International Journal of Research in Undergraduate Mathematics Education
https://doi.org/10.1007/s40753-023-00220-4

Refutations and Reasoning in Undergraduate Mathematics

Lara Alcock[1] · Nina Attridge[2]

Accepted: 6 May 2023

Abstract
This paper concerns undergraduate mathematics students' understandings of refutation and their related performance in abstract conditional inference. It reports on 173 responses to a refutation instrument that asked participants to: 1) state 'true' or 'false' for three statements, providing counterexamples or reasons if they thought these false (all three were false); 2) evaluate possible counterexamples and reasons, where reasons were 'corrected' versions of the statements but not valid refutations; and 3) choose which of the counterexamples and the corrected statements were better answers, explaining why. The data show that students reliably understood the logic of counterexamples but did not respond normatively according to the broader logic of refutations. Many endorsed the corrected statements as valid and chose these as better responses; we analyse their explanations using Toulmin's model of argumentation. The data further show that participants with better abstract conditional inference scores were more likely to respond normatively by giving, endorsing, and choosing counterexamples as refutations; conditional inference scores also predicted performance in a proof-based course.

Keywords Argument · Counterexample · Conditional inference · Reasoning · Logic · Refutation

Introduction

Refutation is an important part of mathematical reasoning. In mathematics education, it is often discussed in terms drawn from Lakatos's seminal *Proofs and Refutations* (Lakatos, 1976). Lakatos characterises mathematical development as a process in which conjectures, proofs and refutations inform one another in multiple ways: mathematicians find global counterexamples to conjectures and clarify definitions to

✉ Lara Alcock
L.J.Alcock@lboro.ac.uk

[1] Department of Mathematics Education, Loughborough University, Loughborough, Leicestershire LE11 3TU, UK

[2] Department of Psychology, University of Portsmouth, Portsmouth, Hampshire, UK

Published online: 10 July 2023

refine concepts; they analyse proofs to identify implicit lemmas, identify local counterexamples to suspect lemmas, and improve conjectures by incorporating unfalsified lemmas as conditions; they use insights so developed to extend concepts, bringing counterexamples into the domains of deeper, deductively generated theorems.

Research in mathematics education that builds on Lakatos (1976) has developed in several directions. Some researchers have studied characteristics of counterexamples that support conceptual change by convincingly refuting naïve conceptions (Balacheff, 1991; Peled & Zaslavsky, 1997; Zazkis & Chernoff, 2008). Some have designed task structures or teacher input to encourage students to generate counterexamples and modify conjectures and/or proofs (Buchbinder & Zaslavsky, 2011; Koichu, 2008; Komatsu, 2016, 2017; Lin, 2005; Reid, 2002; Stylianides & Ball, 2008; Yang, 2012; Yopp, 2013). Some have sought to engage students in the full range of authentic mathematical activity as characterised by Lakatos, designing collaborative inquiry-based learning in which students identify and use local and global counterexamples to clarify conceptual meanings and to test and revise conjectures and proofs (Komatsu & Jones, 2022; Larsen & Zandieh, 2008; Stylianides & Stylianides, 2009; Yim et al., 2008).

In most cases, the focus of this work has been epistemological: researchers want to improve mathematical knowledge, and they seek to elicit and address refutations in service of that goal. Where the focus has been logical, it has usually been on helping students to understand the status of counterexamples in relation to general statements (Peled & Zaslavsky, 1997; Stylianides & Stylianides, 2009; Yopp et al., 2020; Zazkis & Chernoff, 2008) or on the importance of checking for counterexamples to steps in deductive arguments (Alcock & Weber, 2005; De Villiers, 2004; Ko & Knuth, 2013; Weber, 2010). It has less commonly addressed the logic of proposed refutations that are *not* counterexamples, although students and teachers have been observed to produce such alternatives, typically of two types. First, they might claim that a statement is false because no known theorem proves it. This has been reported in cases involving geometry (Potari et al., 2009) and inequalities and absolute values in calculus (Giannakoulias et al., 2010). Second, they might offer amended conjectures, in effect 'correcting' original, false conjectures. Creager (2022), for instance, reported on interviews in which pre-service teachers refuted geometry conjectures; of 32 arguments brought forth, 12 involved providing a conjecture with an alternative conclusion. Lin (2005) reported on a survey in which over 2000 7th and 8th graders considered the conjecture 'a quadrilateral, in which one pair of opposite angles are right angles, is a rectangle'; 32% of 7th graders and 16% of 8th graders disagreed with the conjecture and explained by providing an alternative conclusion.

We believe that this second phenomenon merits deeper investigation, and were initially brought to this view by responses to test items in the first author's real analysis course. Students were asked to examine various statements, to state 'true' or 'false' for each one and, if they stated 'false', to give a counterexample or a brief reason. Three items from the course appear below, with students' purported refutations in the form of corrected statements.

- For all $y \in \mathbb{R}$, the sequence $(y^n) \to \infty$.
  FALSE. Reason: It should be 'For all $y > 1$, the sequence $(y^n) \to \infty$'.

- Every sequence has an increasing subsequence.
  FALSE. Reason: It should be 'Every sequence has a <u>monotonic</u> subsequence'.
- $\sum_{n=1}^{\infty} a_n$ converges if and only if $\left(a_n\right) \to 0$.
  FALSE. Reason: It should be ' $\sum_{n=1}^{\infty} a_n$ converges if and only if its sequence of partial sums converges'.

The statements are indeed all false, and all three reasons are true theorems or correct definitions. They thus have epistemological value: the first two might result from successful heuristic refutation in Lakatos's (1976) sense, and they all show that students have paid attention in the course. However, as refutations per se, they are logically inadequate because none rules out the possibility that the original statement is also true. In the analysis course, such responses were not occasional idiosyncratic curiosities – they were common across these items and others of similar structures.

We believe that this phenomenon deserves research attention because it is not obvious to what extent students providing corrected statements believe them to be logically adequate refutations. The above responses are spontaneous and short, so they give little information on the underlying reasoning. They might come to mind for students who then think no further. They might arise when students fail to find counterexamples and consequently give statements that they at least know are correct. They might arise from successful heuristic refutation in which students identify counterexamples, use these to inform corrections to the statements, and write down the correct versions as outcomes of that process. In the last two cases, we do not know to what extent students engage with the relevant logic. Maybe they do not think about it at all. Maybe they think about it and believe that their answers do logically contradict the original statements. Maybe they think about it and believe that their answers contradict the originals via looser conversational norms (Grice, 1989), assuming that a reader would infer that by correcting to '$y > 1$' they intend to exclude counterexamples among the cases where $y \leq 1$.

This is important because, whether we intend to enforce logical precision or to allow more 'natural' communication, we do want undergraduate mathematics students to know that mathematicians notice cases in which an alternative claim does not logically refute a conjecture and that they might value corrected statements without deeming them valid refutations. For educational purposes, it would therefore be useful to know how accessible this view is based on students' current understandings. This information is not inferable from spontaneous student responses as cited in the above research and anecdotal observations, because these provide no information on understandings of the *relative* value of *different* potential refutations: if a student provides a corrected statement, we learn nothing about whether they would also endorse a counterexample or vice versa, and nothing about which they think is the better refutation and why. We also learn nothing about whether their responses are related to their broader logical reasoning skills and mathematical performance.

We therefore designed a novel refutation instrument to investigate whether the logical issue was relatively easy to address – perhaps students would immediately realise that counterexamples were preferable – or whether corrected statements were endorsed as valid refutations, even in the face of alternatives. We used this, along with a standard abstract conditional inference task as a measure of logical reasoning, to address the following research questions.

1.
    (a) How prevalent are counterexamples and corrected statements when students attempt to refute a statement?
    (b) When presented with both counterexamples and corrected statements, what proportions of students evaluate these as valid?
    (c) When asked to choose between counterexamples and corrected statements, what proportions of students deem each better?

2. How do students explain their reasons for deeming counterexamples or corrected statements better?
3. Are student responses to refutation tasks systematically related to their performance in abstract conditional inference?
4. Do refutation responses and abstract conditional inference performance predict course grades?

Below, we expand on the relevant research and set up a frame for analysis using Toulmin's (1958) model of argumentation.

Theoretical Background

Refutations and Counterexamples

As noted above, research on the logic of refutations has usually focused on counterexamples. There has been concern that students might be reluctant to reject statements based on one or two counterexamples: Galbraith (1981) reported that some students aged 12–17 preferred to classify statements as partially right; Zeybek (2017) observed that pre-service primary teachers with less well-developed deductive reasoning skills tended to believe that more counterexamples would strengthen refutations; and Ko and Knuth (2013) found that two of 16 mathematics majors accepted a proof and a counterexample for the same statement. Stylianides and Al-Murani (2010), in contrast, found that although some high-attaining year 10 students expressed agreement with or uncertainty about arguments for and against the same proposition, when followed up in interviews they recognised the inconsistency.

Overall, research is somewhat mixed but broadly consistent with the idea that in simple enough cases and when not fooled by salient examples in which a statement is true, students can apply the basic logic of counterexamples. Roh and Lee (2017), for instance, reported that undergraduates introduced and used counterexamples when considering possible definitions of sequence convergence. Yopp et al. (2020) reported that eighth graders could be trained to use 'eliminating counterexamples' as a framework for constructing or critiquing arguments for general claims. Hamami et al. (2021) reported that adults without specialist mathematics training

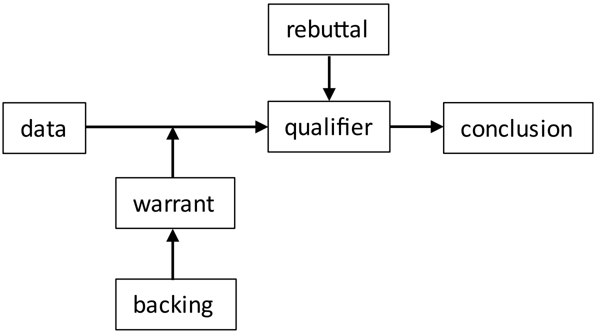


Fig. 1 Toulmin's model of argumentation

produced counterexamples when rejecting inferences about topological relations, though performance was impaired when counterexamples were 'further from' starting configurations. At a larger scale, Hoyles and Küchemann (2002) reported that of 2600 year 8 students presented with a simple false conjecture in number theory, 36% gave valid counterexamples. Küchemann and Hoyles (2006) reported that of 1500 students presented with a simple false conjecture in geometry, 42% of year 8 students and 56% of year 10 students gave valid counterexamples. Lin (2005) reported that of 2200 students presented with false geometry conjectures, between 4 and 34% of seventh and eighth graders gave counterexamples. In teacher education, Peled and Zaslavsky (1997) found that experienced teachers were almost universally able to give valid counterexamples. However, counterexamples might not be frequent in teaching: Zodik and Zaslavsky (2008) reported that of 604 teacher-produced examples across 54 observed lessons, only 18 were counterexamples.

We might therefore expect mathematics undergraduates to have limited experience in working with counterexamples, but to be able to handle the logic in simple cases. We might, however, expect them to have trouble generating counterexamples in more complex cases. Researchers have noted that mathematics undergraduates might fail to identify local counterexamples when validating purported proofs, and thus fail to recognise flaws in those arguments (Alcock & Weber, 2005; Ko & Knuth, 2013). They have also reported that it can be challenging for advanced students or for teachers to coordinate all conditions in a problem in order to produce valid counterexamples, for example in geometry problems with multiple components (Buchbinder & Zaslavsky, 2011; Komatsu et al., 2017; Potari et al., 2009), in problems involving continuous or quadratic functions (Ko & Knuth, 2009; Lee, 2017), and in purported proofs in calculus involving inequalities and absolute values (Giannakoulias et al., 2010).

Arguments and Warrants

To consider counterexamples alongside alternative refutations in a theoretically coherent way, we follow other researchers in employing Toulmin's (1958) model of argumentation. Toulmin's full model appears in Fig. 1, which represents *data* put forth in support of a *conclusion*, perhaps with a suitable *qualifier*; the data and conclusion are linked by a *warrant*, possibly with *backing*, and there might be a *rebuttal* capturing conditions under which the warrant would not adequately justify the conclusion based on the data.

In mathematics education, this model has been used in various contexts, including those involving refutation. Hoyles and Küchemann (2002), for instance, used it to model student arguments about whether an implication and its converse are equivalent. Giannakoulias et al. (2010) modelled arguments teachers formulated to convince hypothetical students that there were errors in written proofs. Inglis et al. (2007) pointed out that although various researchers have modelled mathematical arguments using only data, warrants, and conclusions, mathematical practice is better understood using the full model because both novice and expert mathematicians do use non-absolute arguments that admit rebuttals, so we should avoid assuming on limited evidence that qualifiers are absolute. We respect this point and in theoretical discussions we consider all six components; in analysing arguments that students put forth when evaluating and choosing between refutations, we simplify our diagrams by representing only those components that students invoke and those that are listed in the task.

In theoretical discussion, we also use the distinction Toulmin (2003) draws between warrants and backing, characterising warrants as ‘hypothetical, bridgelike statements’, whereas backing ‘can be expressed in the form of categorical statements of fact’ (p.98). Toulmin (1958) notes (pp.102–103) that an argument can be in the form with ‘data, warrant, so conclusion’, as in ‘(D) Petersen is a Swede, (W) A Swede is almost certainly not a Roman Catholic, so (C) Petersen is almost certainly not a Roman Catholic’, or in the form ‘data, backing, so conclusion’, as in ‘(D) Petersen is a Swede, (B) The proportion of Roman Catholic Swedes is minute, so (C) Petersen is almost certainly not a Roman Catholic’. This distinction is not straightforward – there are different possible relationships between warrants and backing (e.g. Castaneda, 1960; Simpson, 2015) – but in this form it is useful when we consider explanations students give to justify their choices between counterexamples and corrected statements.

We apply Toulmin’s model now to consider arguments about mathematical statements of the form $\forall x S(x)$. Refuting such a statement means arguing for the conclusion that $\forall x S(x)$ is false. When a student offers a counterexample x_0 or a corrected statement $\forall x T(x)$, these function as data in support of that conclusion, and this is all we see of the argument; see Fig. 2.

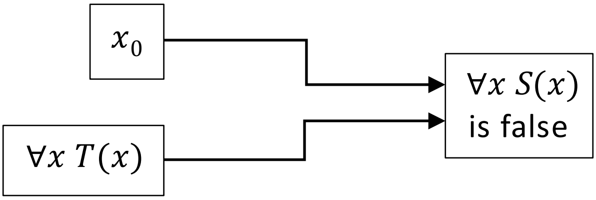


Fig. 2 Arguments that $\forall x S(x)$ is false, where data is a counterexample or a corrected statement

When the data is a counterexample x_0, a mathematically valid argument can be completed. The implicit warrant is the principle that a valid counterexample x_0 refutes a general statement $\forall x S(x)$, with factual backing that $S(x_0)$ is, indeed, false. Such an argument admits no rebuttal so is a valid refutation (the *statement* $\forall x S(x)$ is refuted because the *argument that it is false* admits no rebuttal); see Fig. 3. When the data is a corrected statement $\forall x T(x)$, the rest of the argument is less obvious. The student presumably intends to rely upon the fact that $\forall x T(x)$ is true, which is factual so in Toulmin's terms is understood as backing. The warrant is implicit, and several different warrants might be intended. Students might intend to say that $\forall x S(x)$ and $\forall x T(x)$ cannot both be true, so that the truth of $\forall x T(x)$ implies the falsity of $\forall x S(x)$ in a logical sense. Or they might intend to align with Grice's (1989) maxims for communication, relying on the assumption that if $\forall x T(x)$ is true, it is uncooperative instead to say $\forall x S(x)$, so that the truth of $\forall x T(x)$ implies the falsity of $\forall x S(x)$ in a conversational sense. Or they might rely on an assumed didactic contract (Brousseau, 1997), asserting that, because $\forall x T(x)$ is correct, it should be viewed favourably as a replacement for $\forall x S(x)$, without much notion of implication. In any of these cases, the argument *does* admit a rebuttal, because a true mathematical statement, even if closely related to an original, proves that original false only if the two are

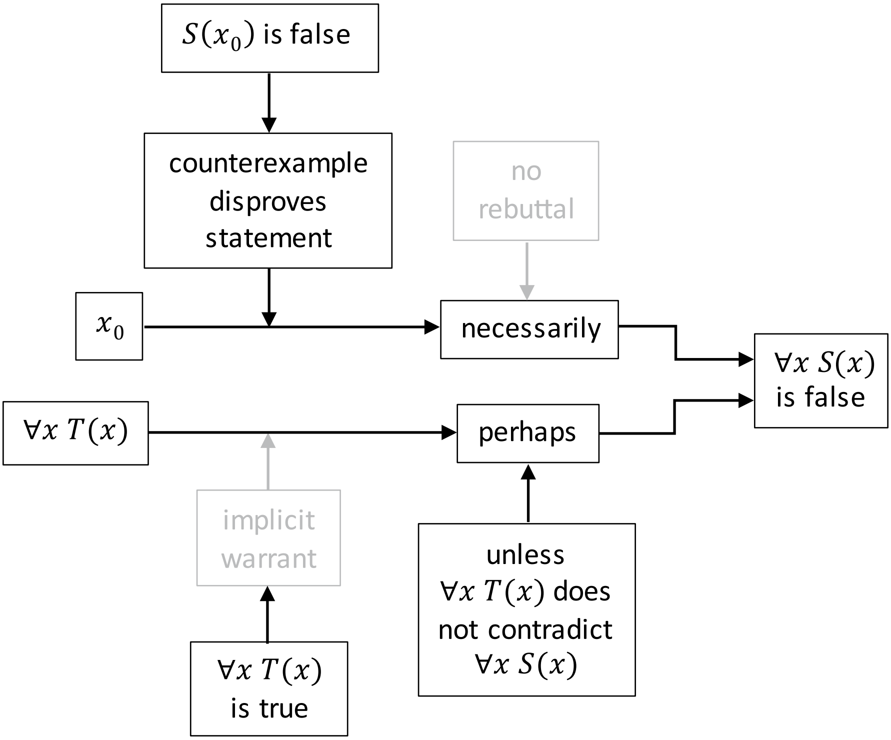

$S(x_0)$ is false
counterexample disproves statement
no rebuttal
x_0
necessarily
$\forall x\ S(x)$ is false
$\forall x\ T(x)$
perhaps
implicit warrant
unless $\forall x\ T(x)$ does not contradict $\forall x\ S(x)$
$\forall x\ T(x)$ is true

Fig. 3 Warrants, backing and rebuttals for arguments that $\forall x S(x)$ is false, where data is a counterexample or a corrected statement

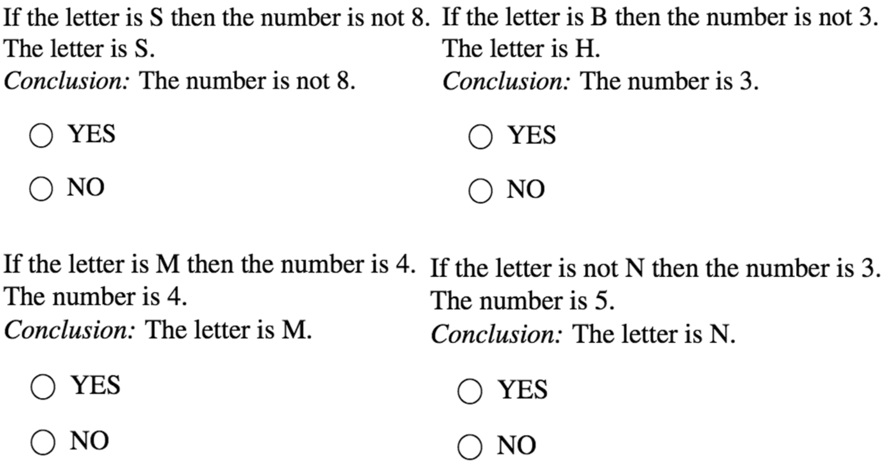
contradictory (the statement is *not* refuted because the argument that it is false *does* admit a rebuttal); again, see Fig. 3.

This breakdown makes clear the complex interactions between mathematical and everyday arguments. If a student refutes a statement by correcting it, this might indicate a failure of logical reasoning. But it might indicate that the student did not attempt logical reasoning due to the salience of an argument with an alternative warrant or backing. Or it might indicate that they did attempt logical reasoning and perhaps get it right, but did not express this conventionally. We think that these distinctions are important because learning in academic disciplines involves learning about the weight given to different types of justification. In mathematics, we care about clear communication, and we certainly care about true statements. But valid logic – in the mathematical rather than the everyday sense – trumps everything.

Conditional Inference

Logical reasoning, then, is crucial in making theoretical sense of refutations. But does it have a statistically significant effect on students' refutation attempts? Are students with better logical reasoning skills more likely to understand refutations in a mathematically normative way?

We investigated this by relating students' refutation responses to their performance on a standard conditional inference task, for which sample items appear in Fig. 4. Each item presents two premises concerning an imaginary letter-number pair, where the major premise is a conditional. The instructions read 'If you think the conclusion necessarily follows, please tick YES, otherwise tick NO'. The inferences are modus ponens (MP, top left), denial of the antecedent (DA, top right), affirmation of the consequent (AC, bottom left), and modus tollens (MT, bottom right).



If the letter is S then the number is not 8.
The letter is S.
Conclusion: The number is not 8.

○ YES

○ NO

If the letter is B then the number is not 3.
The letter is H.
Conclusion: The number is 3.

○ YES

○ NO

If the letter is M then the number is 4.
The number is 4.
Conclusion: The letter is M.

○ YES

○ NO

If the letter is not N then the number is 3.
The number is 5.
Conclusion: The letter is N.

○ YES

○ NO

Fig. 4 Four items from the conditional inference task used in this study (Evans et al., 1995)

Such tasks are very common in research on the psychology of reasoning, though they vary in content and their handling of negation (Evans et al., 1996). Our task had abstract content (as opposed to everyday real-world content) and implicit negation (an explicit negation of 'The letter is B' would be 'The letter is not B' rather than 'The letter is H' as appears in Fig. 4). Across such tasks, educated adults almost always correctly endorse MP inferences. But, beyond that, their responses vary systematically with content (E.g., Evans et al., 2015) and they do not reason in ways that align with the *material conditional* interpretation accepted as normatively correct (e.g., Oaksford & Chater, 2020). For instance, Evans et al. (2007) used an abstract task like ours and found that (non-mathematics) undergraduates endorsed only 50% of valid MT inferences, and endorsed 47% of invalid DA and 74% of invalid AC inferences.

We used this specific task for both theoretical and practical reasons. Theoretically, many mathematical theorems and conjectures can be expressed in the form of a universally quantified conditional '$\forall x$, if $P(x)$ then $Q(x)$'. The quantifier is often omitted (Solow, 2005), which can cause miscommunication when students interpret the conditional predicate 'if $P(x)$ then $Q(x)$' as lacking a truth value (Durand-Guerrier, 2003; Hub & Dawkins, 2018). Nevertheless, a valid counterexample can be understood as an x_0 for which $P(x_0)$ but not $Q(x_0)$, and rejecting invalid inferences is closely related to recognising that '$\forall x$, if $P(x)$ then $Q(x)$' is not necessarily refuted by '$\forall x$, if $U(x)$ then $Q(x)$' or by '$\forall x$, if $P(x)$ then $V(x)$'. We might therefore predict that students with better abstract conditional inference skills would work in more mathematically valid ways with refutations. Our research tests this prediction.

Practically, it has been established that performance in abstract conditional inference is linked to performance in mathematics. Research in the UK and Cyprus (Attridge & Inglis, 2013; Attridge et al., 2015) has shown that studying mathematics intensively at age 16–18 leads to better rejection of DA and AC inferences, but equivalent or poorer endorsement of MT inferences. Research at the undergraduate level has found that better rejection of invalid inferences predicted better performance on a proof comprehension test and in a proof-based course (Alcock et al., 2014). So, mathematics students can be expected to have better than average but imperfect conditional inference skills, and differences in these are known to have measurable effects on their performance in educationally relevant tasks. We extend this research by testing the relationship between abstract conditional inference skill and interpreting refutations.

Method

Participants and Administration

Our research took place in a real analysis course at a UK university. Approximately two thirds of the class of over 200 were first-year students spending 75–100% of their study time on mathematics; the remainder were second-year students on degree programmes with about 50% mathematics. All of the programmes, like many UK mathematics-focused programmes, required high prior mathematical attainment. The majority of participants therefore had an A or A* grade in A level Mathematics and A or B grades in their two[1] other A level subjects (or the equivalent, if they were from overseas); some had taken an extra A level and/or an A level in Further Mathematics, but neither this nor special entry examinations were required. These students were all in one class because the UK higher education system operates a cohort model, meaning that students on a specific programme all attend core courses together.

Our study used two paper instruments, administered at different times so that participants would be less likely to perceive them as linked. Both instruments provided informed consent information and asked for participants' university ID numbers; both ended with tick-boxes for agreement to data being used and to our acquiring their grades from the university's system once the course finished. Participants completed the refutation instrument in 15 min in week 3 of the course; 173 agreed that their data could be used. They completed the abstract conditional inference instrument in 10 min in week 1 of the course; 157 of the 173 agreed that their data could be used, and 151 of these gave permission to access their course scores.

Refutation Instrument

We designed a refutation instrument with items structured similarly to those for which we had previously observed invalid refutations. Because we wanted to provide early collective feedback, our items used content from the first two weeks of the course. We refer to them as the *reciprocal item*, *absolute value item*, and *sequence item* respectively:

- If $x < 3$ then $1/x > 1/3$.
- $\forall a, b \in \mathbb{R}, |a + b| < |a| + |b|$.
- A sequence (a_n) is increasing if and only if $\forall n \in \mathbb{N}, a_{n+2} \geq a_n$.

All three items are false (assuming universal quantification – we provide a note on this in the Results).

The refutation instrument contained four main pages. The *initial response* page showed all three items, spread out with space for responses. Participants were asked to 'Answer TRUE or FALSE to each question. For those that are FALSE, give a counterexample or a brief reason.' This instruction matched that used in the course on weekly non-assessed retrieval practice quizzes, so we used it here to avoid overlooking of or confusion over unfamiliar instructions. We acknowledge, however, that it might be read as implicitly condoning responses of both types. We consider implications of this in our interpretations.

[1] UK students specialise early. Because most students take only three A levels at ages 16–18, they arrive at university having already studied much of the material that would appear in the first 1–2 years of a US calculus sequence. Degree programmes often begin with courses covering the later parts of such a sequence, together with linear algebra and other proof-based courses.

After the initial response page, the refutation instrument had three *evaluation and forced choice* pages. Each showed one item again, together with two possible responses: FALSE with a counterexample and FALSE with a corrected statement. The counterexamples and corrected statements appear below

Reciprocal item
FALSE. Counterexample: If $x = -2$ then $1/x = 1/-2 < 1/3$.
FALSE. Reason: It should be 'If $0 < x < 3$ then $1/x > 1/3$'.
Absolute value item
FALSE. Counterexample: If $a = 1$ and $b = 6$ then $|a + b| = 7$ and $|a| + |b| = 7$.
FALSE. Reason: It should be '$\forall a, b \in \mathbb{R}, |a + b| \leq |a| + |b|$'.
Sequence item
FALSE. Counterexample:
The sequence $1, 3, 2, 4, 3, 5, 4, 6, \ldots$ satisfies the condition but is not increasing.
FALSE. Reason:
It should be 'A sequence $\left(a_n\right)$ is increasing if and only if $\forall n \in \mathbb{N}, a_{n+1} \geq a_n$'.

The layout for the evaluation and forced choice pages is illustrated in the Appendix (Fig. 9). As shown there, participants were asked, separately for each counterexample and reason, to select one option from:

- The answer is correct and the counterexample [reason] is valid;
- The answer is correct but the counterexample [reason] is invalid;
- The answer is incorrect.

They were then asked to make a forced choice, stating which response was better and why.

We intended the three evaluation options to make clear that correct/incorrect applied to the answer 'FALSE' and valid/invalid applied to the refutation provided by the counterexample or corrected statement; as this is arguably the only obvious interpretation for the counterexample, we expected that offering parallel options for the corrected statement would reinforce it. We intended that the forced choice would be interpreted as asking which of the counterexample and corrected statement constituted the better refutation. Under these interpretations, evaluating the corrected statement as valid would indicate reliance upon logically inappropriate warrants – as discussed under "Arguments and Warrants" above – and the explanations provided in the forced choice would provide insight into which warrants were invoked. However, as with many surveys, we cannot guarantee that every participant interpreted the questions as intended. Because participants were not explicitly asked whether the corrected statement proved the original false, it could be that some interpreted 'valid' as asking whether it was mathematically correct in isolation, and 'better' as asking about general mathematical quality rather than value as a refutation. We consider implications of these possibilities across our results sections, and draw the relevant issues together when discussing relationships between refutation responses and conditional inference task scores.

We constructed six versions of the refutation instrument, each with a different permutation of the three items. For each evaluation and forced choice page in each version, copies were created with the order of the counterexample and corrected statement randomised. To discourage participants from going back to change initial 'TRUE' responses, each evaluation and forced choice page ended with a space for participants to state that they would change their initial answers.

Conditional Inference Instrument

To measure conditional inference, we used a short version of the Abstract Conditional Inference Task (Evans et al., 1995), as used previously by Inglis and Simpson (2008) and Attridge and Inglis (2013). This comprised 16 items including those in Fig. 4, four each for MP, AC, DA, and MT inferences. The 16 items form the more difficult half of the original task because negations in the minor premise are implicit. The order of the items was randomised at the participant level.

Results

We present the results in four stages corresponding to our research questions. The section "Initial, Evaluation and Forced Choice Responses" presents quantitative data from the initial response, evaluation and forced choice stages of the refutation instrument. The section "Explanations for Forced Choices" presents qualitative detail on the explanations students gave for their forced choices. The section "Conditional Reasoning and Refutation" presents statistical analyses relating refutation responses to performance on the conditional inference instrument. Section "Conditional Inference, Refutation and Course Scores" presents statistical analyses relating both refutation responses and performance on the conditional inference instrument to course performance.

Table 1 Initial responses for the true/false task

	CEX	CS	Both	True	Other
Reciprocal	119 (69%)	5 (3%)	8 (5%)	38 (22%)	3 (2%)
Absolute value	61 (35%)	43 (25%)	26 (15%)	34 (20%)	9 (5%)
Sequence	25 (14%)	72 (42%)	12 (7%)	35 (20%)	30 (17%)

CEX Stated False and Gave Counterexample, *CS* Stated False and Gave Corrected Statement, *Both* Stated False and Gave Counterexample and Corrected Statement

Initial, Evaluation and Forced Choice Responses

Initial Responses

The 173 initial responses are summarised in Table 1. The two columns on the right show that about one fifth of participants incorrectly answered ‘true’ for each item (these were not the same participants each time – 19 answered ‘true’ twice and none answered ‘true’ three times); smaller numbers gave responses not classifiable according to our main distinctions. The three columns to the left show counts of participants who correctly answered ‘false’, split according to whether they provided counterexamples (including single counterexamples or correctly specified classes), corrected statements, or both.

Table 1 shows that participants were most likely to give normatively valid refutations for the reciprocal item. Of the 132 who gave a classifiable answer while avoiding the incorrect ‘true’ response, 127 (96%) included valid counterexamples (no participant wrote that the statement was neither true nor false, so it appears that these students did assume universal quantification). Only 3% gave a corrected statement alone, and only 5% gave both one or more counterexamples and a reason. This demonstrates understanding of the logic of counterexamples: most participants provided them in this simple case and most did not feel the need to elaborate by providing a corrected statement.

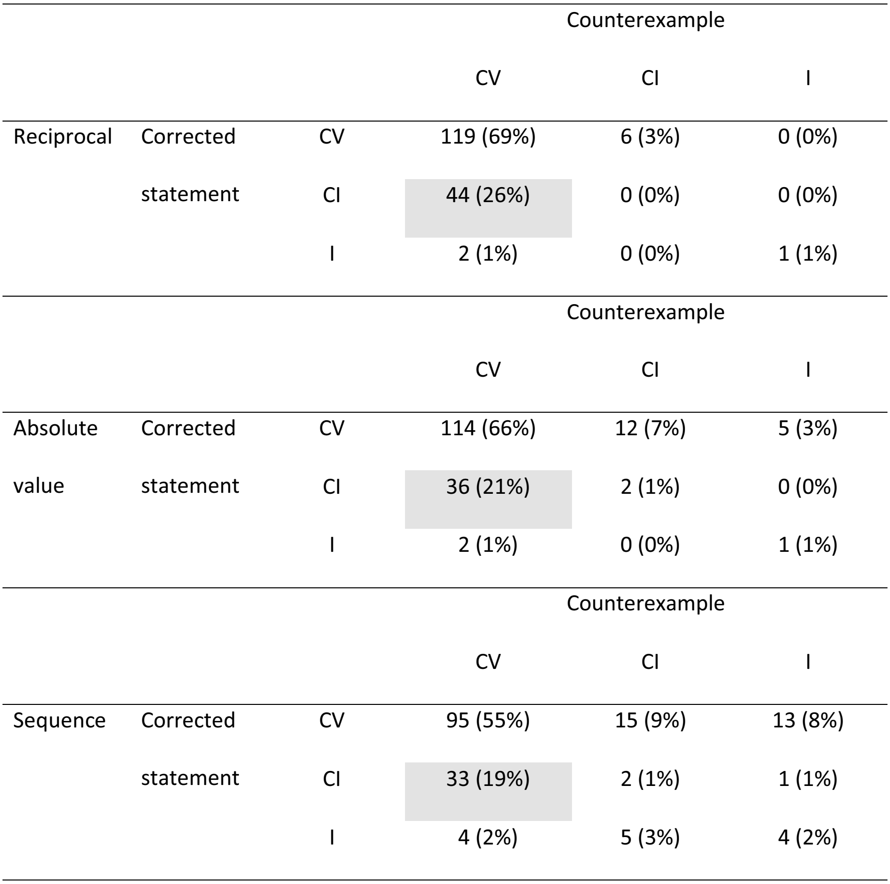
In more complex cases, however, the picture was different. For the absolute value item, 25% gave a corrected statement alone and, of the 130 who gave a classifiable answer while avoiding the ‘true’ response, 87 (67%) included valid counterexamples. This difference could be because a valid counterexample requires two numbers and is thus harder to generate, or because this item is more obviously related to a theorem from the course so that a corrected statement is more accessible. For the sequence item, 42% gave a corrected statement alone and, of the 109 who gave a classifiable answer while avoiding the ‘true’ response, only 37 (34%) included valid counterexamples. Again, this could be because a valid counterexample requires a sequence that is considerably harder to generate, or because this item is obviously related to a definition from the course. It is also worth noting that this item attracted more responses not classifiable according to this simple scheme.

Overall, initial responses were in line with our expectation that mathematics undergraduates would understand the logic of counterexamples but would be less likely to provide them when they are harder to construct or when corrected statements are likely to come to mind. Our interpretation is cautious because, as noted above, the task instruction might have discouraged critical evaluation of corrected statements. The next stage of our instrument addressed that in one way by asking explicitly for evaluations.

Evaluation Responses

The evaluation stage of the refutation instrument confronted every participant with both valid counterexamples and corrected statements that were not valid refutations.

Table 2 Evaluation responses for the true/false task



			Counterexample		
			CV	CI	I
Reciprocal	Corrected	CV	119 (69%)	6 (3%)	0 (0%)
	statement	CI	44 (26%)	0 (0%)	0 (0%)
		I	2 (1%)	0 (0%)	1 (1%)
			Counterexample		
			CV	CI	I
Absolute	Corrected	CV	114 (66%)	12 (7%)	5 (3%)
value	statement	CI	36 (21%)	2 (1%)	0 (0%)
		I	2 (1%)	0 (0%)	1 (1%)
			Counterexample		
			CV	CI	I
Sequence	Corrected	CV	95 (55%)	15 (9%)	13 (8%)
	statement	CI	33 (19%)	2 (1%)	1 (1%)
		I	4 (2%)	5 (3%)	4 (2%)

Blank or unreadable responses account for totals < 100%

CV answer correct and counterexample/reason valid, *CI* answer correct and counterexample/reason invalid, *I* answer incorrect

Table 2 summarises the evaluation data, with normatively correct responses shaded.[2] For all three items, almost all participants agreed that the answer FALSE was correct, at least for the simpler cases; numbers were slightly lower for the mathematically most complex item. However, most participants did not respond in a normatively correct way regarding validity. For instance, for the reciprocal item, 44 participants (26%) gave the normatively correct response, evaluating the counterexample as valid and the corrected statement as invalid.

[2] This table appeared in an early conference presentation of parts of this work, together with qualitative detail on the response types that is not presented here; see Alcock and Attridge (2022).

Far more, 119 (69%), evaluated both the counterexample and the corrected statement as valid. This pattern was repeated across all three items.

At this evaluation stage, participants were not giving spontaneous responses; the design was intended to encourage critical reflection on both possibilities. These data therefore present stronger evidence that participants did not reliably recognise the logical inadequacy of corrected statements as refutations. We remain cautious, because the initial task instruction might have nudged participants toward positive views of both counterexamples and reasons and because some might have applied valid/invalid to the corrected statement in isolation rather than as a refutation. If participants did interpret the questions as asking about validity as refutations, we believe that the scale on which they endorsed both provides educationally useful information and makes it interesting to see whether, when forced to choose, they would come down in favour of counterexamples (see the section "Forced Choice Responses") and how they would explain their choices (see the section "Explanations for Forced Choices"). If participants did not interpret the questions as asking about validity as refutations, this indicates that they were less sensitive to the question of mathematical refutation than we might hope, despite the parallel evaluation options for the counterexample and the corrected statement. This raises a question about whether logical reasoning performance – measured here with our abstract conditional inference task – influences sensitivity to the issue of refutation (see the section "Conditional Reasoning and Refutation").

Forced Choice Responses

The forced choice stage of the refutation instrument asked participants to compare counterexamples and corrected statements directly and decide which was better. Table 3 summarises the responses.

For each item, when forced to choose, a majority judged the counterexample better. But these majorities were not overwhelming: a substantial minority in each case judged the corrected statement better (a small number refused to choose). This held even for the reciprocal item, for which only 3% had initially given a corrected statement alone: 31% nevertheless judged the corrected statement better. This provides yet stronger evidence that some students either do not reason well enough to recognise the logical inadequacy of corrected statements as refutations – perhaps relying upon warrants that should carry less weight than mathematical logic – or are not sensitive to the mathematical issue of refutation (or both). To gain insight into these possibilities, we turn to a qualitative analysis of the explanations students gave for their forced choices.

Table 3 Forced choice responses

Item	Counterexample	Corrected Statement	Both
Reciprocal	112 (65%)	53 (31%)	6 (3%)
Absolute value	100 (58%)	64 (37%)	5 (3%)
Sequence	92 (53%)	71 (41%)	6 (3%)

Blank or unreadable responses account for totals < 100%

Explanations for Forced Choices

The initial, evaluation and forced-choice data show the prevalence of corrected statements in spontaneous responses and the value participants attached to counterexamples and corrected statements individually and comparatively. For information on their underlying reasoning, we turn to the explanations for their forced choices. In the section "Response Types", we qualitatively categorise these explanations, providing illustrations. In the section "Proportions of Explanation Types and Toulmin Analysis", we document the prevalence of explanations of different types and consider common types in relation to Toulmin's model.

Response Types

The forced-choice explanations were typically single sentences and straightforward to classify qualitatively. We identified 11 explanation types, and below we present a descriptive analysis showing illustrations for each type for each item (reciprocal, absolute value, and sequence) where these are pertinent to our analyses. We group the illustrations under explanations from those who judged the counterexample better, explanations from those who judged the corrected statement better, and (smaller in number) explanations exhibiting more serious misunderstandings, or refusals to choose, or nothing beyond the evaluation responses. Because randomisation affected participants' use of referents – the 'first answer' for some was the second for others – we write CEX where they referred to the counterexample and CS where they referred to the corrected statement.

Of the participants who judged a counterexample better, some explained only in relation to the counterexample (Type 1). Others commented additionally on the logical inadequacy of the corrected statement (Type 2).

Type 1: CEX disproves the statement

- *'CEX, because it gives a specific counterexample to disprove the statement.'*
- *'CEX because it uses a clear example to show it's false.'*
- *'CEX – valid counter example.'*

Type 2: CEX disproves the statement + CS does not

- *'CEX is better. It proves that the statement is false, whereas CS just provides an alternative true statement.'*
- *'CS states something with no reasoning or evidence whereas CEX gives evidence in the form of an example proving the statement to be wrong.'*
- *'CEX as it provides a reason for why the original statement is wrong. CS just states the actual definition, it does not prove the statement is wrong.'*

Participants who judged the corrected statement better gave a wider variety of explanations. Some, as anticipated, cited its status as correct (Type 3). Others focused on generality, judging the corrected statement more general (Type 4) or critiquing the counterexample as insufficiently general (Type 5). Some focused on expression, judging the corrected statement to have better terminology or notation (Type 6) or considering the counterexample ‘unfinished’ because it was not explicitly related to the statement (Type 7).

Type 3: CS is correct

- *‘CS because it amends the false statement to make it true.’*
- *‘CS because it gives the true definition of the statement.’*
- *‘CS as it is the correct definition.’*

Type 4: CS is more general

- *‘CS answer is better because it generalises the explanation rather than giving only one example.’*
- *‘CS as it is more general and applies in all cases.’*
- *‘The CS, as it is more general and “mathematicians like to generalise”.’*

Type 5: Single counterexample is unsatisfactory

- *‘CS because it shows a range of values for which x satisfies the equation and not just one value.’*
- *‘CS, since it actually means that the inequality is satisfied $\forall a, b \in \mathbb{R}$, whereas the second just takes two random integers from infinite amount of numbers.’*
- *[N/A for sequence item]*

Type 6: CS uses better terminology or notation

- *[N/A for reciprocal item]*
- *‘CS as it shows the student fully understands the “formula”.’*
- *‘CS: better terminology.’*

Type 7: CEX is unfinished

- *‘CS gives more detail whereas CEX just stops.’*
- *‘CS, as CEX is unfinished, in CEX they haven’t compared the values with the initial statement.’*
- *‘CS is better as, although the counterexample is correct, it doesn’t explain why it disproves it, the actual definition for increasing must be included also.’*

Alternative interpretations were evident in other responses. Small numbers of participants explicitly misinterpreted some aspect of logic in either the counterexample or the reason (Type 8); these are less pertinent to our analyses but are illustrated below for interest. Small numbers responded in terms of personal preferences, or refused to choose, or gave explanations that did not go beyond their evaluations; we refer to these later as types 9, 10 and 11 but omit illustrations.

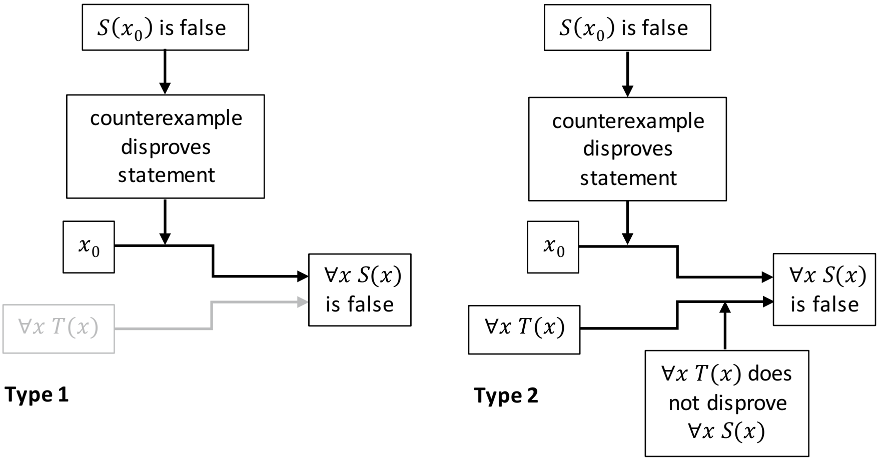
Type 8: Logic misunderstood

Table 4 Proportions of forced-choice response types for reciprocal, absolute value, and sequence items

Type	Reciprocal	Abs value	Sequence
1: CEX disproves the statement	58 (36%)	55 (32%)	56 (32%)
2: CEX disproves the statement + CS does not	46 (27%)	41 (24%)	26 (15%)
3: CS is correct	26 (15%)	25 (14%)	47 (27%)
4: CS is more general	15 (8%)	14 (8%)	8 (5%)
5: Single counterexample is unsatisfactory	2 (1%)	4 (2%)	0 (0%)
6: CS uses better terminology or notation	0 (0%)	5 (3%)	1 (1%)
7: CEX is unfinished	1 (1%)	6 (3%)	1 (1%)
8: Logic misunderstood	8 (5%)	6 (3%)	15 (8%)
9: Personal preference	7 (4%)	4 (2%)	7 (4%)
10: Both are valid	5 (3%)	5 (3%)	3 (2%)
11: Nothing added beyond evaluation response	5 (3%)	8 (5%)	9 (5%)

- *'CS because it gives a restricted domain so out of this range it must not be valid.'*
- *'The CS is better because the CEX implies that* $|a+b|=|a|+|b|$ *and not that can also be* $|a+b|<|a|+|b|$*.'*
- *'The CS isn't true as you could have a sequence* $a_n = 1, 1, 2, 2, 3, 3, \ldots$ *which is increasing but does not have* $a_{n+1} \geq a_n$*.'*

Overall, some responses showed good understanding of the logic of refutations. Some did not, but are reasonable in relation to Toulmin's model of argumentation – see below – or indicate a lack of sensitivity to the issue of mathematical refutation – see the section "Conditional Reasoning and Refutation".



$S(x_0)$ is false
counterexample disproves statement
x_0
$\forall x\ S(x)$ is false
$\forall x\ T(x)$
Type 1
$S(x_0)$ is false
counterexample disproves statement
x_0
$\forall x\ S(x)$ is false
$\forall x\ T(x)$
$\forall x\ T(x)$ does not disprove $\forall x\ S(x)$
Type 2

Fig. 5 Toulmin diagrams for explanations provided by participants who judged counterexamples better

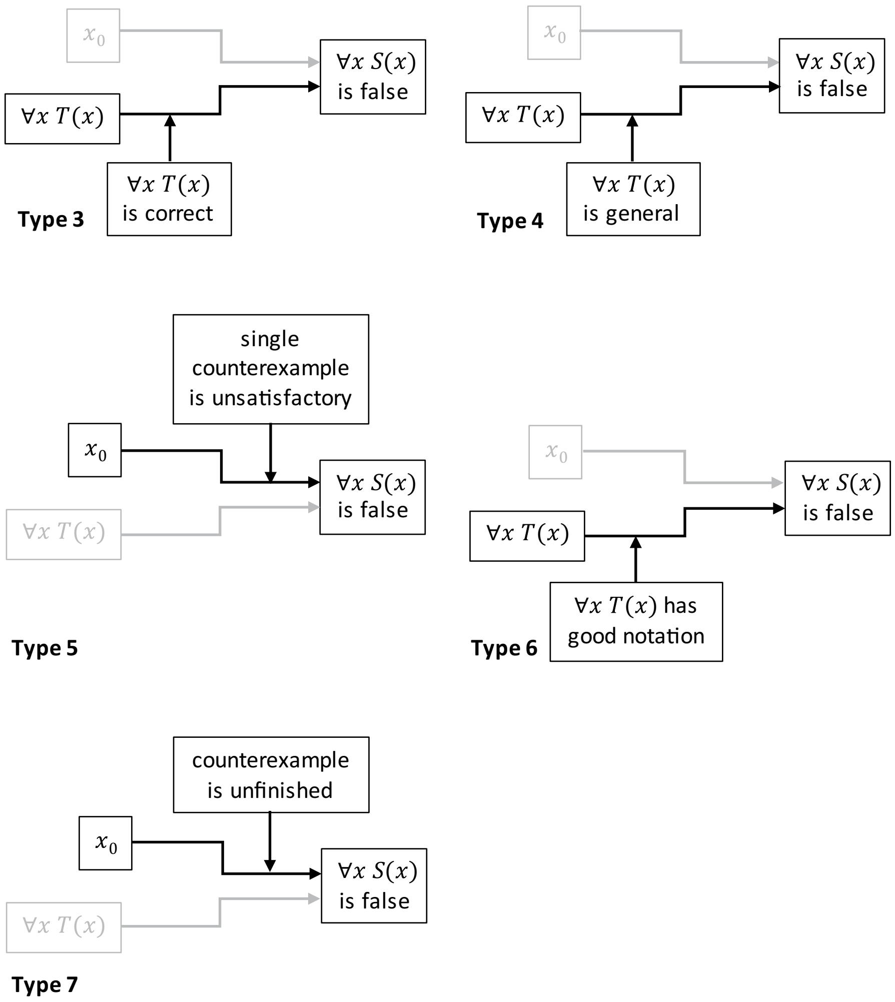


Fig. 6 Toulmin diagrams for explanations provided by participants who judged corrected statements better

Proportions of Explanation Types and Toulmin Analysis

The proportions of explanations of each type were broadly predictable based on participants' evaluation and forced choice responses. They are summarised in Table 4.

If we assume that participants' explanations were about which was the better refutation then, as in the section "Arguments and Warrants", they can be represented using Toulmin's (1958) model with the counterexample and/or corrected statement as data and '$\forall x S(x)$ is false' as conclusion. We make this assumption in Figs. 5 and 6, representing components explicitly addressed in the explanations in black and those offered in the task but not explicitly addressed in grey; all other components we omit.

Figure 5 represents explanations from participants who judged counterexamples better. Most fleshed out the mathematically normative argument discussed in the section "Arguments and Warrants". Encouragingly, up to about 40% of these participants provided not only a warrant for the counterexample-based argument (Type 1) but also an explicit rebuttal of the corrected statement-based argument (Type 2).

Explanations from participants who judged corrected statements better are represented in Fig. 6. The largest proportion focused on correctness of the corrected statement (Type 3). As discussed in the section "Arguments and Warrants", this amounts to providing an argument of the form ‘D, B, so C’, where backing is provided but the warrant is implicit. A warrant could be the relatively naïve notion that correctness is valued in mathematics, or a more sophisticated communicative principle that correcting something about the statement implies that there were counterexamples in the scope of the original statement but outside that of the corrected version. Of participants who focused on generality, most explained that the corrected statement is more general (Type 4), again providing an argument of the form ‘D, B, so C’ – the student who noted that “mathematicians like to generalise” provided a warrant too. Smaller numbers explained instead that a single counterexample is unsatisfactory (Type 5), thereby providing a logically invalid rebuttal to the counterexample-based argument; this doubt over single counterexamples has been noted elsewhere and represents a non-normative understanding of the relevant logic, but it was not prevalent in our data. Of participants who focused on mathematical expression, some explained that the corrected statement uses better mathematical terminology or notation (Type 6), again providing an argument of the form ‘D, B, so C’, presumably with the intended warrant that good terminology and notation are mathematically valued. Others explained that the counterexample was unfinished or inadequate (Type 7), which amounts to a didactic contract-based rebuttal to the counterexample-based argument, addressing presentation rather than logical validity.

Seen in this way, the explanation types – especially those most prevalent – are predictable based on the earlier research and the theoretical analysis presented in the section "Arguments and Warrants". The fact that they all appeared confirms the potential complexity that students face in reasoning about refutations.

This analysis, however, assumes that students were indeed reasoning about refutations, so that their explanations referred to arguments with the conclusions of the form ‘$\forall x S(x)$ is false’. We think this assumption sensible for Type 1 and Type 2 explanations, which explicitly stated that a counterexample proved its original statement false (Type 1) and, in some cases (Type 2), that the corresponding reason did not. It is less sensible for Type 3 – Type 7 explanations. If some participants applied ‘better’ to qualities of the corrected statements in isolation rather than as refutations, then their explanations might be better understood as arguments in which the conclusion is ‘the corrected statement is better mathematics’ rather than ‘$\forall x S(x)$ is false’. This raises the question of what drives sensitivity to refutation as an issue in mathematics, which we address next.

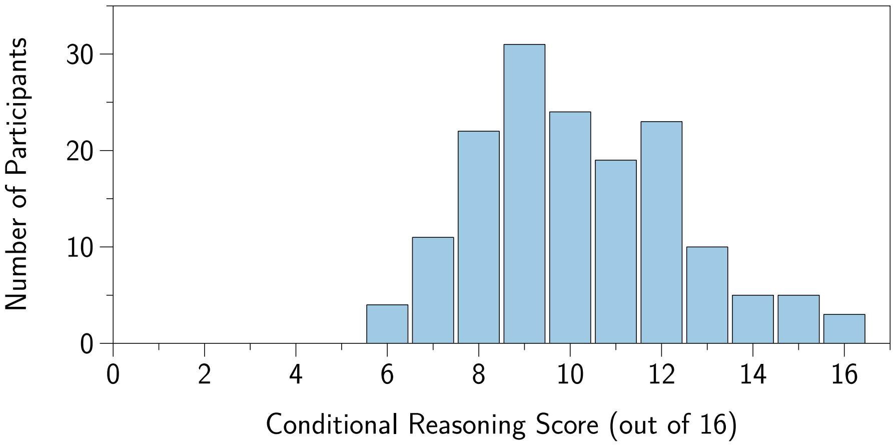


Fig. 7 Counts of total scores out of 16 for normatively correct responses on the conditional reasoning instrument

Conditional Reasoning and Refutation

All undergraduate students have been exposed to mathematical education in which correctness, generality and good presentation have been valued. So all might reasonably cite such issues when judging one mathematical response better than another, thereby failing to show sensitivity to logically valid refutation as an issue that should take mathematical precedence. Our remaining data provide information on whether this happens at random, or whether better attention to mathematically acceptable refutation is systematically associated with better logical reasoning as measured by performance in abstract conditional inference. In the section "Conditional Inference Performance", we summarise participants' performance on the abstract conditional inference task; in the section "Refutation Scores", we relate this to their refutation responses.

Conditional Inference Performance

Conditional inference scores for the 157 participants who allowed use of their data are summarised in Fig. 7 and Table 5. Figure 7 shows a broad distribution of total scores, with mean 10.22 out of 16 (SD = 2.25). Table 5 counts participants responding in a normatively correct way to 0, 1, 2, 3 or 4 items for modus ponens, denial of the

Table 5 Numbers (and percentages) of participants giving 0, 1, 2, 3 and 4 normatively correct responses on MP, DA, AC and MT items, with mean score out of 4

	0	1	2	3	4	Mean (SD)
MP	3 (1.7)	2 (1.2)	11 (6.4)	34 (19.7)	109 (63.0)	3.58 (0.73)
DA	19 (11.0)	35 (20.2)	33 (19.1)	34 (19.7)	38 (22.0)	2.26 (1.34)
AC	34 (19.7)	40 (23.1)	35 (20.2)	26 (15.0)	24 (13.9)	1.81 (1.35)
MT	11 (6.4)	23 (13.3)	33 (19.1)	54 (31.2)	38 (22.0)	2.57 (1.17)

Table 6 Distributions of initial, evaluation and choice scores out of 3

	0	1	2	3	Mean (SD)
Initial	23 (13%)	69 (40%)	62 (36%)	19 (11%)	1.45 (0.86)
Evaluation	104 (60%)	35 (20%)	16 (9%)	18 (10%)	0.70 (1.01)
Choice	33 (19%)	36 (21%)	44 (25%)	60 (35%)	1.76 (1.13)

antecedent, affirmation of the consequent, and modus tollens inferences. Both overall scores and the breakdown by inference type are approximately as expected given earlier research (e.g., Attridge & Inglis, 2013); performance was somewhat better than has been observed in non-mathematics-student samples, but far from normatively perfect.

Refutation Scores

To link conditional reasoning to refutation responses, we constructed three refutation scores, each in the range 0–3. *Initial score* counts the number of items for which participants included a single counterexample or a correctly specified class of counterexamples. *Evaluation score* counts the number of items for which participants endorsed only the counterexample – not the corrected statement – as a valid refutation. *Choice score* counts the number of items for which participants judged the counterexample better. Score distributions are summarised in Table 6 (which takes a by-participant perspective in contrast with the by-item perspective in earlier sections).

The distribution of initial scores shows that most participants included at least one counterexample, confirming that most could apply the logic of counterexamples in some cases; only 13% did not, which is partly attributable to erroneous 'true' responses. However, the distribution of evaluation scores shows that most participants did not make normatively correct evaluations: 60% never accepted the counterexample and rejected the corrected statement (unsurprisingly, given that the most common response on all items was that both counterexample and corrected statement were valid). The distribution of choice scores shows that 81% of participants judged the counterexample better at least once, but only 35% chose it for all three items. In fact, only 9 participants (5%) scored 3 out of 3 for all three stages. Overall, this confirms that most participants responded according to the mathematical logic of refutations sometimes but not reliably.

Conditional Inference and Refutation Scores

To investigate whether better logical reasoning predicts more normatively correct refutation responses, we conducted Spearman correlations between abstract conditional inference score and each of the refutation scores, correcting for multiple comparisons by using $\propto = 0.05/3 = 0.0167$. Abstract conditional inference score was significantly related to initial score ($r_s(157) = .273$, $p = .001$), evaluation score ($r_s(157) = .192$, $p = 0.016$) and choice score ($r_s(157) = .224$, $p = .005$); students with

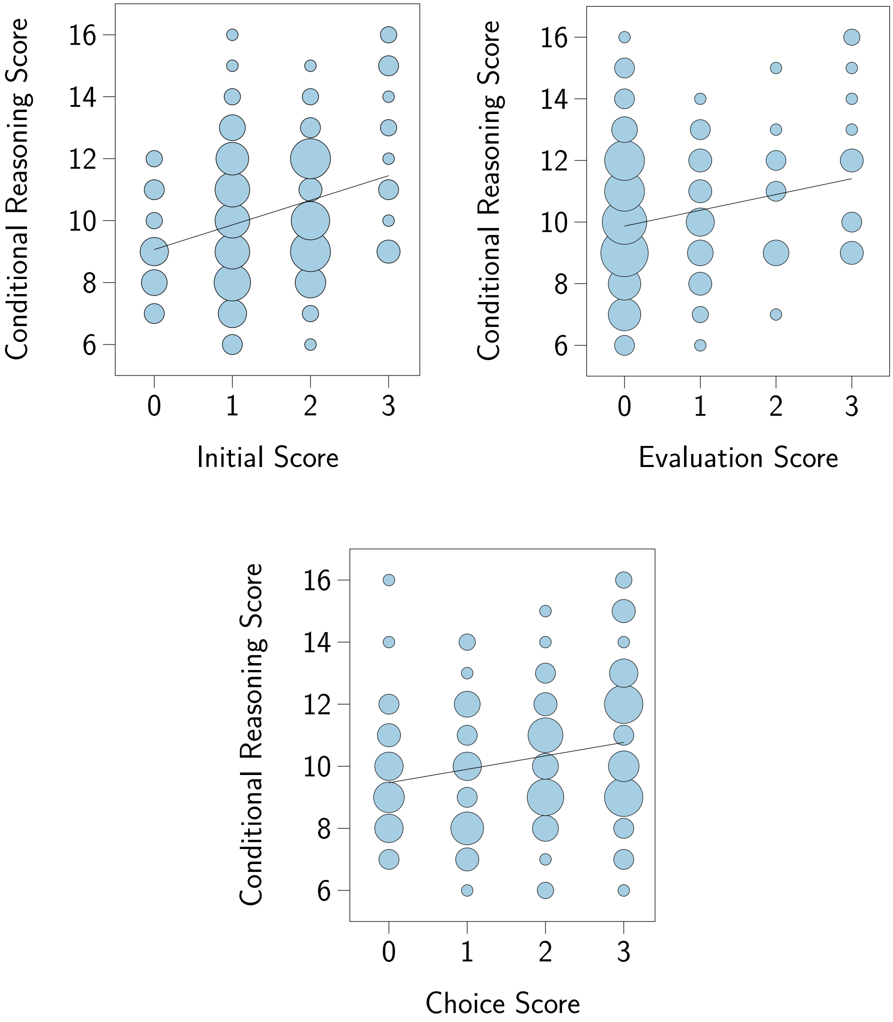


Fig. 8 Bubble plots showing conditional reasoning score against initial, evaluation and choice scores. *Note*. Bubble area represents the number of participants with each pair of scores

higher conditional inference scores were more likely to give counterexamples, evaluate only counterexamples as valid, and choose counterexamples as better refutations. The correlations are fairly small: in Fig. 8, the line of best fit in the Initial Score plot shows that participants who gave three counterexamples on average answered two more conditional inference items correctly than those who gave none, and the bubble plots show the wide spread of conditional inference scores for all four possible refutation scores. However, this shows that participants with better abstract conditional inference scores were more likely to answer across all three tasks in ways in line with mathematically valid refutations. This means that they were more likely to interpret our questions as intended and answer them in normative ways.

Conditional Inference, Refutation and Course Scores

We use this short final Results section to take a step back, relating both conditional reasoning and refutation scores to a standard measure of learning: performance in the analysis course. For the 151 participants who gave permission to access their course scores, we first ran a Pearson correlation between course grade (scored 0–100) and abstract conditional inference score (0–16). This revealed a significant positive correlation with a medium effect size, $r(141) = .355$, $p < .001$; participants with better conditional inference performance tended to achieve higher grades. In line with the results in Alcock et al. (2014), this shows that conditional reasoning predicts grades in a proof-based course. We then ran Spearman correlations between course grade (scored 0–100) and initial score (0–3), evaluation score (0–3) and choice score (0–3). We found small positive correlations between course grade and initial score, $r_s(151) = .150$, $p = .067$, between course grade and evaluation score, $r_s(151) = .125$, $p = .127$, and between course grade and choice score, $r_s(151) = .192$, $p = .018$. These were all in the expected direction but not significant with alpha level corrected to 0.0167.

Discussion

This paper addresses an issue that we believe is important but under-examined: that of whether students understand the logic of refutations. We designed a novel three-stage instrument to assess not only spontaneous refutations, but also how students evaluate counterexamples and corrected statements, and which they think better and why. We related students' explanations for their choices to Toulmin's model of argumentation, and their refutation responses to performance on an abstract conditional inference task.

On the one hand, we found some results indicating mathematically appropriate understandings of refutation. In the initial stage of our refutation instrument, many students gave valid counterexamples. Some gave both valid counterexamples and corrected statements, which might be considered the perfect response (cf. Lakatos, 1976). Where students gave only corrected statements, almost all were true, showing that students could incorporate lemmas appropriately and/or had learned course material. Large majorities of students evaluated counterexamples as valid, and more than half chose them as the better refutations, explaining that they proved the original statements false and in some cases supplying rebuttals to arguments based on corrected statements. This last finding especially is encouraging for classroom practice: if similar tasks were used in discussion-based activities, we would expect a focus on mathematically valid refutation to be common enough to support student-generated normative arguments.

On the other hand, our results provide a sense of where students respond to refutations in mathematically normative ways and where they do not. Students were less likely to provide counterexamples where requirements were more complex and statements were more obviously related to a theorem or definition from the course. Large majorities evaluated corrected statements as valid, and substantial minorities judged them better than the counterexamples. On this evidence, it seems that students might benefit from input encouraging them to recognise that a statement that is correct, general and expressed using good terminology and notation might nevertheless fail to refute a related statement. We consider it important to provide this input. Mathematics instructors might not always want to focus on logical warrants – in classroom communities, norms regarding correctness, generality and expression must be developed too. And they almost certainly will not want to devalue corrected statements – building mathematical theory is an important part of the job. But they will want students to develop a sense of priority in mathematical warrants, recognising that an alternative statement is a valid refutation only if it logically contradicts an original. And they will want students to develop the skills to work out whether and when this is the case.

In research terms, our data leave questions open because they do not allow us to disaggregate the effects of some variables. First, our more complex statements were also more closely related to course material; in future studies, this could be controlled by avoiding statements close to course material or by matching course-based statements and more neutral-content statements for complexity. Second, our task instructions might have implicitly condoned acceptance of both counterexamples and reasons, leading more students to accept invalid refutations than would otherwise have been the case; future studies might use a more neutral request to 'justify your answer', might explicitly ask whether a corrected statement proves the original false, or might use interviews to explore student views in more depth. Third, now that we have a pool of student explanations, a follow-up study might ask participants to evaluate, rank or compare these. Again, interview studies could explore student views in more depth.

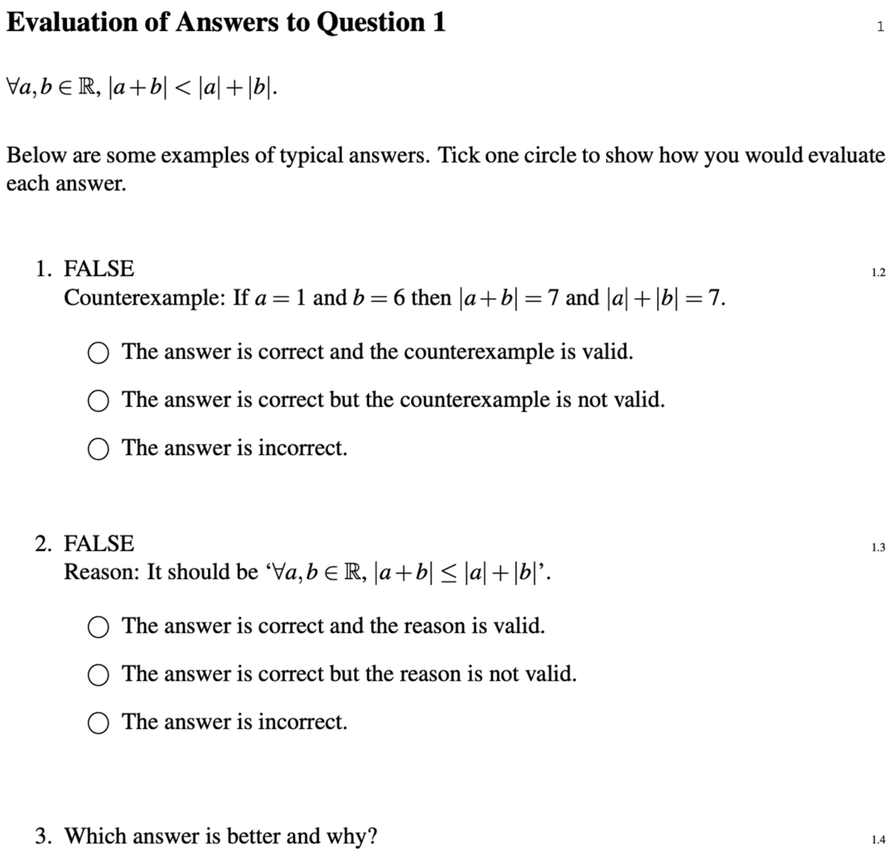
Interview studies do not, however, provide information at scale (usually). We consider scale a strength of our study: while our participant numbers are not in the thousands like some school-based studies of proof and refutation (Hoyles & Küchemann, 2002; Küchemann & Hoyles, 2006; Lin, 2005), we believe that having 173 participants provides a sense of the range and prevalence of arguments that might be produced elsewhere.

More importantly, this scale enabled us not only to study participants' refutation responses but also to relate these systematically to performance on a standard abstract conditional inference task. We established statistically that conditional inference skill predicts refutation responses: students with higher conditional inference scores were significantly more likely to give counterexamples, evaluate only counterexamples as valid, and judge counterexamples to be better. This is theoretically pertinent to the issue of learning about the relative mathematical value of different justifications. As noted earlier, corrected statements have epistemological value: teachers and researchers might recognise their role in mathematical theory-building (De Villiers, 2004; Yang, 2012) and perhaps therefore consider single counterexamples unsatisfactory stopping points (Creager, 2022; Lee, 2017; Peled & Zaslavsky, 1997; Zeybek, 2017). Indeed, researchers have designed tasks with the explicit goal of having students construct amended conjectures (Koichu, 2008; Yang, 2012; Yopp, 2013). Here, the fact that participants with better conditional inference scores were more likely to value counterexamples – despite possible ambiguity in our task instructions – indicates that better logical reasoning makes people more likely to override didactic messages they might have received about correctness, generality and expression, and expect a logically warranted refutation.

This does not mean, of course, that each individual would always respond in the same way, which returns us to considerations central in Toulmin's (1958) work. Toulmin wanted to understand arguments made by real people about situations with varied substantive content. We believe that this might be pertinent in understanding how our results contrast with other extant work. Specifically, our participants produced counterexamples at fairly high rates, where teachers in some studies did not (e.g., Giannakoulias et al., 2010). This could be related to the wider argumentative context. In studies of refutation, teachers have often been asked to refute hypothetical student-produced arguments. While this can be done with counterexamples, it makes sense that teachers might attempt instead to infer students' reasoning and explain why it does not apply, focusing on inferred warrants rather than rebuttals of the main statement. In such contexts and in ours, it might therefore be illuminating to follow other research on proof (e.g., Healy & Hoyles, 2000) by manipulating who students and teachers are asked to address: do they give and endorse different refutations if answering for a mathematician, a fellow student, or a younger student, for instance?

In all, our study addressed understanding of refutations. It used an original three-stage instrument to collect not only spontaneous refutation attempts but also separate and comparative evaluations of counterexamples and corrected statements. By using Toulmin's model of argumentation, we offered a theoretically coherent picture of where students focus when evaluating refutations and how non-logical warrants, backing and rebuttals might draw attention away from mathematically valid logic. We also used the scale of our study to establish that logical reasoning as measured using a standard abstract conditional inference task significantly predicts normatively valid responses to possible refutations. Our research suggests a number of possible avenues for productive future enquiry, as outlined in this discussion. We suggest that research might profitably elucidate what it takes to ensure that students see past non-logical considerations and engage with mathematical logic, and what it takes to help them understand that logic well enough to recognise when two statements are and are not contradictory.

Appendix



Evaluation of Answers to Question 1 1

$\forall a, b \in \mathbb{R},\ |a+b| < |a| + |b|.$

Below are some examples of typical answers. Tick one circle to show how you would evaluate each answer.

1. FALSE 1.2
   Counterexample: If $a = 1$ and $b = 6$ then $|a+b| = 7$ and $|a| + |b| = 7$.

   ○ The answer is correct and the counterexample is valid.

   ○ The answer is correct but the counterexample is not valid.

   ○ The answer is incorrect.

2. FALSE 1.3
   Reason: It should be '$\forall a, b \in \mathbb{R},\ |a+b| \leq |a| + |b|$'.

   ○ The answer is correct and the reason is valid.

   ○ The answer is correct but the reason is not valid.

   ○ The answer is incorrect.

3. Which answer is better and why? 1.4

Fig. 9 Sample evaluation and forced choice page from the refutation instrument

Acknowledgements This research was unfunded. Thank you to the three anonymous reviewers whose comments improved this manuscript.

Author Contributions Lara Alcock conceived and designed the study, collected, analysed, and interpreted the data, and wrote and revised the manuscript. Nina Attridge contributed to the design of the study, the collection, analysis and interpretation of the data, and provided feedback on the manuscript.

Declarations

Conflict of Interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Alcock, L., & Attridge, N. (2022). Counterexamples and refutations in undergraduate mathematics. In *Proceedings of the 2022 Conference on Research in Undergraduate Mathematics Education*. Boston, MA, USA: SIGMAA on RUME.

Alcock, L., Bailey, T., Inglis, M., & Docherty, P. (2014). The ability to reject invalid logical inferences predicts proof comprehension and mathematics performance. In *Proceedings of the 17th Conference on Research in Undergraduate Mathematics Education*. Denver, CO: SIGMAA on RUME.

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*(1), 20–37.

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

Balacheff, N. (1991). Treatment of refutations: Aspects of the complexity of a constructivist approach to mathematics learning. In E. von Glasersfeld (Ed.), *Radical constructivism in mathematics education* (pp. 89–110). Kluwer Academic Publishers.

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

Buchbinder, O., & Zaslavsky, O. (2011). Is this a coincidence? The role of examples in fostering a need for proof. *ZDM: The International Journal on Mathematics Education, 43*, 269–281.

Castaneda, H. N. (1960). On a proposed revolution in logic. *Philosophy of Science, 27*, 279–292.

Creager, M. A. (2022). Geometric refutations of prospective secondary mathematics teachers. *International Journal of Education in Mathematics, Science and Technology, 10*, 74–99.

De Villiers, M. (2004). The role and function of quasi-empirical methods in mathematics. *Canadian Journal of Science, Mathematics and Technology Education, 4*, 397–418.

Durand-Guerrier, V. (2003). Which notion of implication is the right one? From logical considerations to a didactic perspective. *Educational Studies in Mathematics, 53*(1), 5–34.

Evans, J. S. B., Clibbens, J., & Rood, B. (1996). The role of implicit and explicit negation in conditional reasoning bias. *Journal of Memory and Language, 35*(3), 392–409.

Evans, J. S. B., Handley, S. J., Neilens, H., & Over, D. E. (2007). Thinking about conditionals: A study of individual differences. *Memory & Cognition, 35*(7), 1772–1784.

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

Evans, J. S. B. T., Thompson, V. A., & Over, D. E. (2015). Uncertain deduction and conditional reasoning. *Frontiers in Psychology, 6*, 398.

Galbraith, P. L. (1981). Aspects of proving: A clinical investigation of process. *Educational Studies in Mathematics, 12*, 1–28.

Giannakoulias, E., Mastorides, E., Potari, D., & Zachariades, T. (2010). Studying teachers' mathematical argumentation in the context of refuting students' invalid claims. *Journal of Mathematical Behavior, 29*, 160–168.

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

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

Healy, L., & Hoyles, C. (2000). A study of proof conceptions in algebra. *Journal for Research in Mathematics Education, 31*, 396–428.

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.

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*(3), 187–204.

Ko, Y.-Y., & Knuth, E. (2009). Undergraduate mathematics majors' writing performance producing proofs and counterexamples about continuous functions. *Journal of Mathematical Behavior, 28*, 68–77.

Ko, Y.-Y., & Knuth, E. J. (2013). Validating proofs and counterexamples across content domains: Practices of importance for mathematics majors. *Journal of Mathematical Behavior, 32*, 20–35.

Koichu, B. (2008). If not, what yes? *International Journal of Mathematical Education in Science and Technology, 39*, 443–454.

Komatsu, K. (2017). Fostering empirical examination after proof construction in secondary school geometry. *Educational Studies in Mathematics, 96*, 129–144.

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

Komatsu, K., & Jones, K. (2022). Generating mathematical knowledge in the classroom through proof, refutation, and abductive reasoning. *Educational Studies in Mathematics, 109*, 567–591.

Komatsu, K., Jones, K., Ikeda, T., & Narazaki, A. (2017). Proof validation and modification in secondary school geometry. *Journal of Mathematical Behavior, 47*, 1–15.

Küchemann, D., & Hoyles, C. (2006). Influences on students' mathematical reasoning and patterns in its development: Insights from a longitudinal study with particular reference to geometry. *International Journal of Science and Mathematics Education, 4*, 581–608.

Lakatos, I. (1976). *Proofs and refutations*. Cambridge University Press.

Larsen, S., & Zandieh, M. (2008). Proofs and refutations in the undergraduate mathematics classroom. *Educational Studies in Mathematics, 67*, 185–198.

Lee, K. (2017). Students' proof schemes for mathematical proving and disproving of propositions. *Journal of Mathematical Behavior, 41*, 26–44.

Lin, F.-L. (2005). Modeling students' learning on mathematical proof and refutation. In H.L. Chick & J.L. Vincent (Eds.), *Proceedings of the 29th Conference of the International Group for the Psychology of Mathematics Education*, Vol.1, pp. 3–18. Melbourne, Australia: PME.

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

Peled, I., & Zaslavsky, O. (1997). Counter-examples that (only) prove and counter-examples that (also) explain. *Focus on Learning Problems in Mathematics, 19*(3), 49–61.

Potari, D., Zachariades, T., & Zaslavsky, O. (2009). Mathematics teachers reasoning for refuting students' invalid claims. In V. Durand-Guerrier, S. Soury-Lavergne & F. Arzarello, (Eds.), *Proceedings of the Sixth Congress of the European Society for Research in Mathematics Education*. Lyon, France: ERME.

Reid, D. (2002). Conjectures and refutations in Grade 5 mathematics. *Journal for Research in Mathematics Education, 33*, 5–29.

Roh, K. H., & Lee, Y. H. (2017). Designing tasks of introductory real analysis to bridge a gap between students' intuition and mathematical rigor: The case of the convergence of a sequence. *International Journal of Research in Undergraduate Mathematics Education, 3*, 34–68.

Simpson, A. (2015). The anatomy of a mathematical proof: Implications for analyses with Toulmin's scheme. *Educational Studies in Mathematics, 90*, 1–17.

Solow, D. (2005). *How to read and do proofs*. Hoboken, NJ: John Wiley & Sons Inc.

Stylianides, A. J., & Al-Murani, T. (2010). Can a proof and a counterexample coexist? students' conceptions about the relationship between proof and refutation. *Research in Mathematics Education, 12*, 21–36.

Stylianides, A. J., & Ball, D. L. (2008). Understanding and describing mathematical knowledge for teaching: Knowledge about proof for engaging students in the activity of proving. *Journal of Mathematics Teacher Education, 11*, 307–332.

Stylianides, G. J., & Stylianides, A. J. (2009). Facilitating the transition from empirical arguments to proof. *Journal for Research in Mathematics Education, 40*, 314–352.

Toulmin, S. (2003). *The uses of argument* (Updated). Cambridge University Press.

Toulmin, S. (1958). *The uses of argument*. Cambridge University Press.

Weber, K. (2010). Mathematics majors' perceptions of conviction, validity and proof. *Mathematical Thinking and Learning, 12*, 306–336.

Yang, K.-L. (2012). Providing opportunities for students to create mathematics. *Procedia – Social and Behavioral Sciences, 46*, 3905–3909.

Yim, J., Song, S., & Kim, J. (2008). Mathematically gifted elementary students' revisiting of Euler's polyhedron theorem. *The Mathematics Enthusiast, 5*, 125–142.

Yopp, D. A. (2013). Counterexamples as starting points for reasoning and sense making. *The Mathematics Teacher, 106*, 674–679.

Yopp, D. A., Ely, R., Adams, A. E., Neilsen, A. W., & Corwine, E. C. (2020). Eliminating counterexamples: A case study intervention for improving adolescents' ability to critique direct arguments. *Journal of Mathematical Behavior, 57*, 100751.

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

Zeybek, Z. (2017). Pre-service elementary teachers' conceptions of counterexamples. *Inter- National Journal of Education in Mathematics, Science and Technology, 5*, 295–316.

Zodik, I., & Zaslavsky, O. (2008). Characteristics of teachers' choice of examples in and for the mathematics classroom. *Educational Studies in Mathematics, 69*, 165–182.