of the students. This is particularly important if using comparative jdugement for peer assessment, or in collaboration with other schools for standards comparison. For this reason we recommend students write their names on the back of the test, not the front. When the tests are scanned (see next page) only the work will be uploaded and no identifying information.

Assessing the tests

Scanning

The tests need to be scanned so that they can be uploaded to the nomoremarking website. The website supports pdf files as well as image files such as jpegs. Each test needs to be a separate file – you cannot upload the students' tests as one big file. The example tests shown opposite were all designed to be a single page because this makes judging them online easier. However, the website supports multipage tests and this works fine too.

For many teachers, administrative support might be available to do the scanning. Other teachers may have to undertake it themselves. We recommend a multifeed scanner that can process dozens of tests in a few seconds. A flatbed scanner that only allows one sheet at a time would take an age.

It is important to use sensible filenames for the scanned tests. The filenames might be the students' names or ID numbers. The assessors who do the judging will not be able to see the filenames, so it will all be anonymised from their point of view.

There are two ways to ensure sensible filenames. The first is to change the filenames after the scanning is done, going through the students' tests one by one. The second is to record the filenames automatically generated by the scanner on a class list. Most modern scanners generate sequential filenames starting at 1. We advise putting the tests in alphabetical order by student name to reduce the burden of matching students to filenames.

Uploading

It's now time to use the website! Here we give a brief outline of the basics to get you going. A full user-guide can be downloaded from the website, and we respond quickly to requests for help and support.

First visit nomoremarking.com and sign up, which is free and takes just a few seconds. (Click **Sign In** then **Register**.) Once signed in, click **Admin Menu** then **Add Task.** Then fill in the form. You can ignore most of the form for now, just the following are needed.

Name: Give your task a sensible name such as "Fractions test December 2015"

Question for Judges: Describe briefly how you want assessors to make their decisions, such as "The better understanding of fractions?"

Response Type: Select your filetype, i.e. "pdf" or "image".

Now click the **Add Candidates** tab. Click the **Choose Files** button, and find and select your scanned tests. Sit back and watch as the files upload. This might take some time, scanned tests can be quite large. It's a good moment to go and make a cup of tea.

Judges

Now it's time to add the people who will do the judging. This might be teachers or, for peer assessment, students.

How many judges are needed, and how many judgements do they need to do each? Good question, glad you asked! We use a simple rule of thumb: 5 judgements per test is enough to produce a reliable score for every student. For example, if you have uploaded 25 tests then you will need no more than 125 judgements. Usually these will be divided equally among the judges. In this case, if you have 5 judges, you would allocate them 25 judgements each. You can always allocate more: there is no such thing as too many judgements!

For peer assessment, we have found that about 20 or 30 judgements per student is appropriate for an hour's lesson. This will be more than enough to produce reliable scores.

To allocate the number of judgements per judge, click on the **General tab.** Enter your number in **Judgements per judge**.

Now click the **Add Judges** tab. To assign someone as a judge, type in their email address. You can add several judges at once by separating their email addresses with a comma. Always add yourself as a judge so you can test things as you go. Click **Update Judges.**

Do the judging

We're almost ready to go. The tests have been uploaded, the number of judgements calculated, and the judges assigned.

Click the **Do Judging** tab. You will see a list of the assigned judges. Each has a unique judging url which you can see in the **Do Judgements** column. Find yourself and click your url. You can now try judging some tests! Click **Start** and two tests will appear on the screen. Use the **Left** and **Right** buttons to decide which is "better".

Once you're happy it's working you can invite the other judges. Click the Home button in the top-right corner, and then click the **Do Judging tab**. Now tick the checkbox in the **Send** column, and click **Send**. The judges will receive an automated email containing their unique url. (Tell them to check their spam if they don't receive it.)

Results!

Checking it worked

Once the judging is complete you will need to check everything is ok. Click on the **Judge Feedback** tab. Look at the **Judgements** column to check all the judges have completed their judgements. You can also see the **Median Time** it took each judge to make a decision, as well as other data about their performance.

When all the judges are finished click on the **General** tab. Click the **Re-Calculate** button next to **CJ Estimation**. The statistical modelling is now being performed. You may have to wait a few moments if there are a lot of tests.

In the grey bar at the bottom check the **Reliability** figure. This is a measure of how consistent the judges were. The reliability should be greater than 0.70. If it is lower than this get in touch with us!

Getting the results

Now you can download the students' scores. Click the **Downloads** dropdown menu and choose **Candidates**. This will download a csv file which can be opened in a spreadsheet application such as Excel. In the spreadsheet, the **ID** column is the list of students (identifiable by the filenames you used), and the **TrueScore** column is the scores.

Using the results

The statistical modelling produces scores with a mean of about 0 and a standard deviation of about 2.5. You might want to convert the scores to something more meaningful to teachers, students and parents. A mean of 50 and a standard deviation of 15 is typical.

To convert the scores, type **Final Score** at the top of column O in the spreadsheet. In cell O2 carefully type the following formula. (The formula assumes **25** tests have been judged. Change the two occurrences of **25** to the number of tests that were actually judged.)

=50+15*(C2-AVERAGE(C$2:C$**25**))/STDEV(C$2:C$**25**)

Now drag the formula down to fill column O. This is your final set of scores with a mean of 50 and standard deviation of 15.

You can use the scores as you would scores from traditional marking: to apply grades, to compare students, to feedback to students, and so on.

Going further

In this guide we have focussed on the practicalities of using comparative judgement in the classroom. We have kept the technicalities of using the nomoremarking website to a minimum. That's all you need to have a go! For those wishing to delve further there is more guidance available on the website. And do get in touch with any questions. No request is too small, silly or ambitious!

As we have hinted, there are many more things you can do with comparative judgement than described here. It is a simple but infinitely flexible approach to assessment that can be used in different and imaginative ways.

If you have any questions or need help please contact us at
maths@nomoremarking.com

You can find articles about the research behind using comparative judgement for mathematics assessment in the Publications section here:
homepages.lboro.ac.uk/~maij

Ian Jones and Matthew Inglis are academics at the Mathematics Education Centre, Loughborough University. Chris Wheadon and Brian Henderson are directors of NoMoreMarking Ltd. **The research was funded by The Royal Society, The Nuffield Foundation, AQA, MEI, Worshipful Company of Actuaries Charitable Trust and HE STEM Programme.**