Does understanding moderate aesthetic appraisals of proofs?

George Kinnear[1] and Matthew Inglis[2]
[1]University of Edinburgh
[2]Loughborough University

Abstract

The relationship between understanding and aesthetic appraisal in mathematics is an open question, with implications for both the philosophy of mathematics and mathematics education. In this study, we investigated how undergraduate students' understanding of a mathematical proof relates to their perception of its aesthetic value. Participants were asked to evaluate the proof's aesthetics and to complete three different assessments of their understanding. The results reveal that self-reported understanding was moderately associated with aesthetic appraisals, while two performance-based measures of understanding showed close-to-zero relationships. These findings challenge the view that aesthetic judgements in mathematics are merely disguised epistemic judgements, and suggest that future research should focus on exploring the non-epistemic factors that shape aesthetic judgements. We conclude by discussing the implications of these results for educational practices that seek to promote aesthetic experiences.

Keywords: Aesthetics; Proof appraisal; Proof comprehension

Introduction

It is clear that aesthetics is central to mathematicians' experiences of mathematics. Poincaré (1914), for example, asserted that mathematical beauty is a "real aesthetic feeling that all true mathematicians recognize" (p. 59). Consistent with Poincaré's view, many research-level mathematicians have testified that they base decisions on which lines of inquiry to follow on aesthetic factors (e.g., Weyl in Reid, 1986), and prizes for mathematical research are regularly awarded on the basis that the recipient's work is beautiful, deep or profound (Holden & Piene, 2010). Because of the widespread involvement of aesthetics in

George Kinnear https://orcid.org/0000-0003-4191-4258
Correspondence concerning this article should be addressed to George Kinnear, School of Mathematics and Maxwell Institute for Mathematical Sciences, The University of Edinburgh, Edinburgh, UK. E-mail: G.Kinnear@ed.ac.uk

mathematics and, in light of the view that decisions about mathematics education should be, in part, guided by a desire to authentically reflect the discipline of mathematics (Ball & Bass, 2000; Schoenfeld, 2016; Stylianou, 2002; Weber et al., 2014), many mathematics educators have studied and theorised the role that aesthetics could or should have in the classroom (Borasi & Brown, 1985; Dietiker, 2013, 2015a, 2015b, 2016; Dietiker, Riling, et al., 2023; Dietiker, Singh, et al., 2023; Gadanidis & Hoogland, 2003; Koichu et al., 2017; Papert, 1980; Richman et al., 2019; Sinclair, 2001; Wong, 2007; Zazkis et al., 2009).

We begin by briefly characterizing how philosophers conceptualize aesthetics outside of mathematics. Although the exact meaning of the term is, naturally, disputed, all agree that it goes well beyond the notion of beauty. One attempt at distinguishing aesthetic from non-aesthetic concepts is due to Sibley (1959). He suggested that aesthetic concepts cannot be ascribed using objective criteria, but rather that taste and perceptiveness must be used instead. For example, he contrasted the assessment that a novel has "a great number of characters" with the assessment that it has "characters that never really come to life". This latter judgement is aesthetic, suggests Sibley, as it "requires the exercise of taste, perceptiveness, or sensitivity" (p. 421). Hungerland (1962) agreed with Sibley's characterization, but suggested it needed to be supplemented. She pointed out that a key distinction between aesthetic and non-aesthetic terms is that the former admit to an is/looks distinction: a door may look like it is solid, despite actually being flimsy. In contrast, it is hard to see how a mathematical proof could appear to be striking if it is, in truth, unremarkable. Hungerland noted that no tests, checks or criteria can be developed to determine the presence of an aesthetic concept analogous to the kinds of tests, checks or criteria that are routinely developed for non-aesthetic criteria.

How can aesthetic judgements of the type described by Sibley (1959) and Hungerland (1962) be incorporated into mathematics education? One approach is to draw on students' existing aesthetic notions from non-mathematical settings, such as art. Lessons incorporating, say, the golden ratio might be one such method. But Sinclair (2001) argued that this runs the risk of endorsing "the belief that mathematics itself is aesthetically sterile" (p. 25), an approach which would undermine the aesthetic possibilities inherent in mathematics. Instead, she suggested that it was possible to incorporate genuinely mathematical aesthetic experiences into the classroom, and gave several inspiring examples. She characterised aesthetically rich learning experiences as being those which "enable children to wonder, to notice, to imagine alternatives, to appreciate contingencies, and to experience pleasure and pride" (p. 26).

Dietiker (2015b, 2016) went further by proposing strategies by which such experiences could be designed. She suggested that one way of designing aesthetically rich learning experiences, in Sinclair's (2001) sense, is to focus on how mathematical narratives unfold across a lesson or a curriculum. Dietiker, Singh, et al. (2023) designed a series of lessons which aimed to offer aesthetic experiences for students, and compared students' reactions to these lessons with a sample of comparison lessons that had not been designed with this focus in mind. They found that lessons perceived to be aesthetically rich by the students themselves (those which students rated as being more interesting, and which they described using aesthetically loaded adjectives such as 'surprising', 'intriguing', 'fascinating', etc.) tended to have distinct narrative characteristics. These lessons also tended to have questions which remained open for longer, and in which incremental

progress was made towards answering an overarching question throughout the session. Moreover, aesthetically rich lessons typically had more instances of misdirection and a greater number of open questions. In sum, Dietiker et al.'s work provides a framework for designing lessons with high aesthetic value, or at least high perceived aesthetic value.

Although some mathematics education researchers have focused on aesthetics at the lesson level, others have—like most philosophers interested in mathematical aesthetics—concentrated on students' appraisals of mathematical proofs. Sangwin and Kinnear (2024) asked undergraduate students to read and compare different proofs of the fact that the sum of the first n odd numbers is n^2. They found that their participants were able to distinguish between different proofs on a range of subjective dimensions, even though they had limited experience of making subjective judgements about proofs. However, they noted that students gave similar responses regardless of whether they were asked to judge how rigorous the proof was, or the extent to which a proof provided insight (which was characterised as an aesthetic value). Insight and rigour are typically regarded as quite distinct constructs in the philosophy of mathematics literature, raising the question of whether the kinds of distinctions made by philosophers are intuited in the same way by undergraduate students.

These considerations raise an important question. Is the notion of mathematical aesthetics deployed by Dietiker, Singh, et al. (2023), Sinclair (2001) and others, the same notion of mathematical aesthetics described by mathematicians? There are reasons to suppose not. Specifically, some philosophers of mathematics deny that the apparently aesthetic appraisals made by mathematicians are genuinely aesthetic, in the sense used by Sibley (1959) and Hungerland (1962), at all. Instead they suggest that the key to understanding aesthetic appraisal in mathematics is determining what non-aesthetic characteristic is actually being assessed in such situations.

Harré (1958) exemplified this position, sometimes referred to as the reductive or non-literal view of mathematical aesthetics, when he argued that "we are no more entitled to suppose that when someone calls a proof 'elegant' he is appraising it on aesthetic grounds [...] than we would be to suppose he is appraising it on moral grounds when he calls it 'good'" (p. 136). Instead, philosophers who endorse non-literal views have argued that alleged aesthetic judgements in mathematics are, at heart, actually epistemic judgements (Rota, 1997; Todd, 2008, 2018).

The clearest example of a non-literal epistemic account comes from Rota's (1997) classic paper *The phenomenology of mathematical beauty*. For Rota, when mathematicians talk of 'beauty' they are actually referring to the epistemic notion of enlightenment. By 'enlightenment' Rota meant the type of understanding that successfully links a piece of mathematics into a wider context:

> Logical verification alone does not enable us to see the role that a statement plays within the theory. It does not explain how such a statement relates to other results, nor make us aware of the relevance of the statement in various contexts. In short, the mere logical truth of a statement does not enlighten us to the sense of the statement. (p. 181)

Rota went on to say that we have understood that a proof is enlightening "when we see how the theorem 'fits' in its place, how it sheds light around itself, like a Lichtung, a clear-

ing in the woods" (p. 182). For Rota, enlightenment is a property of both the mathematics itself but also the reader. Some routine proofs will not be enlightening no matter how well understood, but equally some readers will miss the enlightenment offered by a proof which is grasped by others. So, on Rota's account, beautiful pieces of mathematics are those that can generate understanding in the sense of deep connections between different mathematical ideas. But why might mathematicians make this strange decision to use aesthetic terms for an epistemic quality? According to Rota it is because enlightenment is a fuzzy concept, and "mathematicians universally dislike any concept admitting degrees" (p. 181). In sum, he argued that mathematical beauty is a 'trick' that "mathematicians have devised to avoid facing up to the messy phenomenon of enlightenment. [...] All talk of mathematical beauty is a copout from confronting the logic of enlightenment." (Rota, 1997, p. 182).

If non-literal theorists such as Rota (1997) are correct to suppose that aesthetic judgements in mathematics are actually disguised epistemic judgements about understanding, then this seems a serious blow to the project of incorporating aesthetic factors into mathematics education. In particular, if non-literal accounts are correct then pedagogy could focus on facilitating student understanding (of the sort that Rota would call 'enlightenment'), confident that the positive affective reactions sought by Dietiker, Singh, et al. (2023) and Sinclair (2001) would follow. If, instead, non-literal accounts are incorrect, and mathematical aesthetic judgements are genuinely aesthetic, then the attempt to incorporate aesthetics into educational settings seems a much more plausible project.

Given this, specifying the relationship between aesthetics and understanding would contribute to our knowledge about how aesthetics can be incorporated into educational practice. However, there has been limited existing empirical work on the relationship between aesthetics and understanding in mathematics. Perhaps the most relevant empirical study was Hayn-Leichsenring, Vartanian and Chatterjee's (2022) investigation of mathematicians' judgements of the beauty of mathematical equations. They asked twenty mathematicians and twenty laypeople to rate the beauty of 64 equations by ordering them from "least aesthetic" to "most aesthetic". Later, they asked the participants to state what factors had influenced their judgements. They found that the mathematically expert group often claimed to consider the meaning of equations, and that they did so more often than the laypeople; Hayn-Leichsenring et al. (2022) interpreted this as evidence that understanding the meaning of an equation contributes to perceiving it as having aesthetic appeal. However, some caution is required. Specifically, Hayn-Leichsenring et al. (2022) relied upon introspective self-reports to assess the factors that influenced their participants' aesthetic judgements. It is far from clear that their participants had access to the factors that underpinned their aesthetic appraisals. After all, if these factors were easily accessible, then it seems implausible that the nature of mathematical aesthetics would still be generating discussion and debate in the philosophy literature. Whether a similar finding would be observed with a more direct measure of understanding is an open question, one we set out to address in the study reported in this paper.

In sum, the relationship between understanding and aesthetics in mathematics is unclear. Educational theorists interested in incorporating aesthetics into the classroom typically believe that they are distinct concepts, whereas an influential line of work in the philosophy of mathematics literature argues that aesthetic judgements are not really

aesthetic, but rather are actually about understanding. Our goal in this paper was to explore this issue further. Specifically, we directly investigated the relationship between undergraduate students' understanding and aesthetic judgement of a mathematical proof. Since students' understanding can be measured in different ways, we also sought to better understand the relationships between these measures.

Background

Assessing proof appraisal

Our approach to assessing proof appraisals is based on the work of Inglis and Aberdein (2015), who investigated the number of dimensions on which mathematical proofs vary. They asked a large number of research-active mathematicians to think of a proof they had recently read or refereed, and then to assess the extent to which eighty adjectives accurately described it. For example, they asked if it would be accurate to describe the proof as 'intricate', as 'ambitious', as 'beautiful' and so on. These ratings were subjected to an exploratory factor analysis, which revealed a five-dimensional structure. Inglis and Aberdein (2015) named these factors *aesthetics*, *intricacy*, *utility*, *precision* and *non-use*. This latter factor was characterised by adjectives that seemed not to accurately characterise any of the proofs the participants thought of, and Inglis and Aberdein (2015) suggested that it was not a genuine dimension. They therefore concluded that proofs vary on four broad independent dimensions: aesthetics, intricacy, utility and precision.

In a follow-up study, Inglis and Aberdein (2016) developed a short 'personality' scale, which they used to assess how mathematicians appraised a particular proof. They took the adjectives that loaded particularly strongly onto the various factors from the earlier study, and across three studies asked participants to assess either proofs they'd recently read, or a specific proof, against these adjectives. For instance, aesthetic appraisals were assessed using the adjectives 'ingenious', 'inspired', 'profound' and 'striking' (all terms which align with the criteria offered by Sibley (1959) and Hungerland (1962)). Importantly, Inglis and Aberdein (2016) used four adjectives from each of the appraisal dimensions, including the non-use dimension, and asked participants about them in a random order. Including the non-use adjectives was intended to ensure that participants would be reluctant to simply select 'very accurate' for most or all adjectives. Across three studies, Inglis and Aberdein (2016) found consistently high internal reliability levels (as indexed by Cronbach's alphas) for the four substantive dimensions.

In a subsequent study, Inglis and Aberdein (2020) used the same short scale to test the extent to which including details about the publication venue of a specific proof influenced its perceived appraisal. They found that pure mathematicians regarded the proof as being higher on the aesthetic dimension if they were told it had been published in *Proofs from the BOOK*, a collection of purportedly beautiful proofs. As with the earlier studies, Inglis and Aberdein (2020) found consistently high Cronbach's alphas for the short 'personality' scale. We adopted the same approach in the current study, although we pay careful attention to the scale's internal reliability, as it has not previously been used with undergraduate students.

Assessing understanding of proofs

We highlight three approaches that have been used in previous research to assess students' understanding of particular proofs. The first is to ask students to report their own perceived level of understanding. For instance, in a study where students engaged with a "proof without words" of the Pythagorean Theorem, Marco (2021) asked students to reflect on their understanding, using two 10-point Likert items: "(1) "To what extent do you feel you understand the proof by now?" and (2) "What grade would you give yourself for the proof you submitted?"" (p. 233). Students gave their self-reported ratings before and after completing a short test about the proof, and there were significant (albeit small) changes in students' self-reported understanding. In a study of students' conceptions about proof, Healy and Hoyles (2000) also relied on self-reports, although the questions were less overtly about self-assessment of understanding. Students were presented with conjectures and various arguments in support of them; the students were asked to self-report "the argument that would be nearest to their own approaches and the argument they believed would receive the best mark from their teachers" (p. 399).

A second approach to assessing understanding of particular proofs is to use a proof-comprehension test. Mejía-Ramos et al. (2017) illustrated a process for developing and validating such tests, by producing proof-comprehension tests for three theorems that might typically be included in a "transition-to-proof" course (including the result that "The open interval $(0, 1)$ is uncountable"). They drew on a model for the assessment of proof comprehension (Mejía-Ramos et al., 2012) that isolated seven distinct aspects of comprehension, grouped into local aspects (such as understanding the meaning of terms used in the proof) and more holistic aspects (such as appreciation of the modular structure of a proof). This model has been used by others to develop tests for various other proofs (Cooley et al., 2024; Hodds et al., 2014; Roy et al., 2017).

The third approach to assessing understanding of a proof is to ask students to write a summary of the proof. Davies et al. (2020) developed this approach, making use of comparative judgement to score the students' summaries. The comparative judgement method is now widely-used in education, including for assessment of conceptual understanding (Jones & Davies, 2024; Jones et al., 2019). To score students' proof summaries using comparative judgement, a group of experts were shown pairs of students' proof summaries and asked to identify which of the pair was the better summary. By having the experts make a large number of these pairwise comparisons, a statistical model could then be used to produce scores for each student. Davies et al. (2020) motivated this approach (in part) by its simplicity relative to the resource-intensive approach of developing and validating a proof-comprehension test for each specific proof. To validate the approach, Davies et al. (2020) also asked students to complete one of the validated proof-comprehension tests developed by Mejía-Ramos et al. (2017). They found a "significant but modest" (p. 189) correlation of $r = .28$ between the scores from the proof summary and the proof-comprehension test. This result suggests that these different approaches may, in some cases at least, assess different aspects of understanding.

Method

We designed a survey to gather students' appraisals of a proof, along with three assessments of their understanding of the proof: a self-reported rating, a validated multiple-choice comprehension test, and a proof summary task. Our planned analysis was pre-registered prior to data collection (https://aspredicted.org/NHC_VT1). Ethical approval for this study was granted through the School of Mathematics at the University of Edinburgh.

Materials

Participants were first asked to read a proof that the open interval $(0,1)$ is uncountable, as shown in Figure 1. The same proof was used in the study conducted by Davies et al. (2020). The proof follows Cantor's famous diagonalisation argument and is widely seen as being an exemplar of beautiful mathematics (e.g., Dutilh Novaes, 2019). It seems clear that, if Rota's (1997) account were correct, then Cantor's proof, if understood, has the potential to offer enlightenment to readers. In particular, the proof offers both a novel idea that can be applied in other uncountability arguments, and a new perspective on the real numbers through the concept of (un)countability.

After reading the proof, participants were asked to complete the short appraisal scale developed by Inglis and Aberdein (2016). They were asked to "indicate how accurately each of the following words describes this proof", followed by the list of the adjectives shown in Table 1. The adjectives were shown in a fixed random order (as shown in the online materials at https://osf.io/9b7hn) together with a five-point Likert scale (very inaccurate, moderately inaccurate, neither inaccurate nor accurate, moderately accurate, very accurate).

Next, participants were asked to complete three assessments of their understanding of the proof:

1. Self-reported understanding: participants were asked to "describe how well you understood this proof" by selecting one of four options: "Did not understand at all", "Vaguely understood", "Understood well", or "Understood very well".

2. Proof-comprehension test: the 12-item test developed by Mejía-Ramos et al. (2017) was presented with the ordering of items and multiple-choice options randomised for each participant.

3. Proof summary task: in line with Davies et al. (2020), participants were asked to "Summarise the proof shown above in 40 words or fewer. Note: You are not being asked to reproduce the proof. The best responses will be those that succinctly communicate the most important aspects/ideas in the proof."

Each of the three tasks appeared on a different page in the online survey (with a copy of the proof at the top of each page), and participants were not able to return to previous pages.

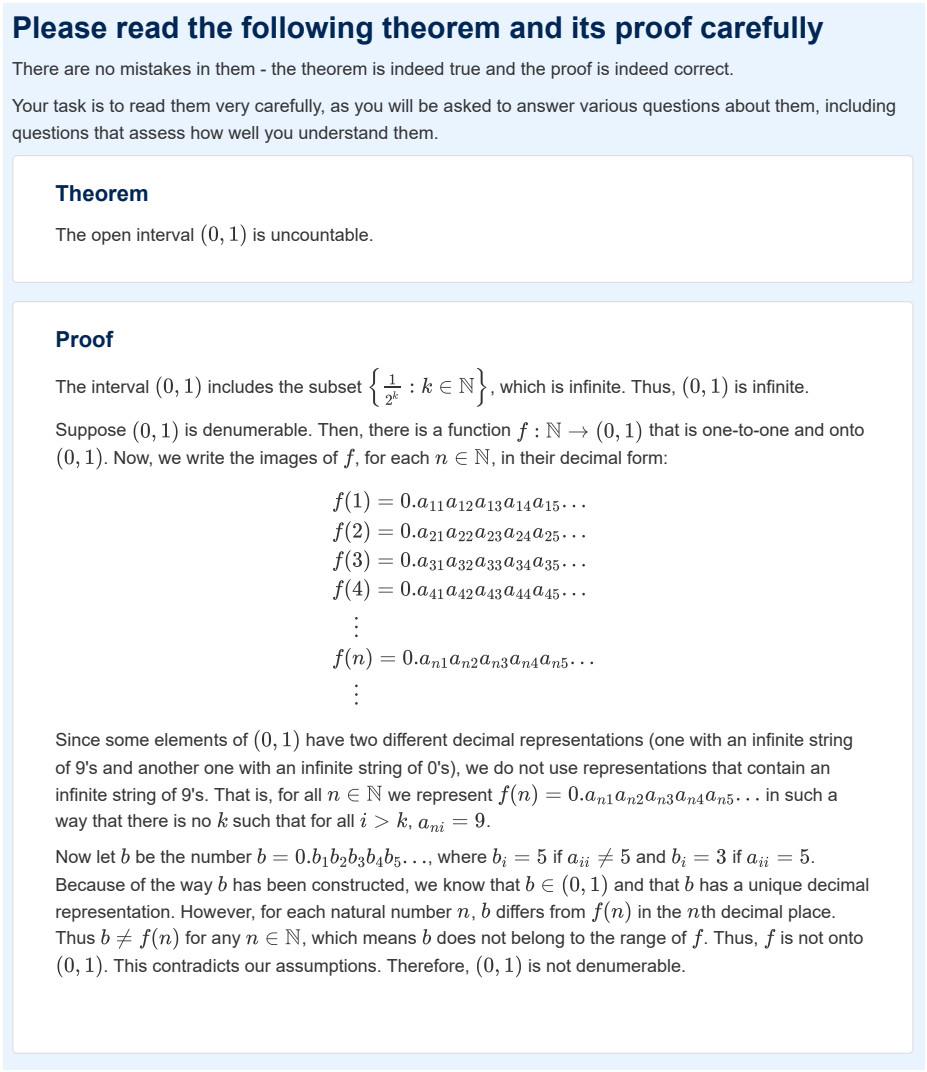

Figure 1

The proof shown to students in the survey.

Aesthetics	Intricacy	Precision	Utility	Non-use
Ingenious	Dense	Careful	Applicable	Careless
Inspired	Difficult	Meticulous	Informative	Crude
Profound	Intricate	Precise	Practical	Flimsy
Striking	Simple[a]	Rigorous	Useful	Shallow

Table 1

The 20 adjectives in the rating scale from Inglis and Aberdein (2016), grouped into the five dimensions.

[a] *Reverse scored*

Participants

We invited all students on a second-year pure mathematics course at a research-intensive UK university to complete the survey. The course provides an introduction to group theory and real analysis, with countable and uncountable sets studied for the first time in week 2. Students were invited to complete the survey online, in their own time, during week 2 of the semester (in January 2023). The proof was therefore likely to be novel but accessible to the students. Out of 454 students on the course, 166 consented to complete the survey. Following the approach used by Inglis and Aberdein (2020), we excluded participants who failed to respond to more than 5 of the 20 adjectives in the appraisal scale, which meant that 146 participants remained for our analysis.

Scoring the responses

For the appraisal scale, again following the approach used by Inglis and Aberdein (2020), we used the mean rating for each adjective to impute a small number of missing values (a total of 11 ratings from 10 participants, out of 2920 ratings in total).

For the proof-comprehension test, responses were scored according to the answer key provided by Mejía-Ramos et al. (2017). In particular, for the three multiple-selection items, responses were only regarded as correct if they matched the answer key exactly. There were 12 items on the test, each scored 0/1, giving a maximum possible score of 12.

For the proof summary task, we produced scores using the same comparative judgement method as Davies et al. (2020). Out of the 146 participants, 105 wrote a proof summary. In line with our pre-registration, participants who did not provide a summary still received a score; we included a single blank summary (presented to the judges as "[Blank]"), giving 106 summaries to be scored. We recruited 10 PhD students studying mathematics (from The University of Edinburgh and Loughborough University) to judge the proof summaries through an online interface. Judges were first asked to read the proof and keep it on hand for consultation during judging. Each judge then completed 106 pairwise comparisons[1], in each case deciding "Which is the best summary of the proof?". The pairs of summaries were selected randomly each time, subject to the constraint that each pair should be judged a similar number of times. The judges took between 22 and 37 minutes to complete their judgements, with an overall median time per judgement of 14.1 seconds. We used the Bradley-Terry model to derive scores for each summary from the 1060 pairwise comparisons. The Scale Separation Reliability (SSR) was .87 and the Split-Halves Reliability (SHR) was .74 (based on the median correlation from 100 split-halves), both of which suggested that there was a good level of between-judge consistency in judgements.

Analysis

To investigate whether there was a relationship between the measures of understanding and aesthetic appraisals, we used a linear regression (Gelman et al., 2020). This

[1]We asked each judge to complete fewer judgements that we had planned in our pre-registration, where we said judges would complete 150 judgements. This was so that we could have at least 10 judges making judgements, while satisfying the overall goal of having the total number of comparisons being 10 times the number of participants.

statistical method examines how several independent variables (in our case, the different measures of understanding) affect one outcome (aesthetic appraisal). For our secondary analysis, regarding relationships between the measures of understanding, we computed Pearson correlations between pairs of measures.

Results

Before proceeding to our pre-registered primary and secondary analyses, we first present exploratory data analyses and descriptive statistics for the appraisal scales and the three assessments of understanding. All of the data and analysis code is available at https://osf.io/7ntzj.

Appraisals and assessments of understanding

The appraisal scale had good internal reliability, with a Cronbach's alpha of .79 for the aesthetics scale. The other scales had lower Cronbach's alpha values, though only intricacy, at .59, was below the traditional 'acceptable' level of .70. Aesthetic appraisals spanned almost the full range of possible scores, with a minimum of 5 and maximum of 20. The other appraisals ranged from 6 to 20.

The scores on each of the assessments of understanding are shown in Figure 2, together with their relationship with aesthetic appraisal scores.[2] The self-reported understanding scores spanned the full range. There were only five students in the "Did not understand at all" group, and most students (79 out of 146) opted for "Vaguely understood".

The proof-comprehension test scores spanned the full range (0-12), with a mean score of 5.76 and standard deviation of 2.98. The test had an acceptable level of internal consistency, with Cronbach's $\alpha = .75$ and a principal component analysis supporting a unidimensional structure (KMO $= 0.75$, Bartlett's test of sphericity having $p < 0.001$, and a 1-factor solution explaining 28% of the variance). This contrasts with the finding of Davies et al. (2020) that, in their context, the test had an unacceptably low internal consistency (Cronbach's $\alpha = .53$).

Scores on the proof summary task ranged from -6.67 to 3.87, with a mean of -1.27 and standard deviation of 2.62. The 41 students who did not write a proof summary were all assigned the same score (-4.31) from the single "[Blank]" entry shown to judges[3]. Excluding these students, the scores had a mean of 0.04 and standard deviation of 2.00. For the results that follow, we included students whose proof summaries were blank (in accordance with our pre-registration), however the results are substantially the same when removing the students with blank summaries.

Primary analysis: relationship between understanding and appraisals

Our primary pre-registered analysis focused on the overall relationship between understanding and appraisals. The analysis consisted of four separate linear regressions,

[2] A version of this figure showing all of the appraisals is available in the online materials.

[3] Note that four non-empty student responses received a lower score than the "[Blank]" entry; each of these was a short expression of a lack of understanding, e.g. "Unsure" or "did not understand".

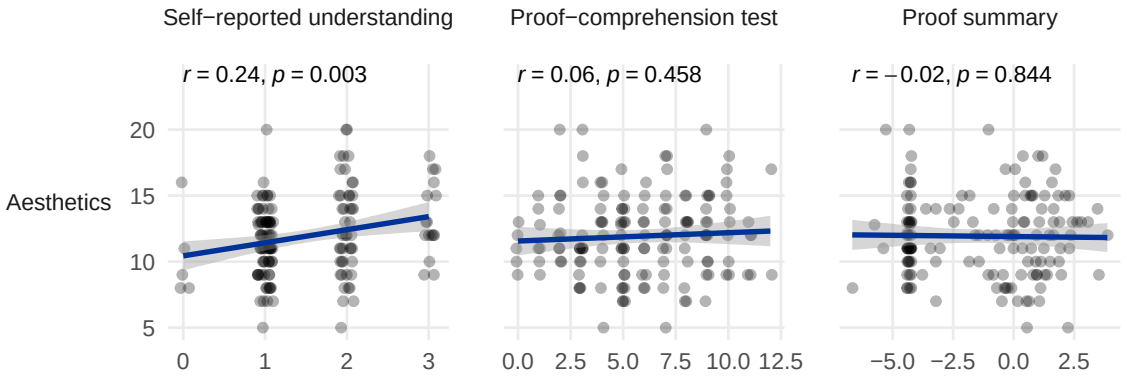

Figure 2

Scores on the three assessments of understanding (self-report, proof comprehension test, and the proof summary task) compared with aesthetic appraisals. Note that a small amount of horizontal jitter has been added to the data points to aid visibility.

where in each case the dependent variable was one of the appraisal scales, and the predictors were the three assessments of understanding. We first checked for possible collinearity in the predictors by computing variance inflation factors (VIFs). The VIFs for all three predictors were below the pre-registered threshold of 5 (Self-report: 1.28, PC score: 1.74, PS score: 1.58), indicating an acceptable level of collinearity for us to proceed with interpreting the regression coefficients. The results of the four linear regressions are shown in Table 2.

Overall, the linear regressions showed that a relationship exists between understanding and appraisals, since at least one of the models was significant at the Bonferroni-corrected alpha level of $.05/4 = .0125$ (in fact three of the four were). To consider each model separately, we used the standard alpha level of .05 (García-Pérez, 2023). Appraisals of Intricacy, Precision and Utility were significantly predicted by the assessments of understanding, with the predictors explaining between 11% and 17% of the variance in appraisals. In each model, surprisingly, self-reported understanding was the only significant predictor.

For the crucial aesthetic appraisals, the regression model was significant ($p = .016$). The assessments of understanding explained only 7% of the variance in aesthetic appraisals, which was considerably lower than the other appraisals. Again, surprisingly, the only assessment of understanding that significantly predicted aesthetic appraisal was self-reported understanding.

Given our focus on students' aesthetic appraisals, we carried out further exploratory analyses, with each of the three assessments of understanding separately as predictors. The results, as shown in Figure 2, are consistent with the results from the overall regression model. In particular, aesthetic appraisals were significantly correlated[4]

[4]Here we have used Pearson correlation coefficients throughout; Spearman correlation coefficients may be more appropriate for the self-reported understanding measure in particular, however the results were very

	Model fit		Coefficients				
Model	R^2	p	Term	Estimate	SE	t	p
Aesthetics	.07	.016	Intercept	9.99	0.82	12.13	$< .001$
			Self-report	1.15	0.38	3.07	.003
			PC score	0.01	0.11	0.05	.960
			PS score	-0.14	0.12	-1.17	.245
Intricacy	.11	.001	Intercept	15.09	0.74	20.26	$< .001$
			Self-report	-1.30	0.34	-3.82	$< .001$
			PC score	0.02	0.10	0.17	.868
			PS score	0.01	0.11	0.07	.942
Precision	.16	$< .001$	Intercept	12.12	0.78	15.54	$< .001$
			Self-report	1.44	0.36	4.05	$< .001$
			PC score	0.03	0.10	0.32	.750
			PS score	0.07	0.11	0.65	.514
Utility	.17	$< .001$	Intercept	11.29	0.72	15.59	$< .001$
			Self-report	1.58	0.33	4.78	$< .001$
			PC score	0.04	0.10	0.40	.688
			PS score	-0.06	0.10	-0.61	.541

Table 2

Results of the four linear regressions.

with self-reported understanding ($r(144) = .24$, $p = .003$, 95% CI $[.08, .39]$) but not with scores on the proof-comprehension test ($r(144) = .06$, $p = .458$, 95% CI $[-.10, .22]$) or the proof summary task ($r(144) = -.02$, $p = .884$, 95% CI $[-.18, .15]$). While the correlation between self-reported understanding and aesthetic appraisals was significant, we note that the correlation of $r = .24$ was relatively weak: in particular, this means that self-reported understanding, although related to aesthetic appraisal, is clearly a distinct construct.

In sum, we found that students' aesthetic appraisals were associated with the level of their understanding, but only as measured by our self-report scale. Performance on neither of our more objective performance-based assessments of understanding—the proof comprehension test and the proof summary task—were predictive of the extent to which students found the proof to have aesthetic properties.

Secondary analysis: relationships between the assessments of understanding

As a secondary pre-registered analysis, we investigated how the different assessments of understanding compared with one another, to replicate and extend the analysis of Davies et al. (2020). A further motivation for this analysis comes from the surprising finding that self-reported understanding was the only significant predictor of appraisals:

similar in each case.

given our interest in the relationship between understanding and appraisals, we wanted to better understand the three different assessments of understanding.

First, we computed the correlation between scores for the proof-comprehension test and proof summary task. We observed a strong correlation ($r(164) = .62$, $p < .001$, 95% CI $[.52, .71]$). For comparison, Davies et al. (2020, p. 189) found a "significant but modest" correlation of $r = .28$ between scores on the proof summary task and a seven-item subset of the proof-comprehension test.

Second, we investigated whether students' self-reported understanding was predictive of their scores on the two performance-based assessments of understanding. The distribution of scores for each level of self-reported understanding (shown in Figure 3) followed the expected pattern, with a visible trend where higher self-reported understanding was associated with higher scores on the proof-comprehension test and proof summary task. These visible trends were confirmed by correlations,[5] with self-reported understanding correlating moderately with proof-comprehension test scores ($r(149) = .48, p < .001, 95\%$ CI $[.34, .59]$) and with proof summary scores ($r(149) = .37, p < .001, 95\%$ CI $[.22, .50]$).

Finally, we investigated whether the scores on the assessments were predictive of overall course results. For self-reported understanding, we had scores and course results for 138 students. There was a moderate correlation between self-reported understanding of the proof and overall course results, $r(136) = .42$, $p < .001$, 95% CI $[.27, .55]$. For the proof-comprehension test, we had scores and course results for 150 students. There was a moderate correlation between the proof-comprehension test scores and course results, $r(148) = .57$, $p < .001$, 95% CI $[.45, .67]$. For the proof summary task, we had scores and course results for 97 students. There was a weaker (albeit still significant) correlation between the scores on the proof summary task and course results, $r(95) = .26$, $p = .009$, 95% CI $[.07, .44]$. The pattern of results was similar when considering only the subset of students with complete data.

In sum, it is clear that the three measures of understanding we used were all related, albeit not so strongly related that they could be said to measure the same construct. It seems that self-reported understanding, understanding as assessed by proof comprehension tests, and understanding as assessed by proof summary tasks are all inter-related but distinct constructs. Given that our findings conflict with those of Davies et al. (2020), further research which aims to uncover exactly when (and why) different measures of proof understanding relate, and when they do not relate, would be worthwhile.

[5] Our pre-registered analysis plan was a Bayesian ANOVA, which gave similar results. We used two Bayesian ANOVA model comparisons (Bergh et al., 2020). In each case, the comparison was between a model with one predictor (self-reported understanding) and the null model. We used the default prior width of $r = .5$. For the models with the proof-comprehension test as the dependent variable, the Bayes factor was $BF_{10} = 940778$ relative to the null model, which indicates strong evidence that the proof-comprehension test scores depend on the levels of self-reported understanding. Similarly, for the models with the proof summary task as the dependent variable, the Bayes factor $BF_{10} = 485$ indicates that the scores depended on self-reported understanding.

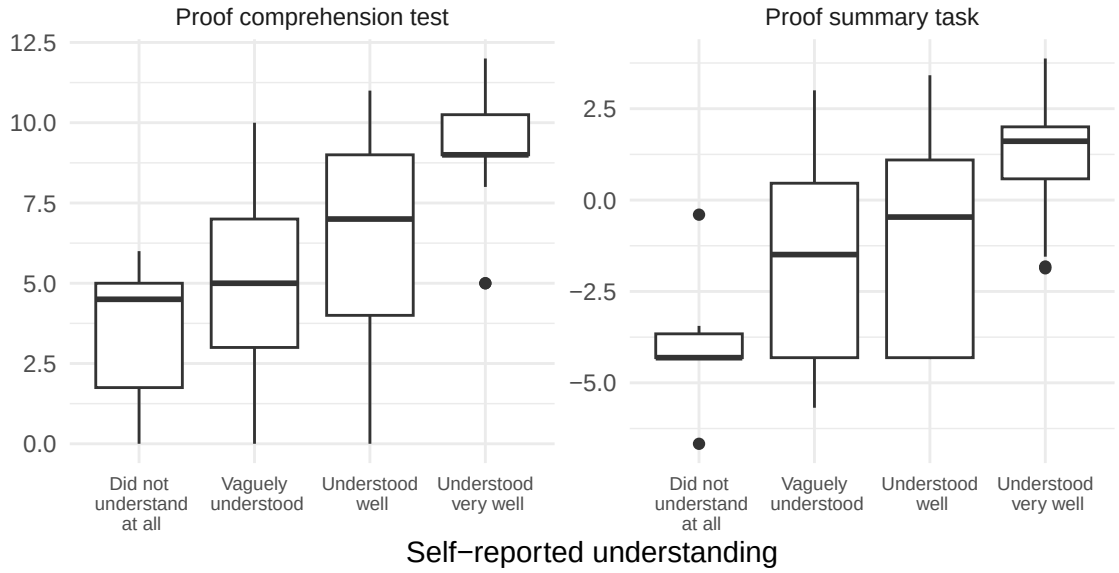

Figure 3

Scores on the proof comprehension test and the proof summary task, summarised by boxplots for each level of self-reported understanding.

Discussion

Summary of Main Findings

We explored the relationship between how well undergraduate students understand a mathematical proof and the extent to which they perceive that it has aesthetic value. Non-literal theorists like Rota (1997) have proposed that aesthetic appraisals in mathematics are not genuinely aesthetic, but instead reflect epistemic factors. Specifically, Rota suggested that aesthetic appraisals track enlightenment, a type of understanding. If such a non-literal account were correct, when students read a mathematical proof we would expect there to be a strong and consistent positive relationship between measures of the extent to which they regard it as having aesthetic value and the extent to which they understand it (at least, we would expect this to be the case for all non-routine proofs that have the potential to offer enlightenment in the sense of Rota).

To this end, we asked a large group of mathematics undergraduates to read a proof, to appraise it on several dimensions including aesthetics, and then to complete several measures designed to assess the extent to which they understood it. In contrast to Rota's (1997) non-literal account of aesthetics we found that neither of our two performance-based measures of understanding significantly predicted students' aesthetic appraisals. However, we found that students' self-reported levels of understanding were significantly, albeit moderately, associated with their aesthetic appraisals. If Rota's account were correct, we would have expected that all three measures would have been related to aesthetic appraisals.

Note that this result cannot be attributed to a general lack of association between

the two performance-based understanding measures and measures of subjective proof appraisal per se. Both our performance-based measures of understanding were correlated with students' subjective appraisals of precision (students who understood the proof better regarded it as being more precise) and non-use (students who understood the proof better were less likely to describe the proof with negative words like 'careless' and 'shallow'). In other words, performance-based understanding measures predicted certain subjective proof appraisals, just not aesthetic appraisals.

Limitations

Our key finding, that performance-based measures of understanding do not correlate with aesthetic appraisals, of course relies upon the validity of our understanding measures. In particular, do either of our measures adequately capture Rota's (1997) notion of enlightenment? Recall that Rota suggested that enlightenment is achieved when a piece of mathematics is seen as shedding light on some wider context.

Our measure of proof comprehension (Mejía-Ramos et al., 2017) was based on a model of proof comprehension that contains seven different dimensions (Mejía-Ramos et al., 2012). Some of these certainly would not relate to Rota's notion of enlightenment. For instance, one of our questions, concerning the *meaning of terms and statements* dimension, asked participants what the definition of uncountable was. However, other dimensions of the model clearly do. For instance, other dimensions focused on transferring the general ideas and methods of the proof to new contexts, and on applying the ideas of the proof to specific novel examples. We identified six items on the test that were aligned to these dimensions. For instance, one of these items asked about applying the ideas of the proof to determine whether the set of infinite binary sequences is uncountable, which fits well with the sort of outward-looking understanding that Rota described. However, participants' scores on this six-item subset were also not significantly correlated with their aesthetic appraisals ($r(144) = .13$, $p = .107$, 95% CI $[-.03, .29]$). This is perhaps not surprising, given that our principal components analysis suggested that all the proof comprehension test items loaded onto one factor, indicating that our data did not support a strong distinction between different types of proof comprehension items. In sum, if Rota's account were correct, it is surprising that we were not able to detect a relationship between scores on our proof comprehension test and aesthetic appraisals.

Of course, it could be that the dimensions of proof comprehension identified by Mejía-Ramos et al. (2012) are not sufficient to describe the sort of understanding that Rota (1997) described as enlightenment. While Mejía-Ramos et al. (2012) used a rigorous approach to identify the dimensions, their focus on assessing comprehension of a particular proof may have meant that issues of how that proof "sheds light around itself" (Rota, 1997, p. 182) were not sufficiently prominent to be included. This would leave open the question of what such measures of enlightenment might look like.

We also used experts' judgements of students' proof summaries as a measure of the students' understanding. If Rota (1997) was correct to assert that enlightenment is fundamental to mathematical practice, then we would have expected the judges who made the comparative judgements would have favoured proof summaries which exhibited some evidence that the author had appreciated how the theorem fits into the wider

context. But we found no relationship between proof summary scores and aesthetic appraisals.

A further limitation of our results is that they derive from a single proof and from students in one particular context. However, much like a single counterexample is all that is needed to disprove a general claim, our results are sufficient to show that Rota's (1997) non-literal account does not apply to all proofs. Of course, replication of our study with different materials and in different contexts would be welcome.

Theoretical and Educational Implications

If our performance-based measures of understanding adequately measured the extent to which students understood the proof (an issue to which we return below), then our findings demonstrate that, contrary to the suggestion of epistemic non-literal theorists, aesthetic appraisals cannot straightforwardly be disguised assessments of the extent to which a proof has been understood. So what exactly are aesthetic appraisals of the type made by the participants in our study? There are at least three broad possibilities.

One possibility is that Rota's (1997) account is correct, but that it applies only to people who accurately assess the extent to which they have understood the proof. Perhaps some people read the proof, gain an illusory sense of understanding (i.e., think that they have understood when in fact they have not) and it is this illusory sense which produces the positive aesthetic appraisal. We think that this is unlikely. Even among those participants who rated their understanding of the proof as high, there were still large individual differences in the extent to which they thought it had aesthetic properties. This can be seen in the distribution of aesthetic judgements in Figure 2; among participants with the highest self-reported understanding, aesthetic appraisal scores ranged from 9 to 18 on the 4 to 20 scale. In other words, even if we restricted our focus to those participants who thought they understood the proof, there were individual differences in aesthetic judgements. If mathematical aesthetics are simply disguised, albeit possibly illusory, epistemic judgements, it is hard to see where these large individual differences came from.

A second possibility is that, contrary to the assertions of non-literal theorists, the aesthetic appraisals made by participants in our study were genuinely aesthetic. But this possibility raises a further question: if participants' aesthetic appraisals were genuinely aesthetic, and not disguised epistemic judgements, then why did they correlate (albeit weakly) with participants' self-reported levels of understanding? Recall that Sinclair (2001) noted that aesthetically rich experiences involve students experiencing "pleasure and pride" (p. 26). Perhaps participants who regarded the proof as having aesthetic value experienced a positive affective response, which led to them responding more positively on the self-report question about understanding (and, vice versa, perhaps those who regarded the proof as having negative aesthetic value exhibited a negative affective response, which led them to respond more negatively on the self-report item). In other words, when asked about the extent to which they understood the proof, perhaps participants responded at least partially based on their more general feelings about the proof, rather than only on their actual levels of understanding. This account is consistent with the more general finding that when asked to self-assess the quality of their learning, students are influenced by a whole host of factors (Emeny et al., 2021).

If this account were correct, then it implies that it is possible to appreciate the aesthetics of a mathematical proof without either understanding it or believing that it is understood. If this is right then, when considering how to present mathematical proofs which have aesthetic value, thought must be given to two quite distinct goals: how student understanding can be promoted, but also how students can be helped to appreciate the proof's aesthetic properties. We return to this issue below.

Assuming that participants' aesthetic judgements of this proof were genuinely aesthetic, then it seems that the proof provoked quite some disagreement about the extent to which it had aesthetic value. Recall that our aesthetic scale ranged from a possible 4 (the lowest possible aesthetic value) to 20 (the highest possible aesthetic value). In fact some participants rated the proof as low as 5, whereas others rated it as high as 20 (median 12, inter-quartile range 4). What accounts for these individual differences? Before speculating we note that the history of this specific proof, Cantor's diagonalisation argument, shows similar disagreements. The initial reaction to Cantor's diagonalisation argument upon publication was extremely mixed. For instance, his contemporary colleague Kronecker reacted very negatively. Schoenflies (1927), in an analysis of Cantor's private letters, wrote that Kronecker's attitude towards Cantor's work was so hostile that it "must have created the impression that Cantor, in his capacity as researcher and teacher, was a corrupter of youth" (p. 2).[6] But, despite this difficult start, the proof is now widely regarded as an exemplar of beauty. Dutilh Novaes (2019) described how it is "often viewed as mesmerizing" (p. 69) and that attributing beauty to it is the "instinctive response" to encountering it (p. 71). Despite this, she also noted that the extent to which these reactions are shared is an open empirical question, and remarked that although a majority of audience members to whom she had presented the proof had shared her intuition that the proof is beautiful, a minority disagreed (p. 86).[7] The distribution of aesthetic appraisals shown in Figure 2 provides an answer to Dutilh Novaes's 'empirical question': while a minority of the undergraduate students in our sample regarded the proof as having high aesthetic value, the majority rated it in the middle of the scale, with some reacting negatively. While it is clear that levels of understanding do not explain these large individual differences in aesthetic appraisal—the R^2 value for our regression was a meagre 7%—it is far from clear what factors do. Investigating how individual differences predict aesthetic appraisal in mathematics would be a worthwhile goal for future researchers to explore (cf. Inglis & Aberdein, 2020).

A final way of accounting for our results would be to deny that the aesthetic appraisals made by our participants were genuinely aesthetic and instead to endorse a non-literal, but also non-epistemic, account of aesthetics in mathematics. For instance, perhaps feelings of aesthetics and self-reported understanding are both driven by non-aesthetic affective responses to the proof, possibly attitudes towards the mathematical domain in which the proof is situated. In other words, maybe those students who have a positive attitude towards set theory are more likely to rate Cantor's proof highly on both aesthetics and self-reported understanding scales, thus creating the correlation we

[6] Translated from the German "die Kroneckersche Einstellung den Eindruck hervorbringen musste, als sei Cantor in seiner Eigenschaft als Forscher und Lehrer ein Verderber der Jugend".

[7] Indeed, Hodges (1998) described reviewing several attempted refutations of Cantor's proof, presumably submitted by people who regarded it as flawed rather than beautiful.

observed. Again, future research which investigates how individual differences in factors such as attitudes towards (subdomains of) mathematics relate to aesthetic appraisals would be worthwhile.

Under the assumption that our performance-based measures capture the kind of understanding that we would like students to develop when they read proofs, then the educational implications of our results are clear. Successfully helping students to understand mathematical proofs will not necessarily lead to them appreciating the aesthetic properties of those proofs. If we believe that mathematical aesthetics are important, and if we care about helping students to develop skills of aesthetic appreciation, then we need to develop appropriate pedagogical strategies to achieve this. Although there are many suggestions of how to do this at the lesson or curricular level (Dietiker, 2013; Dietiker, Riling, et al., 2023; Dietiker, Singh, et al., 2023; Sinclair, 2001), much less has been written about how to present proofs to students in a manner that facilitates aesthetic appraisal. One obvious avenue for exploring this question concerns the extent to which the narrative characteristics identified by Dietiker and colleagues, in the context of lesson or curriculum design, could be incorporated into mathematical proofs. If we are to fully understand the nature of aesthetics in mathematics, and how it can contribute to improving teaching and learning in undergraduate mathematics, work of this sort will be necessary.

References

Ball, D. L., & Bass, H. (2000). Making believe: The collective construction of public mathematical knowledge in the elementary classroom. *Teachers College Record*, *102*(7), 193–224. https://doi.org/10.1177/016146810010200707

Bergh, D. v. d., Doorn, J. v., Marsman, M., Draws, T., Kesteren, E.-J. v., Derks, K., Dablander, F., Gronau, Q. F., Kucharský, Š., Gupta, A. R. K. N., Sarafoglou, A., Voelkel, J. G., Stefan, A., Ly, A., Hinne, M., Matzke, D., & Wagenmakers, E.-J. (2020). A tutorial on conducting and interpreting a Bayesian ANOVA in JASP. *L'Année psychologique*, *120*(1), 73–96. Retrieved September 5, 2024, from https://shs.cairn.info/journal-l-annee-psychologique-2020-1-page-73?lang=en&tab=resume

Borasi, R., & Brown, S. I. (1985). A "Novel" approach to texts. *For the Learning of Mathematics*, *5*(1), 21–23. Retrieved September 30, 2024, from https://www.jstor.org/stable/40247872

Cooley, L., Dorfmeister, J., Miller, V., Duncan, B., Littmann, F., Martin, W., Vidakovic, D., & Yao, Y. (2024). The PRIUM qualitative framework for assessment of proof comprehension: A result of collaboration among mathematicians and mathematics educators. *ZDM Mathematics Education*, *56*(7), 1553–1566. https://doi.org/10.1007/s11858-024-01628-1

Davies, B., Alcock, L., & Jones, I. (2020). Comparative judgement, proof summaries and proof comprehension. *Educational Studies in Mathematics*, *105*(2), 181–197. https://doi.org/10.1007/s10649-020-09984-x

Dietiker, L. (2013). Mathematical texts as narrative: Rethinking curriculum. *For the Learning of Mathematics*, *33*(3), 14–19. Retrieved September 30, 2024, from https://www.jstor.org/stable/43894855

Dietiker, L. (2015a). Mathematical story: A metaphor for mathematics curriculum. *Educational Studies in Mathematics*, *90*(3), 285–302. https://doi.org/10.1007/s10649-015-9627-x

Dietiker, L. (2015b). What mathematics education can learn from art: The assumptions, values, and vision of mathematics education. *Journal of Education*, *195*(1), 1–10. https://doi.org/10.1177/002205741519500102

Dietiker, L. (2016). Generating student interest with mathematical stories. *The Mathematics Teacher*, *110*(4), 304–308. https://doi.org/10.5951/mathteacher.110.4.0304

Dietiker, L., Riling, M., Singh, R., I. Nieves, H., & Barno, E. (2023). The aesthetic effects of a new lesson design approach: Mathematical stories. *The Journal of Educational Research*, *116*(1), 33–47. https://doi.org/10.1080/00220671.2023.2182264

Dietiker, L., Singh, R., Riling, M., Nieves, H. I., & Barno, E. (2023). Narrative characteristics of captivating secondary mathematics lessons. *Educational Studies in Mathematics*, *112*(3), 481–504. https://doi.org/10.1007/s10649-022-10184-y

Dutilh Novaes, C. (2019). The Beauty (?) of Mathematical Proofs. In A. Aberdein & M. Inglis (Eds.), *Advances in Experimental Philosophy of Logic and Mathematics* (pp. 63–93). Bloomsbury Academic.

Emeny, W. G., Hartwig, M. K., & Rohrer, D. (2021). Spaced mathematics practice improves test scores and reduces overconfidence. *Applied Cognitive Psychology*, *35*(4), 1082–1089. https://doi.org/10.1002/acp.3814

Gadanidis, G., & Hoogland, C. (2003). The aesthetic in mathematics as story. *Canadian Journal of Science, Mathematics and Technology Education*, *3*(4), 487–498. https://doi.org/10.1080/14926150309556584

García-Pérez, M. A. (2023). Use and misuse of corrections for multiple testing. *Methods in Psychology*, *8*, 100120. https://doi.org/10.1016/j.metip.2023.100120

Gelman, A., Hill, J., & Vehtari, A. (2020, July). *Regression and Other Stories*. Cambridge University Press. https://doi.org/10.1017/9781139161879

Harré, R. (1958). Quasi-aesthetic appraisals. *Philosophy*, *33*, 132–137. https://doi.org/10.1017/S0031819100038249

Hayn-Leichsenring, G. U., Vartanian, O., & Chatterjee, A. (2022). The role of expertise in the aesthetic evaluation of mathematical equations. *Psychological Research*, *86*(5), 1655–1664. https://doi.org/10.1007/s00426-021-01592-5

Healy, L., & Hoyles, C. (2000). A study of proof conceptions in algebra. *Journal for Research in Mathematics Education*, *31*(4), 396–428. https://doi.org/10.2307/749651

Hodds, M., Alcock, L., & Inglis, M. (2014). Self-explanation training improves proof comprehension. *Journal for Research in Mathematics Education*, *45*(1), 62. https://doi.org/10.5951/jresematheduc.45.1.0062

Hodges, W. (1998). An editor recalls some hopeless papers. *The Bulletin of Symbolic Logic*, *4*(1), 1–16. https://doi.org/10.2307/421003

Holden, H., & Piene, R. (Eds.). (2010). *The Abel prize: 2003–2007 the first five years*. Springer. https://doi.org/10.1007/978-3-642-01373-7

Hungerland, I. C. (1962). The logic of aesthetic concepts. *Proceedings and Addresses of the American Philosophical Association*, *36*, 43–66.

Inglis, M., & Aberdein, A. (2015). Beauty is not simplicity: An analysis of mathematicians' proof appraisals. *Philosophia Mathematica*, *23*(1), 87–109. https://doi.org/10.1093/philmat/nku014

Inglis, M., & Aberdein, A. (2016). Diversity in proof appraisal. In *Mathematical Cultures: The London Meetings 2012-2014* (pp. 163–179). Springer. https://doi.org/10.1007/978-3-319-28582-5_10

Inglis, M., & Aberdein, A. (2020). Are aesthetic judgements purely aesthetic? Testing the social conformity account. *ZDM Mathematics Education*, *52*(6), 1127–1136. https://doi.org/10.1007/s11858-020-01156-8

Jones, I., Bisson, M., Gilmore, C., & Inglis, M. (2019). Measuring conceptual understanding in randomised controlled trials: Can comparative judgement help? *British Educational Research Journal*, *45*(3), 662–680. https://doi.org/10.1002/berj.3519

Jones, I., & Davies, B. (2024). Comparative judgement in education research. *International Journal of Research & Method in Education*, *47*(2), 170–181. https://doi.org/10.1080/1743727X.2023.2242273

Koichu, B., Katz, E., & Berman, A. (2017). Stimulating student aesthetic response to mathematical problems by means of manipulating the extent of surprise. *The Journal of Mathematical Behavior*, *46*, 42–57. https://doi.org/10.1016/j.jmathb.2017.02.005

Marco, N. (2021). The effects of a proof comprehension test on comprehending proofs without words. In M. Inprasitha, N. Changsri, & N. Boonsena (Eds.), *Proceedings of the 44th Conference of the International Group for the Psychology of Mathematics Education* (pp. 222–229, Vol. 3). PME. https://www.igpme.org/wp-content/uploads/2022/04/Volume-3_final.pdf

Mejía-Ramos, J. P., Fuller, E., Weber, K., Rhoads, K., & Samkoff, A. (2012). An assessment model for proof comprehension in undergraduate mathematics. *Educational Studies in Mathematics*, *79*(1), 3–18. https://doi.org/10.1007/s10649-011-9349-7

Mejía-Ramos, J. P., Lew, K., de la Torre, J., & Weber, K. (2017). Developing and validating proof comprehension tests in undergraduate mathematics. *Research in Mathematics Education*, *19*(2), 130–146. https://doi.org/10.1080/14794802.2017.1325776

Papert, S. (1980). *Mindstorms: Children, computers and powerful ideas*. Harvester.

Poincaré, H. (1914). *Science and method*. Thomas Nelson.

Reid, C. (1986). *Hilbert Courant*. Springer.

Richman, A. S., Dietiker, L., & Riling, M. (2019). The plot thickens: The aesthetic dimensions of a captivating mathematics lesson. *The Journal of Mathematical Behavior*, *54*, 100671. https://doi.org/10.1016/j.jmathb.2018.08.005

Rota, G.-C. (1997). The phenomenology of mathematical beauty. *Synthese*, *111*(2), 171–182. https://doi.org/10.1023/A:1004930722234

Roy, S., Inglis, M., & Alcock, L. (2017). Multimedia resources designed to support learning from written proofs: An eye-movement study. *Educational Studies in Mathematics*, *96*(2), 249–266. https://doi.org/10.1007/s10649-017-9754-7

Sangwin, C. J., & Kinnear, G. (2024). Investigating insight and rigour as separate constructs in mathematical proof. *Research in Mathematics Education*, 1–29. https://doi.org/10.1080/14794802.2024.2379301

Schoenfeld, A. H. (2016). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics (reprint). *Journal of Education*, *196*(2), 1–38. https://doi.org/10.1177/002205741619600202

Schoenflies, A. (1927). Die Krisis in Cantor's mathematischem Schaffen. *Acta Mathematica*, *50*, 1–23. https://doi.org/10.1007/BF02421320

Sibley, F. (1959). Aesthetic concepts. *The Philosophical Review*, *68*, 421–450.

Sinclair, N. (2001). The aesthetic "Is" relevant. *For the Learning of Mathematics*, *21*(1), 25–32. Retrieved September 30, 2024, from https://www.jstor.org/stable/40248343

Stylianou, D. A. (2002). On the interaction of visualization and analysis: The negotiation of a visual representation in expert problem solving. *The Journal of Mathematical Behavior*, *21*(3), 303–317. https://doi.org/10.1016/S0732-3123(02)00131-1

Todd, C. (2008). Unmasking the truth beneath the beauty: Why the supposed aesthetic judgements made in science may not be aesthetic at all. *International Studies in the Philosophy of Science*, *22*(1), 61–79. https://doi.org/10.1080/02698590802280910

Todd, C. (2018). Fitting feelings and elegant proofs: On the psychology of aesthetic evaluation in mathematics. *Philosophia Mathematica*, *26*(2), 211–233. https://doi.org/10.1093/philmat/nkx007

Weber, K., Inglis, M., & Mejía-Ramos, J. P. (2014). How mathematicians obtain conviction: Implications for mathematics instruction and research on epistemic cognition. *Educational Psychologist*, *49*(1), 36–58. https://doi.org/10.1080/00461520.2013.865527

Wong, D. (2007). Beyond control and rationality: Dewey, aesthetics, motivation, and educative experiences. *Teachers College Record*, *109*(1), 192–220. https://doi.org/10.1177/016146810710900101

Zazkis, R., Liljedahl, P., & Sinclair, N. (2009). Lesson plays: Planning teaching versus teaching planning. *For the Learning of Mathematics*, *29*(1), 40–47. Retrieved September 30, 2024, from https://www.jstor.org/stable/40248639