Believability in Mathematical Conditionals: Generating Items for a Conditional Inference Task

Lara Alcock
Loughborough University

Ben Davies
University of Southampton

This paper describes design issues for a conditional inference task with mathematical content. The task will mirror those used in cognitive psychology to study inferences from everyday causal conditionals: its items will present a conditional premise (if A then B) and a categorical premise (A, not-A, B, or not-B) and ask participants to evaluate whether a conclusion (respectively, B, not-B, A, not-A) necessarily follows. To assemble items, we asked six mathematics education researchers with expertise in conceptual understanding to generate conditionals covering a range of mathematical topics. To mirror the structure of tasks with everyday causal content, we asked that these conditionals should vary in believability. In this paper, we analyze the content and phrasing of the submitted conditionals in order to assess their suitability for use in a conditional inference task, and describe our planned use of this task to investigate the relationship between logical reasoning and mathematical expertise.

Keywords: conditional, inference, logic, reasoning, proof

Introduction

Logical reasoning is central to mathematics, overtly so at the transition to proof (David & Zazkis, 2019). Students must learn to validate inferences that they make during proof construction and inferences that they read during proof comprehension (Hodds, Alcock & Inglis, 2014). This does not mean that every inference must be explicitly justified – demands vary by context (Weber, 2008). It does mean that every inference could in principle be checked, in a process that in experts involves considerable back-and-forth reading (Inglis & Alcock, 2012).

Inference checking can fail for semantic reasons. Alcock and Weber (2005), for instance, reported that some students were willing to accept an inference from the premise that $(\sqrt{n})$ is an increasing sequence to the conclusion that $(\sqrt{n})$ tends to infinity. This inference is deductively invalid (Smith, 2020) because some increasing sequences do not tend to infinity. The argument, however, has the form below. If the conditional premise (interpreted as a generalized conditional per Durand-Guerrier, 2003) were true, then the inference would be valid.

If (a_n) is an increasing sequence, then (a_n) tends to infinity.
$(\sqrt{n})$ is an increasing sequence.
So $(\sqrt{n})$ tends to infinity.

Inference checking can also fail for syntactic reasons. Selden and Selden (2003), for instance, reported that some students were willing to accept a proof of the first claim below as a proof of the second (see also Hoyles & Kuchemann, 2002; Inglis & Alcock, 2012). This is logically invalid (Smith, 2020) because a conditional and its converse are not equivalent.

For any positive integer n, if n is a multiple of 3 then n^2 is a multiple of 3.
For any positive integer n, if n^2 is a multiple of 3 then n is a multiple of 3.

In the semantic case, failure to detect an invalid inference is readily explained by the fact that individuals' example spaces (Sinclair et al., 2011) or concept images (Tall & Vinner, 1981) do not fully 'match' the defined concepts, so that people might overlook counterexamples. In the syntactic case, failure to detect an invalid inference might occur for what amounts to the same reason at a more abstract level: individuals' example spaces or concept images *for conditionals* are unlikely to 'match' the material or truth-functional interpretation used in mathematics (Dawkins & Norton, 2022; Epp, 2003). This occurs because everyday conditionals lend themselves to multiple distinct interpretations: for 'If it is a dog, then it is an animal', a material conditional interpretation is sensible; for 'If you mow the lawn, I will give you $5', a biconditional interpretation is sensible (Cummins et al., 1991). Hence, mathematics students must learn to restrict their interpretations of conditionals just as they must learn to restrict their interpretations of words like 'limit' and 'group'. As they learn, we can expect to see errors in relation to normative validity. Indeed, we might expect errors to persist – maybe mathematical experts reason normatively across all contexts, but maybe they reason normatively in mathematics only, and maybe their reasoning is vulnerable to errors even there.

To investigate these possibilities and thereby to improve knowledge relevant to the teaching of logical reasoning, we plan to build on the extensive research on everyday conditional reasoning in cognitive psychology. Specifically, we plan to construct a mathematical *conditional inference* task and use this, together with a task with everyday causal content, to study whether and how reasoning differs across contents and levels of mathematical expertise. In the present paper, we explain the theoretical background of our work and report first steps in task design.

Theoretical Background

Cognitive psychology has a long history of studying reasoning with and about conditionals (e.g., Evans & Over, 2004; Oaksford & Chater, 2020). Here we focus on conditional inference tasks, as we believe their structure is clearly relevant to proof validation. A typical conditional inference task presents participants with a conditional premise and one of four categorical premises, as illustrated below with everyday causal content (De Neys et al., 2003); participants are asked to evaluate the conclusion in the third line. Normatively, MP and MT inferences are valid, and DA and AC inferences are invalid.

Modus Ponens (MP)
If John studies hard, then he does well on the test.
John studies hard.
John does well on the test.

Affirmation of the Consequent (AC)
If John studies hard, then he does well on the test.
John does well on the test.
John studied hard.

Denial of the Antecedent (DA)
If John studies hard, then he does well on the test.
John does not study hard.
John does not do well on the test.

Modus Tollens (MT)
If John studies hard, then he does well on the test.
John does not do well on the test.
John did not study hard.

The task instructions might emphasise logic, asking participants to assume that the premises are true and state whether the conclusion necessarily follows, or they might seek to elicit everyday reasoning, asking participants to assume that the premises hold and use a scale to express certainty that the conclusion can be drawn (e.g., Evans, Handley, Neilens & Over, 2010). Instructions that emphasise logic mirror the task of proof validation: participants are asked to assume that some premises are true and decide whether a conclusion necessarily follows, which involves recognizing normatively invalid DA and AC inferences.

However, everyday content introduces complexity, because a conditional like 'If John studies hard, then he does well on the test' is not simply true or false. It is a reasonable assertion about a causal relationship, but studying hard is neither necessary nor sufficient for doing well. *Disablers* (Byrne, 1989) might intervene to prevent the antecedent from causing the consequent: John might study hard but not do well if, for example, the test is very difficult or if he studies the wrong subject (De Neys et al., 2003). *Alternative antecedents* (Cummins et al., 1991) might account for the consequent without the stated antecedent: John might study little but still do well if the test is easy or if he is lucky (De Neys et al., 2003). For causal conditionals, disablers highlight the fact that the antecedent might not be sufficient for the consequent and are known to make people less likely to endorse MP and MT inferences; alternative antecedents highlight the fact that the stated antecedent might not be necessary for the consequent and make people less likely to endorse DA and AC inferences (Byrne, 1989; Cummins et al., 1991). Disablers and alternative antecedents can be more or less accessible – for instance, people typically generate few disablers but many alternatives for 'If Alvin read without his glasses, then he got a headache' (Cummins, 1995). Thus, conditionals are more or less believable and support the four inferences to different extents (De Neys et al., 2003). Also, there are individual differences in ability to prioritise logic over believability, though rates do not approach normative perfection. Evans, Handley, Neilens and Over (2010) reported that for participants of higher cognitive ability, acceptance rates under deductive instructions were 93% for MP inferences, 41% for DA, 41% for AC and 47% for MT; these are far from the normative 100%, 0%, 0% and 100%.

In order to study conditional inference across levels of mathematical expertise, we aim to design a mathematical conditional inference task with structure that parallels tasks with everyday content, using all four inference types and conditionals that vary in believability. However, this raises several issues in task design.

One issue is that to parallel everyday tasks, we would like to use simple phrasing as in 'If x is less than 2, then x is less than 5'. However, in mathematics, such a conditional is considered a predicate, not a proposition, so that formally it has no truth value (Durand-Guerrier, 2003). We consider this relatively unproblematic because sensible universal quantification is commonly assumed (Dawkins & Roh, 2022) and because tasks with everyday content have this feature too (Oaksford & Chater, 2007): participants are implicitly invited to consider 'If John studies hard, then he does well on the test' as a generalized conditional applying across multiple situations.

Another issue is that in mathematical proofs, universal instantiation is routine. In the Introduction, the conditional premise 'If (a_n) is an increasing sequence, then (a_n) tends to infinity' appears not with the general categorical premise '(a_n) is an increasing sequence' but with the instantiation '$(\sqrt{n})$ is an increasing sequence'. Premises of both types can appear in proofs, and instantiations have been used in the few studies to investigate conditional inference in mathematics (Case & Speer, 2021; Durand-Guerrier, 2003). However, premises of the general type are used in typical research on everyday causal tasks so, while noting that this would be an interesting variation for the future, we prioritise comparability across contexts.

A third issue is that with assumed universal quantification and no instantiation, mathematical conditionals are either true or false. For any true conditional (that is not a true biconditional), the antecedent is sufficient but not necessary for the consequent. Sufficiency means that and MP and MT inferences are always valid because standard logic is monotonic (Smith, 2020) so there can be no disablers – nothing could 'intervene' to prevent an x less than 2 from being less than 5. Lack of necessity means that DA and AC inferences are always invalid, though counterexamples differ from the alternative antecedents of everyday tasks because they are often singular ('zero') or in classes of a single type ('all negative numbers').

For these reasons, it is impractical in mathematics to follow the typical research approach (e.g., De Neys et al., 2003) of distinguishing more and less believable conditionals using pre-tests in which participants list possible disablers and alternatives. Nevertheless, we suggest that believability for mathematical conditionals does plausibly vary on a continuous scale. For instance, 'If X is a square then X is a parallelogram' is true but perhaps not immediately or wholly believable. While a square is a parallelogram, it would be conversationally uncooperative (Grice, 1989) to call X a parallelogram if one knew that it was a square. In contrast, 'If $x < 3$ then $1/x > 1/3$' is false but definitely not immediately or wholly unbelievable. Negative x values might easily be overlooked – Alcock and Attridge (2023) reported that about one fifth of mathematics undergraduates initially judged this conditional true. We might, in theory, assess believability by asking participants to score it directly (cf. Evans et al., 2010). However, we think that mathematically experienced people would likely balk at assigning a believability score that is not 1 or 0 when they know that a conditional must be true or false.

We therefore take an alternative approach, asking not about disablers/alternatives or believability scores, but instead about *relative* believability. In the work reported here, we asked experts in conceptual understanding to generate five mathematical conditionals each and to rank their conditionals for believability. If believability does vary continuously, this should ensure that we gather conditionals from across the scale. In work to follow, we will use comparative judgement (see, e.g., Davies et al., 2021) with both experts and undergraduates to assess more robustly whether believability is a reliably shared construct.

Method

Participants, Data Collection and Data Processing

We set out to collect conditionals covering a range of mathematical topics, to assess whether believability might vary as anticipated by the theory above, and to refine our understanding of how language around mathematics reflects or differs from that around everyday causal content. We therefore approached six mathematics education researchers with expertise in secondary-school-level conceptual understanding, who would be able to anticipate what undergraduate students would find more and less believable. These participants were informed about the purpose and theoretical background of the study and were each asked to generate five mathematical conditionals that would be familiar to typical students aged 13-14 and therefore basic for mathematics undergraduates, postgraduates, and experts. They were instructed that to parallel those used in tasks with everyday causal content, their conditionals should:

- Cover a range of mathematical 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 (where the most believable could be clearly true).

They were also asked to:

- State whether each of their conditionals was technically true or false;
- Rank their conditionals from 1 (least believable) to 5 (most believable);
- Give a one-sentence explanation of their ranking for each conditional.

Prior to analysis, we requested amendments where participants had included conditionals with extra connectives (often in symbols, for instance '$\leq$' or '$\neq$'). We checked the stated truth values, inviting participants to explain/adjust if there was any doubt about, for example, the set over which universal quantification was assumed. We then ensured that all conditionals were written in the same way, with an explicit 'then' after a comma, and without extra words. For example, '…then it is also convex' became 'then it is convex', and '…then the median must be 7' became '…then the median is 7'. Overall, participants were able to comply with our instructions and generate conditionals with non-obvious truth values, though some observed that ranking was difficult. Naturally, we do not take individual rankings as inherently reliable; the planned comparative judgement will provide more robust measures.

Analysis: Removing Unsuitable Conditionals

Our analysis focused on suitability for use in a conditional inference task. Seven of the 30 submitted conditionals were not suitable because, although they met our criteria, their consequents could not readily be asserted as categorical premises, even with rephrasing (equivalently, the converses could not readily be stated). Not coincidentally, some were long and some invoked an agent 'you' or 'I' or described physical actions as well as abstract relationships. These conditionals, listed below, were therefore removed.

- If I flip a fair coin twice, then the probability of getting two heads is 1/3.
- If you sum the first *n* odd numbers, then you get n^2.
- If a circle is divided into regions by straight lines connecting *n* dots, then the maximum number of regions is 2^n.
- If coin lands on tails 42 times out of 100 flips, then there is not enough information to tell whether it is biased.
- If £12 is shared in the ratio 1:3, then the smaller share is 1/3 of the total.
- If two normal 6-sided dice are thrown and I tell you that one of the dice shows a 2, then the probability of both being a 2 is 1/6.
- If the perimeter of a square is the same as the circumference of a circle, then the area of the square is less than the area of the circle.

Along similar lines, we also removed the conditional 'If $a = -1$, then $|a| = -|1|$' because its consequent is false. We then removed a further four conditionals were not suitable because, although they met our criteria and were phrased in ways better fitted a conditional inference task, they had true converses. This would potentially create a confound because it would mean that DA and AC inferences would be valid for semantic reasons; it also makes these conditionals less 'like' everyday causal conditionals in which there is a clear cause in the antecedent and effect in the consequent. The following conditionals were therefore also removed.

- If $x^3 < x^2$ then $x < 0$.
- If $x = 24$ is the solution to $4x + 9 = 105$, then $x = 12$ is the solution to $8x + 9 = 105$.
- If the diagonals of a quadrilateral bisect each other, then the quadrilateral is a rectangle.
- If *n* is a positive integer, then it can be written as a product of prime factors in exactly one way.

Analysis: Content and Form in Suitable Conditionals

This left 18 conditionals potentially suitable for a conditional inference task. We list these in Table 1, ordered by participants' individual believability rankings (most to least believable) and together with their truth values and mathematical topics. The final column captures the way in which scope of quantification is handled, with codes (discussed further below):

- 'I' where the scope is implicit, usually where a variable is assumed to be a real number;
- 'E' where the scope is explicit, usually where a set is specified in the antecedent;
- 'S' where the antecedent involves a single specific number or situation.

Table 1. Conditionals potentially suitable for a conditional inference task

Conditional	Rank	Truth	Topic	Scope
If a quadrilateral is cyclic, then it is convex.	5	T	geometry	E
If a polygon is a square, then it is a rhombus.	5	T	geometry	E
If $x^2 = y^2$, then $xy = yx$.	5	T	algebra	I
If $x = -4$, then $x^2 + x - 12 = 0$.	5	T	algebra	S
If a quadrilateral has a reflex angle, then it will tesselate.	4	T	geometry	E
If x is positive, then $\tan x > \sin x$.	4	F	trig	E
If $x - 12{,}345 = .67$ then $x > -12{,}345.67$.	4	T	number	S
If a fraction has denominator 7, then it is equivalent to a non-terminating decimal.	4	F	number	E
If $\sin x > 0$ then $\cos x < 1$.	3	T	trig	I
If a number is a multiple of 13, then it has an even number of factors.	3	F	number	E
If a composite number ends in a 3, then it is a multiple of 3.	3	F	number	E
If a rectangle has area 10cm^2, then its perimeter is greater than 10cm.	2	T	geometry	E
If an equation is a quadratic, then it has exactly two roots.	2	F	algebra	E
If $a = 42$, then $a \times b > 42$.	2	F	number	S
If four consecutive numbers are added, then the result is a multiple of four.	2	F	number	E
If the mean of a dataset is 7, then the median is 7.	1	F	statistics	E
If $a > b$, then $ac > bc$.	1	F	number	I
If a rectangle is stretched so that its side lengths double, then its area doubles.	1	F	ratio	E

In terms of topics, truth values, and believability rankings, we find this initial collection satisfactory. Topics are spread across believability rankings. More believable conditionals are more likely to be true but, in line with our suggestion that truth and believability might not align perfectly, there is considerable overlap in the middle of the table. Our removal of unsuitable conditionals did not disproportionately remove those at specific ranks, so there is no evidence of a problematic interaction between believability and phrasing challenges or truth of the converse.

The scope codes highlight new issues in designing a task to parallel those involving everyday causal content, in terms of substance, phrasing and the interactions between the two.

Conditionals with scope code I are most straightforwardly like those used in everyday causal tasks. Their 'making sense' without explicit scope means that they lend themselves to presentation with all four corresponding categorical premises. For instance, for the conditional 'If $x^2 = y^2$ then $xy = yx$', an AC item would appear as follows.

If $x^2 = y^2$ then $xy = yx$.
$xy = yx$.
$x^2 = y^2$.

Conditionals with scope code E usually have a specified set in the antecedent and a deictic 'it' or 'its' in the consequent, as in 'If a quadrilateral is cyclic, then it is convex'. In these cases, categorical premises cannot simply use the consequent as is ('It is convex' makes no sense in isolation). We do not want to lose these items, which make up the majority of those submitted and which are clearly natural in mathematics. As a sensible solution, we will likely mirror the phrasing typically used for everyday causal items, naming the object and using the name in both categorical premise and conclusion, as below. Mathematical conditionals are at least simpler in that they do not require manipulating tenses to accommodate the temporal aspects of causality.

If a quadrilateral Q is cyclic, then it is convex.	If John studies hard, then he does well on the test.
Quadrilateral Q is convex.	John does well on the test.
Q is cyclic.	John studied hard.

Conditionals with scope code S – whose antecedents focus on specific objects – have the interesting property that they do not 'feel' like generalized conditionals. We could avoid using such items, but we note that they have interesting properties around phrasing, necessity and sufficiency. For instance, the antecedent of the conditional 'If $a = 42$, then $a \times b > 42$' seems to be about a, when the conditional actually quantifies over b. It is false but invokes the common misconception 'multiplication makes things bigger', and the antecedent is so far from necessary for the consequent that accepting the AC inference from $a \times b > 42$ to $a = 42$ would seem nonsensical. This raises questions about how plausible false conditionals must be in order to work in our task, and we will discuss this further at the conference.

Discussion and Next Steps

The analysis above clarifies design issues for a mathematical conditional inference task. Despite the normative interpretation of conditionals in mathematics, 'natural' uses of if-then to express mathematical relationships vary considerably. Some uses do not lend themselves to a conditional inference task due to length and overall complexity in describing agents and/or actions. Some have true converses, and therefore do not match items used for causal conditionals, where the cause appears in the antecedent and the effect in the consequent (Cummins et al., 1991). For those that are in principle suitable, scope might be fully implicit – commonly in the case of quantification over sets of numbers – or explicit in the sense that a set is specified in the antecedent. In the latter case, rephrasing is necessary to construct an item.

We will use this analysis to inform our ongoing work. We aim to collect a total of approximately 40 suitable conditionals, with a view to constructing a 16-item task using eight conditionals that are more believable and eight that are less believable (should believability turn out not to be a reliable construct, we will simply use a spread of topics). We will therefore ask two more experts to generate five conditionals each, with the additional criterion that the converse of each should be readily articulated and false. We will then collect conditionals from the mathematics education literature and textbooks to complete a set that balances topics, truth and implicit/explicit quantification, and will seek to make phrasing as uniform as is practical. We will report on this work at the conference, together with the comparative judgement study of believability and a first test of the conditional inference task with undergraduates.

Acknowledgments

This work is funded by Leverhulme Trust Research Fellowship RF-2022-155 titled "Does Mathematics Develop Logical Reasoning?".

References

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

Byrne, R.M.J. (1989). Suppressing valid inferences with conditionals. *Cognition*, *31*, 61–83.

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.

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

David, E. J. & Zazkis, D. (2019). Characterizing introduction to proof courses: A survey of U.S. R1 and R2 course syllabi. *International Journal of Mathematical Education in Science and Technology*, *51*, 388–404.

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

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.

Epp, S. (2003). The role of logic in teaching proof. *American Mathematical Monthly*, *110*, 886-899.

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.

Evans, J.St.B.T. & Over, D.E. (2004). *If*. Oxford: Oxford University Press.

Grice, H. P. (1989). *Studies in the way of words*. Boston, MA: Harvard University Press.

Hodds, M., Alcock, L., & Inglis, M. (2014). Self-explanation training improves proof comprehension. *Journal for Research in Mathematics Education*, *45*, 62-101.

Hoyles, C. & Küchemann, D. (2002). Students' understanding of logical implication. *Educational Studies in Mathematics*, *51*, 193–223.

Inglis, M. & Alcock, L. (2012). Expert and novice approaches to reading mathematical proofs. *Journal for Research in Mathematics Education*, *43*, 358-390.

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

Oaksford, M. & Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford: Oxford University Press.

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.

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

Smith, P. (2020). *An introduction to formal logic: Second Edition*. Cambridge: Cambridge University Press.

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.

Weber, K. (2008). How mathematicians determine if an argument is a valid proof. *Journal for Research in Mathematics Education*, *39*, 431–459.