Logical Reasoning in Mathematics Students:
Conditional Inference with Mathematical, Abstract and Everyday Content

Lara Alcock[1], Ben Davies[2] and Matthew Inglis[1]

[1]Loughborough University

[2]University of Southampton

Corresponding Author: Lara Alcock, Department of Mathematics Education, Loughborough University, Loughborough, LE11 3TU, UK, l.j.alcock@lboro.ac.uk

Declarations:

This research was conceived by Lara Alcock and designed by Lara Alcock and Matthew Inglis. Material preparation and data collection were performed by Lara Alcock. Data analysis was performed by Ben Davies and Matthew Inglis. The first draft of the manuscript was written by Lara Alcock; Ben Davies and Matthew Inglis contributed to theoretical formulation and commented on previous versions. All authors read and approved the final manuscript. Ethical approval for the studies reported here was given by the Loughborough University Ethics Review Subcommittee (references 15945, 16091, 16357, 16723, 17575, 17501). Data, analyses and materials associated with this manuscript are available in the online supplementary materials at https://figshare.com/s/2ed67504674c50512bb3.

Short Title: Logical Reasoning in Mathematics Students

Keywords: believability; comparative judgement; conditional inference; logic; reasoning; undergraduate mathematics

Logical Reasoning in Mathematics Students:
Conditional Inference with Mathematical, Abstract and Everyday Content

Abstract

A core aim of mathematics education is to develop mathematical reasoning. This often involves making and evaluating inferences from *conditionals* of the form 'if $A(x)$, then $B(x)$'; some inferences are valid and others are not. We report five studies (total $N = 546$), building on contemporary theories of reasoning with everyday conditionals to show that 1) the perceived believability of mathematical conditionals affects inference acceptance, 2) mathematics undergraduates reason similarly from mathematical, abstract and everyday conditionals but more normatively from mathematical conditionals with high perceived believability, and 3) perceived believability, rather than perceived easiness, accounts for this effect. We relate this empirical work to a mathematics-specific extension of the standard theoretical account of conditional inference, considering implications for mathematics educators.

Introduction

An important goal of mathematics education is to develop mathematical reasoning. This is reflected in research across primary, secondary and tertiary education on problem solving (Lampert, 2001; Mason et al., 1982; Schoenfeld, 1985), student and teacher justifications and proofs (Bell, 1976; Harel & Sowder, 1998; Stylianides, 2007), and classroom norms that support reasoning development (Brousseau, 1997; Yackel & Cobb, 1996; Yackel et al., 2000). It is also reflected in policy: one practice standard in the Common Core State Standards for Mathematics is 'construct viable arguments and critique the reasoning of others' (NGA & CCSSO, 2010), and England's National Curriculum (DfE, 2021) states that pupils should 'reason mathematically by following a line of enquiry, conjecturing relationships and generalisations, and developing an argument, justification or proof using mathematical language'. Indeed, developing reasoning is an important reason for teaching mathematics. Both historically (Locke, 1706/1971) and more recently (Smith, 2004), high-profile thinkers and policymakers have asserted that mathematical study develops not only subject-specific knowledge but also domain-general reasoning skills.

To successfully construct and critique mathematical arguments, students must learn to judge what can legitimately be inferred from accepted claims. It is therefore important to understand what affects students' evaluations of inference validity. In this paper, we consider *inference content* and *inference form*.

Inference content should affect evaluations of validity due to individual differences in content knowledge (Braithwaite & Rafferty, 2025). People with better-developed *personal example spaces* (Sinclair et al., 2011) are better placed to generate valid inferences by generalising across examples, and to challenge invalid inferences by identifying counterexamples. Using counterexamples requires learning mathematics-specific norms. Where everyday argumentation often tolerates exceptions, one counterexample is enough to render a mathematical inference invalid (Zazkis & Chernoff, 2008).

Inference form should affect evaluations of validity because some inference forms are valid and others are not. However, this is not obvious in everyday communication because an everyday conditional 'if A then B' is often interpreted as the biconditional 'A if and only if B'. Many promises have this quality: 'if you mow the lawn, then I will give you $5' is routinely understood as implicating its inverse 'if you do not mow the lawn, then I will not give you $5' (Cummins et al., 1991). In contrast, mathematics treats a conditional 'if $A(x)$, then $B(x)$' as logically distinct from its inverse and converse, but equivalent to its contrapositive. In Table 1, 'If x is less than 2, then x is less than 5' is true, but its inverse and

converse are false. Its contrapositive is true, as illustrated in Figure 1 – when 'if $A(x)$, then $B(x)$' is true, an x outside the truth set of $B(x)$ must be outside the truth set of $A(x)$.

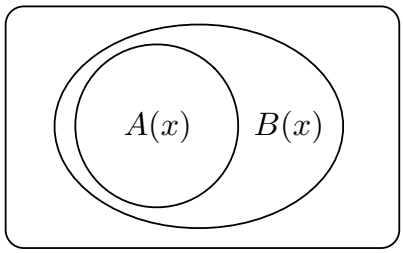
Table 1

Conditionals with their Inverses, Converses and Contrapositives.

	Mathematical example	General case
Conditional	If x is less than 2, then x is less than 5.	If $A(x)$, then $B(x)$.
Inverse	If x is not less than 2, then x is not less than 5.	If not $A(x)$, then not $B(x)$.
Converse	If x is less than 5, then x is less than 2.	If $\mathrm{B}(x)$, then $\mathrm{A}(x)$.
Contrapositive	If x is not less than 5, then x is not less than 2.	If not $\mathrm{B}(x)$, then not $\mathrm{A}(x)$.

Figure 1

Relationship Between Truth Sets for a True Conditional 'if $A(x)$ then $B(x)$'.



Students often miss these logical distinctions. Hoyles and Küchemann (2002), for instance, asked high-achieving 13 year-olds whether the conditional and converse below say 'the same thing':

If the SUM of two whole numbers is EVEN, their PRODUCT is ODD.

If the PRODUCT of two whole numbers is ODD, their SUM is EVEN.

Of 2,663 students, 71% said yes, and a further 15% initially said yes before changing their answers. Undergraduates do not necessarily fare better: they do not interpret logical language consistently, and they struggle to work with conditionals' contrapositives and negations (e.g., Hub & Dawkins, 2018; Norton et al., 2025).

This matters for mathematical reasoning because these logical distinctions are closely related to inference validity. One valid *inference form* is *modus ponens*, illustrated here.

Conditional premise: If x is less than 2, then x is less than 5.

Categorical premise: x is less than 2.

Conclusion: x is less than 5.

Modus ponens inferences form the backbone of mathematical reasoning: arguments often combine a conditional claim 'if $A(x)$, then $B(x)$' with a categorical claim that $A(x)$ is true to conclude that $B(x)$ is true. But other inferences are tempting, and only some are valid. Table 2 exhibits valid modus ponens (MP) and *modus tollens* (MT) inference forms, where the latter underpin proofs by contradiction and contraposition. It also exhibits invalid *denial of the antecedent* (DA) and *affirmation of the consequent* (AC) inference forms (cf., Evans et al., 2010).

Table 2

Conditional Inferences.

	Mathematical example	General case
Modus ponens MP (valid)	If x is less than 2, then x is less than 5. x is less than 2. *Conclusion*: x is less than 5.	If $A(x)$, then $B(x)$. $A(x)$. *Conclusion*: $B(x)$.
Denial of the antecedent DA (invalid)	If x is less than 2, then x is less than 5. x is not less than 2. *Conclusion*: x is not less than 5.	If $A(x)$, then $B(x)$. Not $A(x)$. *Conclusion*: Not $B(x)$.
Affirmation of the consequent AC (invalid)	If x is less than 2, then x is less than 5. x is less than 5. *Conclusion*: x is less than 2.	If $A(x)$, then $B(x)$. $B(x)$. *Conclusion*: $A(x)$.
Modus tollens MT (valid)	If x is less than 2, then x is less than 5. x is not less than 5. *Conclusion*: x is not less than 2.	If $A(x)$, then $B(x)$. Not B(x). *Conclusion*: Not A(x).

Making or accepting invalid DA or AC inferences has consequences for reasoning because conditionals are ubiquitous in mathematics. By upper-secondary and tertiary education, they appear routinely in theorems such as 'If $f: \mathbb{R} \to \mathbb{R}$ is differentiable at a, then f is continuous at a'. This theorem allows us to infer that a differentiable function must be continuous (MP) and that a non-continuous function cannot be differentiable (MT). It does

not allow us to infer that a non-differentiable function is non-continuous (DA) or that a continuous function is differentiable (AC).

We want students to avoid such errors, so distinguishing valid from invalid inferences is key to developing mathematical reasoning. Yet there has been little research on conditional inference in mathematics education. This paper therefore follows others (e.g., Datsogianni et al., 2020; Leron & Hazzan, 2006; Vamvakoussi et al., 2012) in leveraging extensive work on non-mathematical reasoning. We investigate a question important for mathematics education (which inferences do students accept as valid?), using a well-established general method (conditional inference tasks) adapted to mathematical content. This approach isolates one aspect of reasoning, permitting its systematic study.

Theoretical Background

To frame our research, we first review findings on non-mathematical conditional inference, discussing contemporary theoretical accounts and believability effects. We then discuss two strands of work on conditional inference in mathematics students, arguing that mathematical conditionals are likely to vary in believability and considering associated methodological issues.

Conditional Inference with Abstract and Causal Conditionals

Conditional inference has been studied extensively in psychology (Oaksford & Chater, 2020), where tasks usually spread items equally across MP, DA, AC and MT inferences. Abstract tasks often use conditionals about imaginary letter-number pairs ('if the letter is A, then the number is 3'). For these, educated adults who are not mathematics specialists accept nearly all valid MP inferences, but also accept many invalid DA and AC inferences, and reject many valid MT inferences. Evans et al. (2007), for instance, reported acceptance rates of 98%, 47%, 74% and 50% for MP, DA, AC and MT respectively. This is some distance from the normative 100%, 0%, 0%, 100%, and is fairly typical – AC acceptance is often over 50% and higher than both DA and MT acceptance (Nickerson, 2015).

Tasks with everyday content often involve *causal* conditionals ('if John studies hard, then he does well on the test') (e.g., De Neys et al., 2003). Instructions vary: some are 'deductive', requiring a yes/no response on whether conclusions follow necessarily (e.g., Evans et al., 2010); others ask for scaled ratings of belief/certainty that conclusions follow (e.g., Cummins et al., 1991). Either way, response patterns broadly reflect those for abstract tasks, except that people typically accept more inferences from conditionals that are more *believable*.

Belief refers to 'the attitude we have, roughly, whenever we take something to be the case or regard it to be true' (Schwitzgebel, 2024). Accounts for conditional inference assume that beliefs can be held with more or less certainty[1]: the perceived believability of a conditional 'if $A(x)$, then $B(x)$' rests on the extent to which $B(x)$ appears to a reasoner to follow from $A(x)$. This means that perceived believability depends upon a reasoner's familiarity with situations or objects x for which $A(x)$, and $B(x)$ are true or false, and that beliefs might depart from accepted evidence or truth. This conceptualisation is consistent with the tripartite model (de Grefte, 2023; Ichikawa & Steup, 2024) in which knowledge is *justified true belief*, so that an individual's knowledge about a concept forms a subset of their beliefs about that concept.

In empirical studies of conditional inference, believability is usually operationalised by asking participants to provide believability ratings (e.g., Evans et al., 2010) or to list distinct counterexamples (maybe John cheats, etc.); list lengths are then averaged (e.g., Cummins, 1995). Research has reported both individual differences and results on *counterexample accessibility*. Evans et al. (2010), for instance, found that under deductive instructions, participants with lower AH4 general intelligence scores were affected by believability, whereas those with higher scores ignored believability and rejected higher proportions of invalid inferences. Cummins (1995) reported that if people can readily think of counterexamples to a conditional, they are less likely to accept MP inferences; if they can readily think of cases in which a conditional's consequent occurs without its antecedent ($B(x)$ without $A(x)$), they are less likely to accept AC inferences. Acceptance decreases further with the number of counterexamples generated or presented in the task (De Neys et al., 2003).

Such findings inform *probabilistic* accounts of conditional inference, in which the subjective probability P(if p then q) is the modelled as the subjective conditional probability P($q|p$), with inferences accepted when, given the premises, the probability of the conclusion is high (Over et al., 2007). For instance, Oaksford et al. (2000) had participants estimate P(p), P(q) and P($q|p$) for both experimentally manipulated frequencies and everyday conditionals ('if a person is a politician then they are privately educated'); they found that theoretical probabilities compared well with actual inference acceptance.

[1] Because of this assumption, some philosophers would prefer the term 'credence' rather than 'belief' (see Jackson, 2020).

Dual process/strategy versions of these accounts propose that individuals can reason using both probabilistic and *counterexample* strategies. A probabilistic strategy generates rapid estimates based on a wide variety of information; a counterexample strategy rejects inferences based on specific counterexamples, is slower, and demands working memory (Markovits et al., 2013). Strategy seems to be influenced by time and cognitive capacity: Markovits et al. (2013) found that reasoners were more likely to use a counterexample strategy with unlimited than limited time; De Neys et al. (2005) found that for everyday causal conditionals, individuals with higher working memory capacities responded more normatively.

Overall, there is widespread agreement that a dual strategy account is necessary to explain decades of findings on everyday reasoning (for reviews see Evans, 2019; Leron & Hazzan, 2006; Oaksford & Chater, 2020; Over & Evans, 2024). We therefore take this account as a starting point, assuming that people use the same cognitive apparatus when reasoning about mathematics as when reasoning about everything else. Indeed, a probabilistic strategy aligns well with mathematical intuition: reasoners often get a 'first impression' about truth or validity that is not informed by conscious consideration of logic or specific examples (Fischbein, 1987). A counterexample strategy is important because a single counterexample renders a mathematical claim or inference invalid.

However, there also exist strategies viable *only* in mathematics. In mathematics, we can often do better than observing that no counterexamples come to mind – we can be certain that none exist, that a conclusion follows *deductively*. Mathematically trained people might be able to construct deductive arguments; where they cannot do this, or cannot do it quickly, they might nevertheless think it more or less plausible that such arguments exist (Inglis et al., 2007). Such thinking informs expert choices in problem solving (Pólya, 2014); Corfield (2001) described it as Bayesianism in mathematics, and Gowers (2023) provided an extended account of factors that make unproven conjectures more or less plausible. Here, we hypothesize that just as accessible counterexamples should decrease inference acceptance, accessible deductive arguments – or accessible signs that deductive arguments exist – should *increase* it.

Moreover, mathematics affords the possibility of ignoring content altogether and focusing on inference form. This is rarely viable in everyday reasoning, where a rational approach takes account of knowledge and experience (Oaksford & Chater, 2007). In mathematics, inference forms are valid or invalid independently of content, and – at least in

advanced mathematics – the ability to focus on form is desirable. It is possible, therefore, that mathematically trained people might respond normatively to conditional inference tasks.

Conditional Inference in Mathematics Students

Evidence pertinent to these various possibilities exists, but is somewhat limited. On abstract conditional inference tasks, mathematics undergraduates are more likely than peers in other subjects to reject invalid inferences: Inglis and Simpson (2008) reported acceptance rates of 98%, 13%, 22% and 69% for MP, DA, AC and MT inferences, and we see progress toward this in those studying mathematics intensively at age 16-18 (Attridge et al., 2015). This is consistent with the idea that mathematics trains students to check for counterexamples or to recognise inference forms (deductive arguments are not relevant for abstract content). Notably, the improvement is not matched for valid MT inferences: Attridge and Inglis (2013) reported that specialist mathematical study at 16-18 led students to reject slightly higher proportions of these. This weighs against the possibility that students recognise inference form, but Attridge and Inglis suggested that it might reflect greater alertness to the possibility of counterexamples, making students more sceptical and more likely to reject inferences that are difficult to justify (Inglis & Attridge, 2016).

Abstract tasks provides insight into the effects of mathematical training on general reasoning. But they might not reflect how students reason *in mathematics*. Mathematics is abstract but not content-free – on the contrary, it has meaningful semantic content. So it could be that students who specialise in mathematics perform imperfectly in abstract tasks, but are cued by mathematical content to make more normatively valid evaluations. This would be consistent with Braithwaite's (2025) finding that even people without specialist training judge mathematical conditionals falsified by rare exceptions.

The possibility that people reason differently about mathematics has, however, been little investigated. Conditional inference tasks with mathematical content have usually involved only small numbers of items, often with non-standard phrasing. For instance, Stylianides et al. (2004) studied one DA and one MT item with everyday content, and one mathematical contraposition item with non-standard phrasing and notation: participants were asked to consider '$x = y \Rightarrow x^2 = y^2$' as a proof for 'If $x^2 \neq y^2$ then $x \neq y$ (where $x, y \in$ **N**)'. Datsogianni et al. (2020) studied all four inferences for everyday and mathematical conditionals, but again with non-standard phrasing: for instance, 'if the box contains exactly two blue diamonds and three red diamonds, then the diamonds in the box are worth 12 gold coins' has an irrelevant 'and' clause in the antecedent and, unlike typical conditionals in mathematics, lacks an inferential link between antecedent and consequent. Case and Speer

(2021) studied all four inferences using calculus theorems ('For all functions f, if f is differentiable at a point $x = c$, then f is also continuous at the point $x = c$') and abstract but mathematical-sounding propositions ('For integers a and b, if $a§b$ then $aba§bab$'). Their tasks were non-standard in that categorical premises were instantiated ('Suppose h is a function that is continuous at $x = 7$'; 'Suppose that (7)(4)(7)§(4)(7)(4) is true'). Finally, Durand-Guerrier (2003) studied all four inferences for three of her six items, which came from textbooks so departed considerably from standard phrasing. She also used open 'What can be said…?' prompts and classified answers as positive, negative or contingent (*can't tell* or *neither necessarily true nor necessarily false*).

These methodological differences mean that most existing work with mathematical content is not directly comparable with broader research on everyday reasoning. An exception is work by Braithwaite and Rafferty (2025), who studied inferences from conditionals with algebraic content ('if m and n are both even, then $m + n$ is even'). Across four studies, they provided evidence that individual differences in knowledge of mathematical examples affected responses. That result aligns with Durand-Guerrier's (2003) claim that students treat conditional inferences differently depending upon their mathematical content. It also suggests that mathematical conditionals – like everyday conditionals – might vary in believability.

On mathematical believability and its possible variation, there is additional direct evidence from students' acceptance of invalid but plausible-sounding inferences (Alcock & Weber, 2005), and from findings that people are more likely to reject invalid geometric inferences when counterexamples are more accessible from provided diagrams (Hamami et al., 2021). There is also substantial indirect evidence, because individuals' concept images (Tall & Vinner, 1981) do not correspond to formally defined sets, leading to errors of omission or overgeneralisation. A conditional like 'if X is a square, then it is a rhombus' is true but might have relatively low believability, because everyday pragmatics tends to treat the sets of squares and rhombi as mutually exclusive. Conversely, the conditional 'If $x < 3$ then $1/x > 1/3$' is false but not immediately unbelievable – Alcock and Attridge (2023) reported that about one fifth of mathematics undergraduates initially judged this true.

In the present research, we measured perceived believability for a range of mathematical conditionals, and investigated believability effects in mathematical conditional inference. We also compared mathematics students' acceptance of inferences from

mathematical, abstract and everyday causal conditionals. We addressed two main research questions:

1. Is mathematical conditional inference affected by perceived believability?
2. Do mathematics undergraduates treat conditional inference similarly across mathematical, abstract and everyday content?

We also followed up with a third question, designed to address a possible confound:

3. Are apparent believability effects in mathematical conditional inference better understood as arising from perceived believability or perceived easiness?

Methodology

To investigate believability, we must measure it. Unfortunately, measurement approaches used for everyday conditionals do not transfer well to mathematics. Using scaled believability ratings is arguably not viable because mathematically informed people will likely interpret mathematical conditionals as either true or false. There are subtleties – mathematicians are often willing to describe a conjecture 'if p then q' as 'probably true' (Gowers, 2023). However, this does not mean believing that $P(q|p)$ is high; it means believing that $P(P(q|p) = 1)$ is high (Inglis, 2006). Even if allowed to rate believability on that basis, we anticipated that participants would struggle; quantifying the probability might push them into zero-or-one judgements.

Also not viable is asking for distinct counterexamples. For a true mathematical conditional (that is not a true biconditional), MP and MT inferences have no counterexamples: nothing could prevent a number less than 2 from being less than 5. DA and AC inferences are invalid, but mathematical counterexamples often occur singularly ('the claim fails for zero') or in infinite sets ('the claim fails for the negative numbers'). The task of listing counterexamples would therefore be artificial or impossible.

We therefore measured believability relativistically, using *comparative judgement* (CJ). CJ asks judges to consider pairs of stimuli, stating which is 'better' in relation to a construct. Multiple judgements made by multiple judges are then used to generate scores via the Bradley-Terry model (Bradley & Terry, 1952), where scores represent the judges' collective views. In mathematics education, CJ has been used to assess constructs including conceptual understanding (Jones et al., 2019), problem solving (Jones & Inglis, 2015), and conceptions of proof (Davies et al., 2021). All are hard to define, but we expect experts to broadly agree on their application. Believability shares these characteristics.

When using CJ, reliability is an issue: judges must agree enough for results to be meaningful. Misfitting judges can be identified and, if there is evidence that they have not taken the task seriously – perhaps very short judgement times – their data might be removed. For the retained judges and resulting scores, reliability is usually reported using Scale Separation Reliability (SSR) and Inter-Rater Reliability (IRR) (Jones & Davies, 2024). SSR is seen as analogous to Cronbach's alpha and is interpreted using the same .70 threshold. IRR is often lower (Verhavert et al. 2019) but again is often compared with .70. The main driver of reliability is the number of judgements (to a point – more judgements cannot artificially 'make' scores reliable). Verhavert et al. (2019), in a meta-analysis of 49 CJ assessments, found that reliability of .70 usually required 10-14 comparisons per response, and reliability of .90 required 26-37. This informed our design in Study 1.

Study 1: Believability and Conditional Inference in Mathematics

Study 1 investigated believability in mathematical conditionals and its effect on inference acceptance. In Study 1a, mathematics education researchers and mathematics undergraduates comparatively judged 40 mathematical conditionals, yielding group-level scores of relative perceived believability (henceforth, 'perceived believability'). We report reliability within groups, agreement across groups, and relationships between perceived believability and truth. In Study 1b, we constructed a conditional inference task using true mathematical conditionals with higher and lower perceived believability. We report effects of perceived believability on inference acceptance.

Study 1a Method

Eight mathematics education researchers were each asked to generate five mathematical conditionals with content that would be familiar to typical students aged 13-14, and therefore basic for mathematics undergraduates. The researchers were instructed that, to parallel everyday tasks, their conditionals should:

- Cover a range of topics;
- Have plausibly related antecedent and consequent;
- Not be obviously false;
- Not use additional connectives ('not', 'and', 'or') in the antecedent or consequent;
- Vary in believability.

Seven of the resulting conditionals were discarded because their consequents could not readily be asserted as categorical premises. Six were discarded because they had true converses, which would make DA and AC inferences valid for semantic reasons. One

converse was itself suitable, so we kept that and made the collection up to 40 using conditionals from the literature (Alcock, 2013; Alcock & Attridge, 2023; Dawkins & Norton, 2022; Durand-Guerrier, 2003; Houston, 2009; Hoyles & Küchemann, 2002; Selden & Selden, 2003). We standardised by phrasing each conditional with a comma and 'then' between antecedent and consequent, and by removing extra words ('must', 'also') in the consequent. The 40 conditionals (see Table 3) were typeset in a large LaTeX font and uploaded to nomoremarking.com, an online CJ engine.

To balance reliability with avoiding judge fatigue (Verhavert et al., 2019), we aimed for 20 judgements per conditional, requiring $40 \times 20/2 = 400$ in total (each judgement involves two conditionals). The same eight researchers were therefore each asked to make 50 judgements, comparing randomly selected pairs of conditionals under the prompt 'Which is more believable?'. We also recruited 12 mathematics undergraduates from elective mathematics education courses at Nottingham and Loughborough Universities; each was asked to make 50 judgements. There were no time restrictions, and judgement times varied substantially (mean and median response times 18.1 and 11.3 seconds respectively, interquartile range 11.1 seconds), suggesting differing strategy use in line with the dual strategy account. For each group, the judgements were fitted to the Bradley-Terry Model to generate scores as z-transformed parameter estimates.

Study 1a Analysis and Results

All analyses reported in this paper were pre-registered to assist readers in judging the severity of our hypothesis tests (Lakens, 2019). For this study, pre-registered at https://aspredicted.org/ts39g.pdf, the researchers' scores had SSR = .84 and IRR = .73, so were sufficiently reliable. The undergraduates' scores had SSR = .83 and IRR = .63; given the high SSR, we judged reliability adequate. Thus, CJ generated meaningful group-level scores of perceived believability. Moreover, the researchers and undergraduates broadly agreed: the correlation between their perceived believability scores was $r = .74$, $p < .001$ (see Figure 2). Table 3 lists all 40 conditionals with their truth values and with researcher and undergraduate perceived believability ranks and scores.

Figure 2

Undergraduate Against Researcher Group-Level Perceived Believability Scores; Error Bars Show ± 1 *Standard Error of the Mean.*

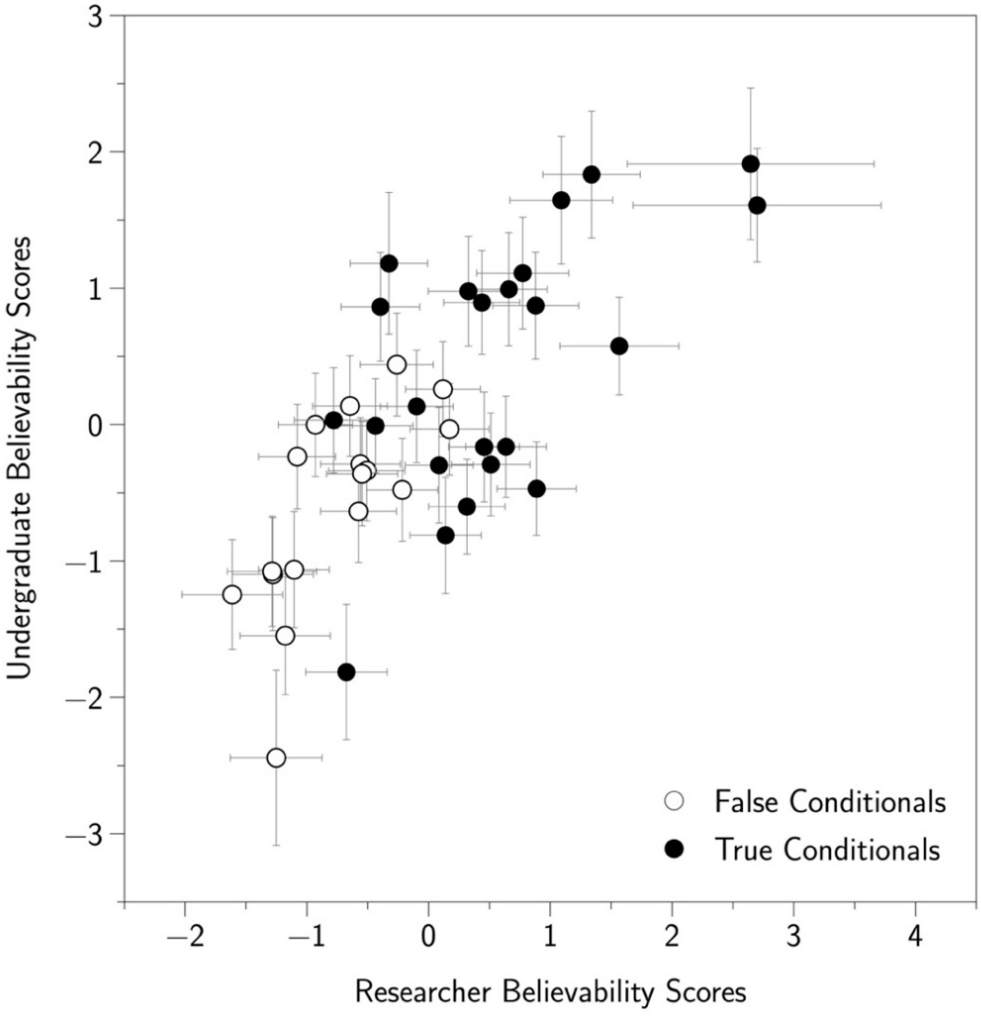


Table 3

Conditionals with Believability Rankings and Scores (as z-scores) for Researchers and Undergraduates, Ordered by Researchers' Scores.

Conditional	Truth Value	Researcher Rank	Researcher Score	Undergrad Rank	Undergrad Score
If $x < 2$, then $x < 5$.	True	1	2.70	4	1.61
If n is a multiple of 6, then n is a multiple of 3.	True	2	2.64	1	1.91
If line L is tangent to circle C, then L is perpendicular to a radius of C.	True	3	1.57	12	0.58
If n is a multiple of 4, then n^2 is a multiple of 4.	True	4	1.34	2	1.83
If C is a circle, then its width is the same when measured in any direction.	True	5	1.09	3	1.64
If polygon P is a square, then it is a rhombus.	True	6	0.89	29	-0.47
If $a = b$, then $an = bn$.	True	7	0.88	10	0.87
If $x = \sqrt{y}$, then $x^2 = y$.	True	8	0.77	6	1.11
If $x = -4$, then $x^2 + x - 12 = 0$.	True	9	0.66	7	0.99
If X is a circle, then X is an ellipse.	True	10	0.64	21	-0.16
If circle C and square S have the same perimeter, then C has bigger area than S.	True	11	0.51	25	-0.29
If polygon P is a rhombus, then it has perpendicular diagonals.	True	12	0.46	22	-0.16
If n is the product of two consecutive integers, then n is even.	True	13	0.44	9	-0.90
If $x < 0$ then $x^3 < x^2$.	True	14	0.33	8	0.98
If $x^2 = y^2$, then $xy = yx$.	True	15	0.32	31	-0.60
If fraction x has denominator 7, then it is equivalent to a non-terminating decimal.	False	16	0.17	20	-0.03
If quadrilateral Q is cyclic, then it is convex.	True	17	0.14	33	-0.81
If x is an integer, then $x^2 > x$.	False	18	0.12	14	0.26
If $x - 12{,}345 = 0.67$, then $x > -12{,}345.67$.	True	19	0.09	26	-0.30
If polygon P is a rectangle, then every line through its centre cuts its area in half.	True	20	-0.10	16	0.13
If n is the sum of four consecutive numbers, then n is a multiple of 4.	False	21	-0.22	30	-0.48

If line L is tangent to curve C, then L intersects C at only one point.	False	22	-0.26	13	0.44
If $x = 3$, then $2(x - 3) = 5x - 3(x + 2)$.	True	23	-0.33	5	1.18
If the product of two whole numbers is odd, then their sum is even.	True	24	-0.40	11	0.86
If rectangle R has area 10cm^2, then its perimeter is greater than 10cm.	True	25	-0.44	19	-0.01
If $a > b$, then $a^2 > b^2$.	False	26	-0.51	27	-0.34
If $x < 3$, then $1/x > 1/3$.	False	27	-0.55	28	-0.36
If $a > b$, then $ac > bc$.	False	28	-0.56	24	-0.29
If equation E is quadratic, then it has exactly two roots.	False	29	-0.58	32	-0.64
If n is prime, then $n + 1$ is even.	False	30	-0.65	15	0.14
If quadrilateral Q has a reflex angle, then it will tesselate.	True	31	-0.67	39	-1.81
If $\sin x > 0$, then $\cos x < 1$.	True	32	-0.78	17	0.03
If n is a multiple of 13, then it has an even number of factors.	False	33	-0.93	18	0.00
If x is positive, then $\tan x > \sin x$.	False	34	-1.08	23	-0.23
If the side lengths of rectangle R are doubled, then its area is doubled.	False	35	-1.10	34	-1.06
If function f is polynomial, then f has a real root.	False	36	-1.18	38	-1.55
If the mean of dataset D is greater than 100, then the median is greater than 100.	False	37	-1.25	40	-2.44
If the mean of dataset D is 7, then the median is 7.	False	38	-1.28	36	-1.10
If $a = 42$, then $a \times b > 42$.	False	39	-1.29	35	-1.08
If composite number c ends in a 3, then it is a multiple of 3.	False	40	-1.61	37	-1.25

Regarding truth, the results aligned with our theoretical suggestion that some true conditionals might have relatively low perceived believability and some false conditionals might have relatively high perceived believability. For researchers, true conditionals (N = 23) on average received higher scores (M = 0.554, SD = 0.907) than false conditionals (N = 17, M = −0.749, SD = 0.519); this difference was significant, $t(38) = 5.31$, $p < .001$. For undergraduates, true conditionals (N = 23) on average received higher scores (M = 0.435, SD = 0.952) than false conditionals (N = 17, M = −0.589, SD = 0.742); this difference was also significant, $t(38) = 3.68$, $p < .001$. However, both groups collectively judged some false conditionals more believable than some true conditionals (again, see Figure 2).

Study 1a Discussion

Study 1a established that mathematically educated people agree about which conditionals are more believable, both within and across expertise levels[2]. It also established that perceived believability is distinct from mathematical truth. However, it is theoretically possible that people can judge relative believability for mathematical conditionals without perceived believability affecting their reasoning. Study 1b investigated this possibility.

Study 1b Method

Using the results from Study 1a, we constructed a mathematical conditional inference task to parallel everyday causal tasks (e.g., Cummins, 1995). We avoided conditionals for which the antecedent referred to a specific object ('If $x = -4$, then $x^2 + x - 12 = 0$') as these are unlike typical mathematical theorems. To avoid a truth-value confound, we selected four true conditionals with higher perceived believability and four with lower perceived believability, using sets that were similar in length and in the groups' perceived believability ranks.

High perceived believability:

- If x is less than 2, then x is less than 5.
- If n is a multiple of 6, then n is a multiple of 3.
- If line L is tangent to circle C, then L is perpendicular to a radius of C.
- If $x = \sqrt{y}$, then $x^2 = y$.

Low perceived believability[3]:

[2] For undergraduates, Study 3a effectively replicated this study with a larger sample ($N = 105$). For the 17 conditionals common to Studies 1a and 3a, the correlation between undergraduate perceived believability scores was high, $r = .823$.

[3] The first of the lower-believability conditionals is true due to its true consequent (assuming that x and y are real numbers). Analyses for Study 1b with this item removed did not yield substantively different outcomes.

- If $x^2 = y^2$, then $xy = yx$.
- If rectangle R has area 10cm^2, then its perimeter is greater than 10cm.
- If quadrilateral Q has a reflex angle, then it will tesselate.
- If $\sin x$ is greater than 0, then $\cos x$ is less than 1.

Each conditional was combined with categorical premises and conclusions to create MP, DA, AC and MT inferences, yielding a 32-item task. Item order was randomised at the participant level, and participants responded to all 32 items so that inference type and perceived believability were within-subjects variables. Task instructions were deductive, requiring a yes/no response to whether the conclusion follows necessarily. This provided a strong test of the influence of perceived believability in mathematics.

Participants were 57 mathematics undergraduates at Loughborough University in the UK, where students specialize early. At age 16-18, English students usually study only three subjects; at university, they register for three- or four-year degree programmes involving one or two subjects. Data were collected during a lecture for an introduction-to-proof course taken by students in the first year of a degree with at least 75% mathematics or in the second year of a degree with at least 50% mathematics. The preceding week had covered content on logic including true and false propositions, negation, conjunction and disjunction, Boolean algebra, implication, necessary and sufficient conditions, and proof by contradiction.

Attendees were asked to spread out so that no-one was in adjacent seats. Each completed a booklet individually and in silence, reading a participant information sheet and completing a consent form if willing to contribute their data; 57 consented. They were allowed up to 20 minutes for the conditional inference task and were asked to note their finish time according to a clock on the screen. One participant completed fewer than half of the items, so their data were excluded according to our pre-registered criteria[4] (https://aspredicted.org/WT9_WV2). Responses from the remaining 56 were included in the analyses; 52 completed all items. This sample size gave us 80% power to detect a two-way interaction effect of $\eta_p^2 = .016$, with an alpha of 0.05, assuming a moderate correlation between repeated measures.

Study 1b Analysis and Results

The time allowed was comfortably adequate: mean reported time for participants who completed all items was 14.3 minutes (SD = 3.13 minutes). Scoring for normative validity

[4] Our pre-registration anticipated 80-100 participants, but attendance was reduced by a storm and an earlier lecture cancellation.

yielded a mean of 25.7/32 (SD = 3.21) or 80% (SD = 10.0%). Scores were not significantly correlated with time ($r = -.15$, $p = .308$), so there was no evidence that early finishers quickly generated normative answers or rushed at the expense of accuracy.

Mean percentage acceptance rates for inferences from conditionals with high and low perceived believability are shown in Table 4. Overall, participants accepted almost all valid MP inferences. They rejected around three quarters of invalid DA and AC inferences; reflecting results for mathematics undergraduates (Inglis & Simpson, 2008) and contrasting with non-specialists (Nickerson, 2015). They accepted high proportions of valid MT inferences – for conditionals with high perceived believability, the acceptance rate reached 90%. This contrasts with previous results: acceptance rates for non-specialists are often around 50%, and studies with mathematics undergraduates and abstract content have reported rates of around 60-70% (albeit with tasks made more challenging by a negations paradigm[5], e.g., Attridge & Inglis, 2013; Inglis & Simpson, 2008, 2009). Therefore, although not normatively perfect, these mathematics undergraduates applied standard logic relatively well to mathematical content.

Table 4

Mean Percentage Acceptance Rates for Inferences from Conditionals with High and Low Group-Level Perceived Believability (with Standard Deviations).

Believability	MP	DA	AC	MT
High	97.0 (8.3)	30.3 (22.7)	28.2 (22.4)	90.1 (18.2)
Low	92.5 (17.1)	23.2 (22.8)	26.0 (20.1)	70.0 (26.8)

To address effects of perceived believability on inference acceptance, responses were subjected to a 4 (inference type: MP, DA, AC, MT) × 2 (believability: high, low) within-subjects Analysis of Variance (ANOVA), with a Greenhouse-Geisser correction where assumptions of sphericity were violated. This revealed significant main effects of inference type, $F(1.97, 142.21) = 333.85$, $p < .001$, $\eta_p^2 = .859$, and perceived believability, $F(1,165) = 13.23$, $p < .001$, $\eta_p^2 = .261$, and a significant inference type × perceived believability interaction effect, $F(2.59, 142.21) = 6.32$, $p < .001$, $\eta_p^2 = .103$. The main effect of perceived believability reflected the fact that participants accepted more inferences from conditionals

[5] A negations paradigm counterbalances use of 'If A then B', 'If not-A then B', 'If A then not-B' and 'If not-A then not-B'. This washes out negative conclusion effects but is not usually applied with everyday content.

with high perceived believability; the interaction reflected the fact that this was especially so for MT inferences.

Study 1b Discussion

Study 1b showed that mathematical believability is not an artificial construct: it was not mentioned to Study 1b participants, yet it predicted inference acceptance. Thus, Durand-Guerrier (2003) was right that semantic content affects mathematical conditional inference; our task, with its standard structure, demonstrates this cleanly. Moreover, while individual differences in, for example, content knowledge might also be expected to affect inference acceptance, our within-subjects design means that these cannot easily account for this effect. We expand upon this point when discussing Study 2.

Study 2: Conditional Inference Across Content Types

Study 2 investigated whether Study 1's results are specific to mathematical content. We used an inference task with mathematical, everyday causal, and abstract conditionals.

Study 2 Method

From the true mathematical conditionals used in Study 1b, we selected the two with highest and the two with lowest perceived believability. From Evans et al.'s (2010) causal conditionals, we selected two rated as having relatively high and two rated as having relatively low believability. We constructed two abstract conditionals using letter-number pairs.

Mathematical high perceived believability

- If x is less than 2, then x is less than 5.
- If n is a multiple of 6, then n is a multiple of 3.

Mathematical low perceived believability

- If quadrilateral Q has a reflex angle, then it will tesselate.
- If $\sin x$ is greater than 0, then $\cos x$ is less than 1.

Causal high believability

- If car ownership increases, then traffic congestion will get worse.
- If nurses' salaries improve, then recruitment of nurses will increase.

Causal low believability

- If divorce is made more difficult, then the number of marriages will decrease.
- If foreign investment is encouraged, then the UK car industry will revive.

Abstract

- If the letter is C, then the number is 7.

- If the letter is N, then the number is 2.

Each conditional was paired with categorical premises and conclusions to give MP, DA, AC and MT inferences, yielding a 40-item task. As in Study 1b, participants responded to all items so inference type and perceived believability were within-subjects variables. Subtask order was counterbalanced so that approximately one sixth of participants responded in the order abstract-causal-mathematical, one sixth abstract-mathematical-causal, and so on. Each subtask had an instruction page asking for yes/no responses to whether the conclusion follows necessarily. Within each subtask, item order was randomised at the participant level. Booklets were organised so that adjacent participants would see subtasks in different orders.

Participants were 81 undergraduates from the University of Nottingham in the UK. All were in the first year of a degree comprising at least 50% mathematics. Data were collected in an introduction-to-proof course that had covered content on logic including true and false propositions, negation, conjunction and disjunction, necessary and sufficient conditions, quantifiers, and direct and indirect proof. The class was taught in three groups; in all three, attendees were asked to spread out so that they had no-one in adjacent seats. Each received a booklet, which they completed individually and in silence, reading a participant information sheet and completing a consent form if willing to contribute their data; 81 consented. They were asked to close their booklets once finished, and all three groups finished in approximately 26 minutes.

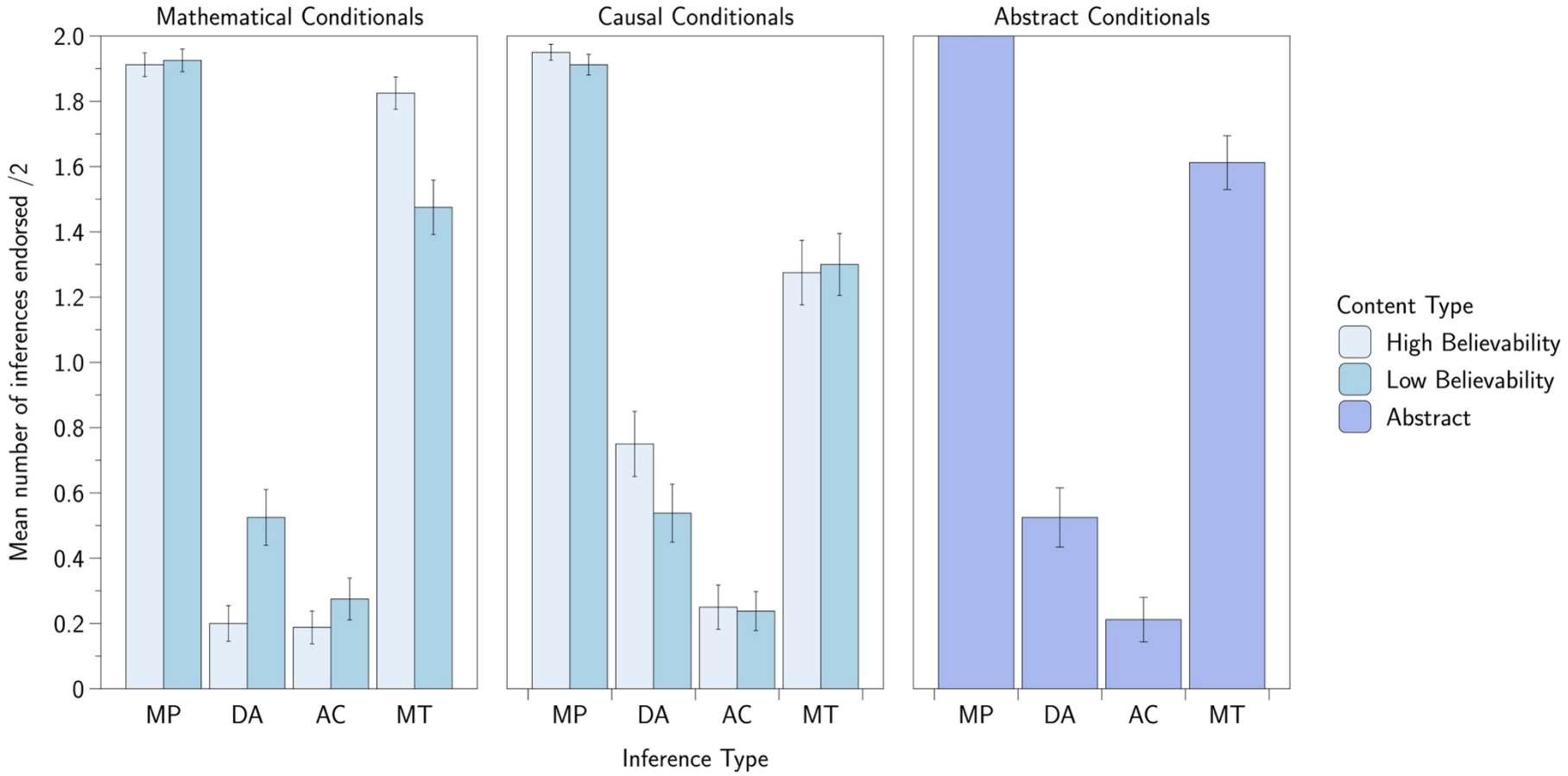
Analyses followed our pre-registered plan (https://aspredicted.org/LXM_7FS). One participant completed fewer than 90% of the items, leaving 80 participants for the analyses. This sample size gave us 80% power to detect a three-way interaction effect of $\eta_p^2 = .007$, with an alpha of 0.05, assuming a moderate correlation between repeated measures.

Study 2 Analysis and Results

Figure 3 shows mean acceptance rates for the four inference types across the five content types. As in Study 1b, responses were more normative than is typical for non-specialists, with MP acceptance close to 100%, MT acceptance high, and invalid inference acceptance low, especially for AC items. Again, this should not be over-interpreted for abstract items because we did not use a negations paradigm. For causal items, comparisons can be more direct.

Figure 3

Mean Numbers of Inferences Accepted from High and Low Believability Mathematical and Causal Conditionals and from Abstract Conditionals; Error Bars Show ±1 *Standard Error of the Mean.*



For our main analyses, we first investigated perceived believability effects for inferences from mathematical and everyday causal conditionals. We used a 4 (inference type: MP, DA, AC, MT) × 2 (content: causal, mathematical) × 2 (believability: high, low) within-subjects ANOVA, applying a Greenhouse-Geisser correction where assumptions of sphericity were violated. This revealed a significant main effect of inference type, $F(1.93,212.32) = 217.889$, $p < .001$, $\eta_p^2 = .734$, significant interactions of both content and perceived believability with inference type (content × inference type, $F(2.30,212.32) = 19.517$, $p < .001$, $\eta_p^2 = .198$; perceived believability × inference type, $F(2.64,212.32) = 4.753$, $p = .005$, $\eta_p^2 = .057$), and a significant three-way inference-type × content × perceived believability interaction, $F(2.69,212.32) = 16.936$, $p < .001$, $\eta_p^2 = .177$. As shown in Figure 3, the two-way interaction between perceived believability and inference type changed across content, with the effect driven by close-to-normative responses for inferences from mathematical conditionals with high perceived believability.

We then compared inferences from mathematical and abstract conditionals using a 4 (inference type: MP, DA, AC, MT) × 3 (content: abstract, mathematical high perceived believability, mathematical low perceived believability) within-subjects ANOVA, again applying a Greenhouse-Geisser correction where assumptions of sphericity were violated. This revealed a significant main effect of inference type, $F(2.08,367.89) = 317.05$, $p < .001$, $\eta_p^2 = .801$, and a significant inference-type × content interaction, $F(4.66,367.89) = 9.16$, $p < .001$, $\eta_p^2 = .104$. The interaction was again driven by close-to-normative responses for inferences from mathematical conditionals with high perceived believability; response patterns for abstract content resembled those for other content types.

These results, however, masked individual differences, which we investigated in a novel way. Usually, analyses of individual differences in conditional inference have not yielded clean results. Attridge and Inglis (2013), for instance, assessed the extent to which mathematics students' responses aligned with four theoretical interpretations of the conditional:

- Material (accept MP, reject DA, reject AC, accept MT);
- Defective (accept MP only);
- Biconditional (accept all four inferences);
- Conjunction (accept MP, reject DA, accept AC, reject MT).

They reported a shift from biconditional toward defective interpretations, but this meant that responses consistent with a biconditional interpretation dropped from around 75% to 67%,

and responses consistent with a defective interpretation rose from around 45% to 55%. Students neither began nor ended with interpretations matching theoretical profiles.

We therefore took a bottom-up approach. For participants who completed all items, the 20 acceptance rates (acceptances out of two for each of four inference types by five content types) were subjected to a hierarchical cluster analysis, using Ward's (1963) clustering criterion. This supported a three-cluster solution, with mean acceptances across content, perceived believability and inference types shown in Table 5.

Table 5

Mean Percentage Acceptance Rates for Inferences from Conditionals with Different Content, Broken Down by Cluster; H = High Perceived Believability, L = Low Perceived Believability.

	Cluster 1 (N = 38)				Cluster 2 (N = 24)				Cluster 3 (N = 18)			
	MP	DA	AC	MT	MP	DA	AC	MT	MP	DA	AC	MT
Math H	100	5	0	100	98	19	8	79	84	8	31	89
Math L	100	5	3	92	98	44	8	42	86	47	44	78
Caus H	100	14	1	92	100	54	13	29	89	64	36	50
Caus L	100	5	5	94	100	54	2	29	81	36	39	53
Abs	100	4	0	94	100	38	8	48	100	59	36	97

This contrasts starkly with typical results. Cluster 1 responded almost normatively across all content (with the possible exception of DA inferences from highly believable causal conditionals). This has not been reported elsewhere, so is likely an effect of mathematical training (or of selection into mathematics by students inclined to reason normatively, cf. Inglis & Attridge, 2016). The smaller Cluster 3 responded relatively normatively for mathematical conditionals with high perceived believability, but similarly across all other content, accepting invalid and MT inferences at rates similar to those seen in nonspecialists. Cluster 2 formed an interesting middle case. They too responded relatively normatively for mathematical conditionals with high perceived believability. For other content types, they accepted low proportions of AC inferences but much higher proportions of DA inferences. This is unusual (Nickerson, 2015) and has educational implications, as discussed below.

Study 2 Discussion

Study 2 provided further insight into believability effects in mathematical conditional inference. Across mathematical, everyday and abstract content, it revealed clusters of participants with distinct response profiles. It could be that such profiles were overlooked in earlier work: studying theoretically informed profiles would pick up Cluster 1 (responding normatively), but would pick up Cluster 3 only implicitly (as 'more biconditional'), and would overlook Cluster 2. This, we think, vindicates our bottom-up approach.

The clusters show that some mathematics undergraduates apply the normative logic of conditionals almost perfectly across all content – perhaps these students are alert to inference form. Others do not: they apply standard logic moderately well to mathematical conditionals with high perceived believability, but otherwise respond like non-specialists. Still others – our Cluster 2 – respond in a way that might reflect educational practice. Many transition-to-proof texts stress the distinction between a conditional and its converse (e.g., Houston, 2009). Fewer discuss the distinction between a conditional and its inverse (Alcock & Sa, 2025), which suggests that Cluster 2 have paid attention and consequently become good at rejecting AC inferences but not DA inferences. It also suggests that instruction on inverses might be effective for these students.

Overall, it seems that first-year UK mathematics undergraduates reason similarly across mathematical, abstract and everyday causal content, but more normatively from mathematical conditionals with high perceived believability. As in Study 1b, our within-subjects design means that this believability effect cannot easily be accounted for by individual differences on other factors such as content knowledge (or confidence, mathematical self-efficacy, etc.). It could be that individual factors *interact* with perceived believability – we consider this in our General Discussion. We rounded off this work, however, by testing a different possibility: that group-level perceived believability might index some more fundamental group-level variable, where the obvious candidate is 'easiness' of the mathematical content. Where content is more elementary, reasoners might work more fluently, readily thinking of relevant counterexamples or constructing deductive arguments, and feeling more confident that a conclusion does or does not follow. Also, true conditionals with more elementary content might seem more believable. This is relevant for interpreting the results of Study 2 because its mathematical conditionals with higher perceived believability involved more elementary concepts, so that apparent believability effects might be better understood as easiness effects. We addressed this possibility in Study 3.

Study 3: Believability or Easiness?

In Study 3, we asked whether apparent effects of perceived believability on conditional inference are better understood as effects of perceived easiness. Study 3a used CJ to score mathematical and everyday causal conditionals for both perceived believability and perceived easiness; Study 3b used a by-items regression to establish which variable better predicts inference acceptance.

Study 3a Method

We assembled 20 mathematical and 20 everyday causal conditionals with varying perceived believability. The mathematical conditionals came from Study 1a, which generated 23 true conditionals. We eliminated three for which the antecedent referred to a specific object and two for which the consequent could not, under the obvious interpretation, be false; the latter two we replaced by conditionals with related content[6]. We selected eight causal conditionals with varying believability from Evans et al. (2010), avoiding those that involved past political situations or that would likely be judged false. All had a future-tense 'will' in the consequent, which we removed for consistency with the mathematical conditionals. We selected a further 12 causal conditionals from Cummins (1995), three from each category of high/low average numbers of counterexamples to the conditional and its converse. All 40 conditionals were typeset in a large LaTeX font and uploaded to nomoremarking.com; see the Supplementary Materials at https://figshare.com/s/2ed67504674c50512bb3.

Participants were 215 first-year mathematics undergraduates taking an introduction-to-proof course at the University of Edinburgh in the UK. About three-quarters were studying mathematics or mathematics with another subject; most of the remainder were studying sciences or engineering. They had no prior instruction in logic and took part during their first course workshops. They were warned that comparing mathematical and everyday sentences might seem peculiar, then each was asked to complete 15 comparative judgements in response to the prompt 'Which is more believable?' or 'Which is easier to think about?'. They read a participant information sheet, signed a consent form if willing, and completed their judgements individually, in silence, on their own devices. Since each participant completed 15 randomly selected judgements, it was unlikely that neighbouring students would see the same conditionals at the same time. The task took under ten minutes.

Study 3a Analysis and Results

[6] One conditional removed was 'If $x^2 = y^2$, then $xy = yx$', for which this potential problem was noted in a footnote to Study 1b. Swapping out these items introduces no methodological problems because we investigated relationships between believability and easiness in general, not for specific items.

Analysis followed our pre-registered plan (https://aspredicted.org/xj4ar.pdf). We removed data from participants who did not complete 15 judgements, leaving participant and judgement numbers as in Table 6. Table 6 also shows reliability measures[7], demonstrating that mathematics undergraduates broadly agree about both believability and easiness[8]. There was no significant difference between mathematical and causal conditionals for perceived easiness ($t(38) = 0.166$, $p = .869$, $d = 0.052$) or perceived believability ($t(38) = 1.739$, $p = .090$, $d = 0.550$). There were, however, significant correlations between perceived believability and perceived easiness: for mathematical conditionals, the correlation was high ($r = .858$, $p < .001$); for causal conditionals it was moderate ($r = .520$, $p = .019$). This is in line with our theoretical observation that believability and easiness might be related.

Table 6

Participants, Judgements and Reliabilities for Conditional Believability and Easiness.

Prompt	Participants	Judgements	Approx judgements per item	SRR	IRR
Believability	105	1575	79	.88	.76
Easiness	110	1650	83	.84	.70

Study 3a Discussion

Study 3a confirmed that both perceived believability and perceived easiness can be scored reliably using CJ. It also showed that these constructs were too correlated to be treated as distinct predictors in a regression analysis, at least for our full lists of conditionals. We therefore selected subsets for Study 3b.

Study 3b Method

To disaggregate perceived believability and perceived easiness, we narrowed our lists to 10 mathematical and 10 causal conditionals, maintaining reasonable score ranges while reducing covariation. Table 7 lists the conditionals used, for which the correlation between perceived believability and perceived easiness was moderate for the mathematical conditionals ($r = .435$, $p = .208$) and low for the causal conditionals ($r = .141$, $p = .697$).

[7] Calculated from CJ with the entire set of conditionals and their converses; we do not use converses here.
[8] On believability, these participants also agreed with the undergraduates in Study 1a. For the 17 mathematical conditionals used in both studies, the correlation between CJ believability scores was $r = .823$.

Table 7

Mathematical and Causal Conditionals Used in Study 3b, with Believability and Easiness Scores. Numbered Mathematical Conditionals on Shaded Rows are Discussed under Study 3b Method.

Conditional	Content	Believability	Easiness
If n is a multiple of 4, then n^2 is a multiple of 4. [1]	Math	1.91	0.81
If line L is tangent to circle C, then L is perpendicular to a radius of C.	Math	1.54	0.99
If the product of two whole numbers is odd, then their sum is even. [3]	Math	1.29	-0.10
If C is a circle, then its width is the same when measured in any direction.	Math	1.07	0.90
If n is a multiple of 6, then n is a multiple of 3. [2]	Math	1.05	2.63
If n is the product of two consecutive integers, then n is even.	Math	0.85	0.11
If polygon P is a rectangle, then every line through its centre cuts its area in half.	Math	0.51	0.36
If $x^3 = y^3$, then $x^2 = y^2$.	Math	0.28	0.37
If polygon P is a rhombus, then it has perpendicular diagonals.	Math	0.26	-0.30
If $\sin x$ is greater than 0, then $\cos x$ is less than 1. [4]	Math	-0.47	-0.12
If Larry grasped the glass with his bare hands, then his fingerprints were on it.	Causal	1.39	0.17
If the gong was struck, then it sounded.	Causal	1.07	0.53
If the match was struck, then it lit.	Causal	0.79	0.01
If Joe cut his finger, then it bled.	Causal	0.73	0.84
If nurses' salaries improve, then recruitment of nurses increases.	Causal	0.27	0.43
If John studied hard, then he did well on the test.	Causal	-0.11	0.27
If the apples were ripe, then they fell from the tree.	Causal	-0.23	-0.15
If fertilizer was put on the plants, then they grew quickly.	Causal	-0.24	0.50
If foreign investment is encouraged, then the UK car industry revives.	Causal	-0.44	-0.63
If the trigger was pulled, then the gun fired.	Causal	-0.62	0.87

The distinction between the constructs is highlighted by Table 7's shaded rows: conditional [1] had high perceived believability and moderate perceived easiness; conditional [2] had moderate perceived believability and (very) high perceived easiness; conditional [3] had similarly moderate perceived believability but low perceived easiness; conditional [4] had low perceived believability and low perceived easiness. We as usual paired each conditional with categorical premises and conclusions to construct MP, DA, AC and MT inferences, yielding 40 mathematical and 40 causal items.

Participants were 175 first-year undergraduates from Durham University, UK. About two thirds were mathematics students; most of the remainder were studying natural sciences, physics or computer science. They participated in a lecture for a course called *Analysis I*, which involved introductory logic, including truth tables and proof by contraposition and contradiction. Each attendee received a booklet containing an information sheet, a consent form, task instructions, and five mathematical and five causal conditional inference items selected randomly from the banks of 40. Randomisation was such that each item appeared in an equal number of booklets; half of the booklets began with mathematical items and half began with causal items. Participants were asked to spread out so that, where possible, there was no-one in adjacent seats. Booklets were organised so that neighbouring participants would receive different subtask orders, and participants worked individually and in silence. The task took about 10 minutes, and 175 students consented to contribute data, giving approximately 22 responses per item.

Study 3b Analysis and Results

Analyses followed our pre-registered plan (https://aspredicted.org/pz8x9.pdf). For each of the 80 items, we calculated the proportion of acceptances among participants who saw that item. We then used a by-items regression to predict those proportions using content type (mathematical, causal), inference type (MP as reference, dummy variables for DA, AC and MT), perceived believability, perceived easiness, and the content-type × perceived believability and content-type × perceived easiness interaction terms. The resulting model is shown in Table 8. Unsurprisingly, inference type was a significant predictor of acceptance. Of the other predictors, only perceived believability was significant. This means that for mathematics undergraduates, semantic effects on inference acceptance are sensibly understood as rooted in believability rather than easiness.

Table 8

Regression Model Predicting Proportion of Acceptances; $R^2 = .886$.

Predictor	Estimate	SE	95% CI for Estimate	β	p
Intercept	0.903	0.035			< .001
DA	-0.665	0.040	[-0.910, -0.714]	-0.812	< .001
AC	-0.775	0.040	[-1.044, -0.849]	-0.946	< .001
MT	-0.146	0.040	[-0.276, -0.081]	-0.178	< .001
Content	0.035	0.041	[-0.066, -0.164]	0.049	.397
Believability	0.094	0.031	[0.065, 0.314]	0.189	.003
Easiness	0.062	0.047	[-0.059, 0.291]	0.116	.190
Content x Believability	-0.082	0.046	[-0.305, 0.017]	-0.144	.078
Content x Easiness	-0.053	0.054	[-0.290, 0.100]	-0.095	.336

Study 3b Discussion

Study 3b established that although perceived believability and perceived easiness are correlated – especially for mathematical conditionals – the effects in Studies 1b and 2 are better understood as driven by perceived believability. Thus, it appears that students at this level are able to look beyond experiences of ease or fluency when evaluating mathematical inferences.

General Discussion

Summary and Theoretical Account

This paper addressed three research questions:

1. Is conditional inference in mathematics affected by perceived believability?
2. Do mathematics undergraduates treat conditional inference similarly across mathematical, abstract and everyday content?
3. Are apparent believability effects in mathematical conditional inference better understood as arising from perceived believability or perceived easiness?

The answer to Question 1 is yes: conditional inference in mathematics is affected by perceived believability, though not uniformly. Studies 1b and 2 found overall believability effects, and Study 2 found individual differences: some students' responses were affected by perceived believability and others' responses were not.

The answer to Question 2 is yes, but with an important exception. Mathematics undergraduates evaluate conditional inferences more normatively overall than typical non-

specialist populations: they accept high proportions of MT as well as MP inferences, and reject high proportions of DA and AC inferences. This provides a conceptual replication of earlier studies (Attridge & Inglis, 2013; Inglis & Simpson, 2008, 2009). Also, mathematics undergraduates respond similarly across mathematical, abstract and everyday causal content, but evaluate inferences from mathematical conditionals with high perceived believability more normatively. Again, there are individual differences: some students respond almost normatively across all content; others are more influenced by content and perceived believability.

The answer to Question 3 is that perceived believability and perceived easiness are correlated, but the former rather than the latter predicts conditional inference acceptance. This means that mathematical believability is distinct from both easiness (Study 3) and truth (Study 1a), which raises a theoretical question: what exactly is mathematical believability?

As in the Theoretical Background, we suggest that mathematical believability can be explained by assuming the standard dual-strategy account of reasoning about conditionals (Evans, 2019; Oaksford & Chater, 2020) and hypothesizing a third strategy viable only in mathematics. Under the dual-strategy account, a probabilistic strategy gives a fast estimate of a conditional's likely truth; a counterexample strategy might refine this, if invoked. This accounts for believability judgements: people judge which of two conditionals is more believable based on global intuition or on how readily counterexamples come to mind. It also accounts for Study 3a's overall correlation between perceived believability and perceived easiness: presumably, the easier the content in a true conditional, the more readily reasoners see that there are no counterexamples.

The utility of the dual-strategy account, however, is limited to cases where truth is in doubt. For everyday reasoning, this is unproblematic because truth is always in doubt: even for conditionals that most people would view as sensible claims, it is possible to concoct exceptions. In mathematics, however, true conditionals admit no counterexamples, so the dual strategy account predicts that people with sufficient knowledge should rate all true conditionals equally believable. Including our hypothesized third strategy predicts something different: believability should vary not only with accessibility of counterexamples but also with accessibility of deductive arguments from a conditional's antecedent to its consequent.

We did not directly manipulate or measure argument accessibility. Nevertheless, this hypothesis is consistent with our data. Consider again these conditionals, with perceived believability and easiness *z*-scores from Study 3b:

[1] If n is a multiple of 4, then n^2 is a multiple of 4 (believability 1.91, easiness 0.81);

[2] If n is a multiple of 6, then n is a multiple of 3 (believability 1.05, easiness 2.63).

We would expect mathematics undergraduates to treat both of these as uncontroversially true. Yet our participants rated [1] considerably less easy than [2], but considerably more believable. Our hypothesis suggests that this occurs because the conditional 'if n is a multiple of 4, then n^2 is a multiple of 4' admits an accessible deductive argument ('squaring n squares the factors'). The conditional 'if n is a multiple of 6, then n is a multiple of 3' has this quality to a lesser extent: a proof requires at least informally introducing another variable ('n is 6 times something, so it is 2 times 3 times something').

Limitations and Open Questions

The sequence of studies reported here constitutes a first demonstration that believability affects mathematical conditional inference. But, as is natural for a first demonstration, it has limitations and leaves many open questions. We consider these under four themes: measurement, related constructs, generalisability, and culture.

Regarding measurement, we took a relativistic, bottom-up approach, using CJ to measure perceived believability. We think this likely to provide more meaningful group-level scores than, say, *a priori* researcher judgements. But there are different ways to operationalise group-level believability, including averaging direct ratings or counterexample counts. More importantly, we did not collect individual believability ratings, leaving open the question of how an individual student's assessments of believability uniquely influenced their conditional inference responses. This was a deliberate methodological choice: we thought that such ratings would be difficult to provide, and that the requirement to judge believability might influence subsequent inference evaluation. Nevertheless, a study of this nature could be attempted, perhaps with a time delay between the two tasks. In work using individual measures, we would expect believability effects to appear, if anything, more strongly than seen here; we note also that such measures could be used to study differences across populations (experts or younger students, for instance) or the stability of believability effects over time. Research might also investigate what does and does not change longitudinally: learning should shift individual

absolute believability ratings up for true conditionals and down for false conditionals, but this might have little effect on group-level *relative* perceived believability – some counterexamples and deductive arguments might remain more accessible than others.

Regarding related constructs, future studies could address believability and conditional inference in relation to fluency and confidence with content knowledge. They could measure fluency via performance and/or speed on standard mathematical tasks such as calculation or example generation. They could address potential individual differences in confidence thresholds: some students might accept a conditional inference if they feel 70% confident in its validity, where others might reject all inferences for which they are not 100% confident. Studies involving confidence could request Likert-scale confidence judgements or scaled ratings of certainty that a conclusion follows.

Regarding generalisation, our studies were each conducted with undergraduates at one of four UK universities, which permits efficient data collection but introduces challenges. We are fairly confident that our results would generalise to similar students, because participants in the conditional inference studies (1b and 2) had recently arrived from secondary schools around the UK and the world, and because their different universities had typical entry requirements. Also, the mathematical content they had encountered is broadly similar to that taught in other English-speaking countries, albeit that UK students specialise early so that they might be more comparable to more advanced undergraduates elsewhere. However, it is likely that they were more conscientious than average (as attendance was not enforced), and the clusters we identified could reflect their knowledge on arrival or their courses' more recent teaching of logic. Certainly, further empirical work would be needed to investigate generalisation more broadly, to younger students or to those not specialising in mathematics.

Generalisation is also central to questions of culture. Culture affects terminology: one of our conditionals used the term 'cyclic quadrilateral', where 'inscribed quadrilateral' might be more common in the USA. Consequently, we would not expect perceived believability scores to transfer between cultures – local calibration would be necessary. More importantly, there are substantive and interesting questions about the extent to which results on reasoning generalise across cultures (Yama, 2018). Different languages handle conditionals differently: Mandarin Chinese, for instance, has different connectives for necessary conditionals 'if A, then B' and sufficient conditionals, 'only if A, B' (Shao et al., 2022). This might promote more normative

conditional inference responses. On the other hand, there is evidence of a greater focus on rule-based reasoning in Western than in Eastern cultures (Peng & Nisbett, 1999). Although these results have not always replicated (e.g., Mercier et al., 2015), such differences might lead to cultural effects regarding the influence of semantic content, including believability. Either way, without further work, it would be unwise to assume that our findings generalise to non-Western, non-English speaking cultures.

Educational Implications

Our work does, however, have implications for instructors of undergraduate mathematics in comparable systems, and it raises questions for teachers and curriculum designers in earlier grades. Mathematics undergraduates' reasoning reflects the cumulative effects of their education. The finding that our participants responded more normatively for inferences from mathematical conditionals perceived as highly believable likely reflects the fact that the UK education system – which a majority of our participants had experienced – does not include work on formal logic. We suggest that these successful students had therefore learned to judge *deductive* validity: for conditionals perceived as highly believable, they could identify relevant counterexamples or deductive arguments and respond normatively. But most had not yet learned to judge *logical* validity: for conditionals is perceived as less believable, they were less able to identify relevant counterexamples or deductive arguments, and they did not have the option of relying on inference form.

This is not necessarily a problem – arguably, formal logic belongs in undergraduate mathematics, and the close-to-normative responses of a cluster in Study 2 suggest that some undergraduates, at least, can learn it relatively quickly. But we suggest that mathematics educators working in earlier grades, in the UK mathematics educators and elsewhere in similar systems, could consider developing students' reasoning via two broad strategies. First, to support normative reasoning by strengthening judgements of deductive validity, they could consider using tasks like those suggested by Watson and Mason (2005) and Komatsu (2016) to improve students' knowledge of examples, counterexamples and related deductive arguments. Second, to support normative reasoning by anticipating work on formal logic, they could follow or adapt the approach advocated by Hub and Dawkins (2018) and Dawkins and Norton (2022). That is, they could encourage students to compare arguments that they and their peers judge valid or invalid, and to notice regularity in their logical forms.

Acknowledgements

This work was supported in part by Leverhulme Trust Research Fellowship RF-2022-155 entitled ‘Does Mathematics Develop Logical Reasoning?’, in part by Research England via an Expanding Excellence in England grant to the Centre for Mathematical Cognition, and in part by Economic and Social Research Council grant ES/W002914/1.

The authors thank Cerian Brewer, Tom Wicks, Evgeny Ferapontov, Ian Jones, Dirk Schuetz and Wei En Tan for facilitating participant recruitment, and Rentuya Sa for completing data entry for Studies 1b, 2 and 3b. They also thank the reviewers of earlier versions of the manuscript for their substantive and useful comments.

References

Alcock, L. (2010). Mathematicians' perspectives on the teaching and learning of proof. In F. Hitt, D. Holton & P.W. Thompson (Eds.) *Research in Collegiate Mathematics Education VII* (pp. 63-91). MAA.

Alcock, L. (2013). *How to study for a mathematics degree*. Oxford University Press.

Alcock, L. & Attridge, N. (2023). Refutations and reasoning in undergraduate mathematics. *International Journal for Research in Undergraduate Mathematics Education.* https://doi.org/10.1007/s40753-023-00220-4

Alcock, L. & Inglis, M. (2008). Doctoral students' use of examples in evaluating and proving conjectures. *Educational Studies in Mathematics*, *69*, 111–129.

Alcock, L. & Sa, R. (2025). How do introduction-to-proof textbooks explain conditionals and implications? *International Journal of Research in Undergraduate Mathematics Education*. https://doi.org/10.1007/s40753-024-00263-1

Alcock, L. & Weber, K. (2005). Proof validation in real analysis: Inferring and checking warrants. *Journal of Mathematical Behavior*, *24*, 125–134.

Attridge, N., Doritou, M., & Inglis, M. (2015). The development of reasoning skills during compulsory 16 to 18 mathematics education. *Research in Mathematics Education*, *17*, 20–37.

Attridge, N. & Inglis, M. (2013). Advanced mathematical study and the development of conditional reasoning skills. *PLoS ONE*, 8:e69399.

Bell, A. W. (1976). A study of pupils' proof conceptions in mathematical situations. *Educational Studies in Mathematics*, *7*, 23–40.

Bradley, R. & Terry, M. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika, 39*, 324–345.

Braithwaite, D. W. (2025). Domain effects on interpretations of general conditionals: The case of mathematics. *Thinking & Reasoning*, *31*, 214–236.

Braithwaite, D. W. & Rafferty, A. N. (2025). Knowledge of examples affects conditional reasoning with mathematical content. *Thinking & Reasoning*. https://doi.org/10.1080/13546783.2025.2560995

Brousseau, G. (1997). *Theory of didactical situations in mathematics*. Kluwer.

Case, J. & Speer, N. (2021). Calculus students' deductive reasoning and strategies when working with abstract propositions and calculus theorems. *Problems, Resources, and Issues in Mathematics Undergraduate Studies*, *31*, 184–201.

Corfield, D. (2001). Bayesianism in mathematics. In D. Corfield & J. Williamson (Eds.), *Foundations of Bayesianism*, pp. 175–202. Springer-Science+Business Media, B.V.

Cummins, D. D. (1995). Naive theories and causal deduction. *Memory & Cognition*, *23*, 646–658.

Cummins, D. D., Lubart, T., Alksnis, O., & Rist, R. (1991). Conditional reasoning and causation. *Memory & Cognition*, *19*, 274–282.

Datsogianni, A., Sodian, B., Markovits, H., & Ufer, S. (2020). Reasoning with conditionals about everyday and mathematical concepts in primary school. *Frontiers in Psychology*, *11*, 531640.

Davies, B., Alcock, L., & Jones, I. (2021). What do mathematicians mean by proof? A comparative-judgement study of students' and mathematicians' views. *Journal of Mathematical Behavior*, *61*, 100824.

Dawkins, P. C. & Norton, A. (2022). Identifying mental actions for abstracting the logic of conditional statements. *Journal of Mathematical Behavior*, *66*, 100954.

de Grefte, J. (2023). Knowledge as justified true belief. *Erkenntnis*, *88*, 531–549.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2003). Inference suppression and semantic memory retrieval: Every counterexample counts. *Memory & Cognition*, *31*, 581–595.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2005). Working memory and everyday conditional reasoning: Retrieval and inhibition of stored counterexamples. *Thinking & Reasoning*, *11*, 349–381.

Department for Education (2021). National curriclum in England: Mathematics programmes of study. https://www.gov.uk/government/publications/national-curriculum-in-england-mathematics-programmes-of-study/national-curriculum-in-england-mathematics-programmes-of-study

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. (2019). Reflections on reflection: The nature and function of type 2 processes in dual-process theories of reasoning. *Thinking & Reasoning*, *25*, 383–415.

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.

Evans, J. St. B. T., Handley, S.J., Neilens, H., & Over, D. (2010). The influence of cognitive ability and instructional set on causal conditional inference. *Quarterly Journal of Experimental Psychology*, *63*, 892–909.

Fischbein, E. (1987). *Intuition in Science and Mathematics*. Reidel, Dordrecht.

Gowers, T. (2023). What makes mathematicians believe unproved mathematical statements? *Annals of Mathematics and Philosophy*, *1*, 57–110.

Hamami, Y., Mumma, J., & Amalric, M. (2021). Counterexample search in diagram-based geometric reasoning. *Cognitive Science*, *45*, e12959.

Harel, G. & Sowder, L. (1998). Students' proof schemes: Results from exploratory studies. In A.H. Schoenfeld, J. Kaput, & E. Dubinsky (Eds.), *Research in Collegiate Mathematics III*, pp.234–282. American Mathematical Society.

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.

Hub, A. & Dawkins, P. C. (2018). On the construction of set-based meanings for the truth of mathematical conditionals. *Journal of Mathematical Behavior*, *50*, 90–102.

Ichikawa, J. & Steup, M. (2024) The analysis of knowledge. In E.N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy (Fall 2024 Edition)*. https://plato.stanford.edu/archives/fall2024/entries/knowledge-analysis/

Inglis, M. (2006). *Dual processes in mathematics: Reasoning about conditionals.* (Doctoral dissertation, University of Warwick). https://wrap.warwick.ac.uk/2847/

Inglis, M. & Attridge, N. (2016). *Does mathematical study develop logical thinking? Testing the theory of formal discipline*. London: World Scientific.

Inglis, M., Mejía-Ramos, J. P., & Simpson, A. (2007). Modelling mathematical argumentation: The importance of qualification. *Educational Studies in Mathematics*, *66*, 3–21.

Inglis, M. & Simpson, A. (2008). Conditional inference and advanced mathematical study. *Educational Studies in Mathematics*, *67*, 187–204.

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

Jones, I., Bisson, M.-J., Gilmore, C., & Inglis, M. (2019). Measuring conceptual under- standing in randomised controlled trials: Can comparative judgement help? *British Educational Research Journal*, *45*, 662–680.

Jones, I. & Davies, B. (2024). Comparative judgement in education research. *International Journal of Research & Method in Education*, *47*, 170-181.

Jones, I. & Inglis, M. (2015). The problem of assessing problem solving: Can comparative judgement help? *Educational Studies in Mathematics*, *89*, 337–355.

Komatsu, K. (2016). A framework for proofs and refutations in school mathematics: Increasing content by deductive guessing. *Educational Studies in Mathematics*, *92*, 147–162.

Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. *Japanese Psychological Review, 62,* 221-230

Lampert, M. (2001). *Teaching problems and the problems of teaching*. Yale University Press.

Leron, U. & Hazzan, O. (2006). The rationality debate: Application of cognitive psychology to mathematics education. *Educational Studies in Mathematics*, *62*, 105–126.

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

Markovits, H., Brunet, M.-L., Thompson, V., & Brisson, J. (2013). Direct evidence for a dual process model of deductive inference. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *39*, 1213–1222.

Mason, J., Burton, L., & Stacey, K. (1982). *Thinking Mathematically*. Addison-Wesley.

Mercier, H., Zhang, J., Qu, Y., & Lu, P., & Van der Henst, J-B. (2015). Do Easterners and Westerners treat contradiction differently? *Journal of Cognition and Culture, 15*, 45–63.

National Governors Association Center for Best Practices, Council of Chief State School Officers (2010). *Common core state standards for mathematics*. National Governors Association Center for Best Practices, Council of Chief State School Officers

Nickerson, R. S. (2015). *Conditional reasoning: The unruly syntactics, semantics, thematics, and pragmatics of "if"*. Oxford University Press.

Norton, A., Antonides, J., Arnold, R., & Kokushkin, V. (2025). Logical implications as mathematical objects: Characterizing epistemological obstacles experienced in introductory proofs courses. *Journal of Mathematical Behavior*, *79*, 101253.

Oaksford, M. & Chater, N. (2020). New paradigms in the psychology of reasoning. *Annual Review of Psychology*, *71*, 305–330.

Oaksford, M., Chater, N., & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *26*, 883–899.

Over, D. E., & Evans, J. St. B. T. (2024). *Human reasoning*. Cambridge, UK: Cambridge University Press.

Over, D.E., Hadjichristidis, C., Evans, J. St. B. T., Handley, S.J., & Sloman, A.A. (2007). The probability of causal conditionals. *Cognitive Psychology*, *54*, 62–97.

Peng, K. & Nisbett, R. E. (1999). Culture, dialectics, and reasoning about contradiction. *American Psychologist, 54*, 741–754.

Poincaré, H. (1969). Classics in mathematics education: Intuition and logic in mathematics. *The Mathematics Teacher*, *2*, 205–212.

Pólya, G. (2014). *Mathematics and plausible reasoning*. Martino Fine Books.

Schoenfeld, A. H. (1985). *Mathematical problem solving*. Academic Press.

Schwitzgebel, E. (2024). Belief. In E.N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy (Spring 2024 Edition)*. https://plato.stanford.edu/archives/spr2024/entries/belief/.

Selden, A. & Selden, J. (2003). Validations of proofs considered as texts: can undergraduates tell whether an argument proves a theorem? *Journal for Research in Mathematics Education*, *34*, 4–36.

Shao, J., Tikiri Banda, D., & Baratgin, J. (2022). A study on the sufficient conditional and the necessary conditional with chinese and french participants. *Frontiers in Psychology*, *13*, 757588.

Sinclair, N., Watson, A., Zazkis, R., & Mason, J. (2011). The structuring of personal example spaces. *Journal of Mathematical Behavior*, *30*, 291–303.

Smith, A. (2004). *Making Mathematics Count: The report of Professor Adrian Smith's Inquiry into Post-14 Mathematics Education*. London: The Stationery Office.

Stylianides, A. J. (2007). Proof and proving in school mathematics. *Journal for Research in Mathematics Education*, *38*, 289–321.

Stylianides, A. J., Stylianides, G. J., & Philippou, G. N. (2004). Undergraduate students' understanding of the contraposition equivalence rule in symbolic and verbal contexts. *Educational Studies in Mathematics*, *55*, 133–162.

Tall, D.O. & Vinner, S. (1981). Concept image and concept definition in mathematics with particular reference to limits and continuity. *Educational Studies in Mathematics*, *12*, 151–169.

Vamvakoussi, X., Van Dooren, W., & Verschaffel, L. (2012). Naturally biased? In search for reaction time evidence for a natural number bias in adults. *Journal of Mathematical Behavior, 31*, 344–355.

Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice*, *26*, 541–562.

Ward Jr., J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*, 236-244.

Watson, A. & Mason, J. (2005). *Mathematics as a constructive activity: Learners generating examples*. Lawrence Erlbaum Associates.

Yackel, E. & Cobb, P. (1996). Sociomathematical norms, argumentation, and autonomy in mathematics. *Journal for Research in Mathematics Education*, *27*, 458–477.

Yackel, E., Rasmussen, C., & King, K. (2000). Social and sociomathematical norms in an advanced undergraduate mathematics course. *Journal of Mathematical Behavior*, *19*, 275–287.

Yama, H. (2018). Thinking and reasoning across cultures. In L. J. Ball & V. A. Thompson (Eds.), *The Routledge international handbook of thinking and reasoning* (pp. 624–638). Routledge/Taylor & Francis Group.

Zazkis, R. & Chernoff, E. J. (2008). What makes a counterexample exemplary? *Educational Studies in Mathematics*, *68*, 195–208.