Are Aesthetic Judgements Purely Aesthetic? Testing the Social Conformity Account

Matthew Inglis

Loughborough University

Andrew Aberdein

Florida Institute of Technology

Address for correspondence:

Matthew Inglis, Centre for Mathematical Cognition, Loughborough University, Loughborough, Leicestershire. LE11 3TU. United Kingdom.

Email: m.j.inglis@lboro.ac.uk.

Author note:

This work was first presented at the 15th Congress of Logic Methodology and Philosophy of Science. We are grateful to Dirk Schlimm for organising the "Mathematical beauty: A challenge for empirically informed philosophy of mathematics" symposium. Data and analysis code associated with this manuscript are available at https://doi.org/10.17028/rd.lboro.c.4679399

Abstract

Many of the methods commonly used to research mathematical practice, such as analyses of historical episodes or individual cases, are particularly well-suited to *generating* causal hypotheses, but less well-suited to *testing* causal hypotheses. In this paper we reflect on the contribution that the so-called hypothetico-deductive method, with a particular focus on experimental studies, can make to our understanding of mathematical practice. By way of illustration, we report an experiment that investigated how mathematicians attribute aesthetic properties to mathematical proofs. We demonstrate that perceptions of the aesthetic properties of mathematical proofs are, in some cases at least, subject to social influence. Specifically, we show that mathematicians' aesthetic judgements tend to conform to the judgements made by others. Pedagogical implications are discussed.

Key words: mathematical aesthetics; hypothetico-deductive method; social conformity; hypotheses

Are Aesthetic Judgements Purely Aesthetic?

1. The Hypothetico-Deductive Method in the Study of Mathematical Practice

Understanding expert mathematical practice is an interdisciplinary endeavour involving philosophers, historians, mathematics educators, psychologists and sociologists. Inevitably this implies that a diverse array of different research approaches has been used to gain insights into the behaviour of mathematicians. An indication of this diversity can be seen by studying the contents of Larvor's (2016) recent edited volume. The contributors to that collection drew conclusions about mathematical practice by studying historical episodes (e.g., Barany, 2016), by interviewing mathematicians (e.g., Johansen & Misfeldt, 2016), by analysing cultural artefacts (e.g., Pantsar, 2016) and by conducting detailed case studies of a particular mathematical notation (e.g., De Toffoli & Giardino, 2016). All these approaches are well-suited for generating insights about mathematical practice but are not always useful strategies if one wishes to test hypotheses that already exist, especially if those hypotheses involve causal claims.

Our goal in this paper is twofold. First, we outline the so-called hypothetico-deductive method (H-D), an approach that relatively few empirical studies focused on mathematical practice have adopted to date. Second, we exemplify the approach by reporting a study of mathematical aesthetics that explicitly set out to test a pre-existing causal hypothesis.

The philosopher of science Nancy Cartwright draws a useful distinction amongst empirical methods for warranting causal claims: "clinchers" are methods which decisively clinch the conclusion and "vouchers" are methods that can only vouch for it (Cartwright, 2007, p. 25). Clinchers bring deductive certainty but at the price of restrictive assumptions about their scope; vouchers are much wider ranging but their conclusions are inherently defeasible. However, as she observes, H-D is a "straddler": it can function as either a clincher or a voucher depending on the result of the test to which it is applied. The essence of H-D is

extremely simple. From some hypothesis, let us say H, an outcome, O, is deduced (hence the name hypethetico-deductive). This deduction permits us to assert as a necessary truth the conditional H→ O (assuming there are no confounding variables or other threats to internal validity). Now we perform an observation of some kind and either we get the outcome (O) or we don't (¬O). The second case is the clearest: we may reason with deductive validity from H→ O and ¬O to ¬H (an inference familiar to logicians as modus tollens). This is a clincher: H *must* be false.[1] The first case is more ambiguous: to reason from H→ O and O to H is a deductive fallacy (affirming the consequent). Some philosophers of science, notably Karl Popper, would see this as the end of the matter: no good, they aver, can come from such a step; we must not fool ourselves into thinking we can confirm hypotheses in this manner and should instead concentrate on disconfirming them. But most philosophers of science are more permissive; they accept a role for such inferences within H-D, albeit a much weaker one: vouching, not clinching. For example, although Carl Hempel (1966) emphasised that observing a predicted outcome does not prove that the hypothesis is correct, he argued that it may in certain circumstances raise its plausibility. In Hempel's terms, a sequence of successfully observed predictions derived from a hypothesis "provides at least some support, some partial corroboration or confirmation for" that hypothesis (p. 8). One important feature

[1] This is an idealization. The outcome seldom depends exclusively on the hypothesis; in addition, some auxiliary assumptions, say A, are usually required. Hence the conditional we can assert is not H→ O, but (H&A)→ O, and modus tollens with ¬O thereby gives us ¬H∨ ¬A: that is, *either* the hypothesis *or* (at least one of) the auxiliary assumptions must be false. (This insight is due to Pierre Duhem (1904, p. 183); for a helpful exposition, see Gillies (1993, pp. 98 ff.). Duhem's thesis lies behind and, once recognized, helps to resolve the so-called Kuhn–Popper debate over the falsification of theories by counterexamples (Worrall, 2003, p. 72). Our present concern, however, is with single hypotheses, rather than whole theories.) Nonetheless, if the auxiliary assumptions are true, we can still arrive at ¬H by the further deductively valid step of disjunctive syllogism. Of course, if we *knew* them to be true, they would not be assumptions, but it is an element of the art of hypothesis formation to choose hypotheses from which experimental outcomes can be deduced with the fewest, most reliable auxiliary assumptions.

of both the vouching and clinching modes of H-D is that the source of the hypothesis being tested is not specified. It may be based on prior research, the wider literature, or simply the researcher's intuition.[2]

Here we give two examples of where an H-D approach has been used to investigate mathematical practice. The first followed the vouching H-D mode of reasoning, the second the clinching mode. Weber and Mejía-Ramos (2011) reported data from a series of 17 interviews with research mathematicians. In each case they were asked to reflect on their behaviour as they read mathematical proofs. Weber and Mejía-Ramos identified three general strategies: appealing to the authority of other mathematicians who had read the proof (for example reviewers), line-by-line reading (where the reader proceeds linearly through the text) and modular reading (where the reader chunks up the text into logically coherent sections and evaluates how they fit together).

As Weber and Mejía-Ramos noted, they could not draw strong general conclusions from their interview study in view of the small number of participants involved. However, they could use it to generate hypotheses about the mathematical reading behaviour of expert mathematicians. In a subsequent study, these hypotheses were tested using the vouching mode of H-D. Mejía-Ramos and Weber (2014) surveyed 118 research-active mathematicians, asking them about their reading habits. They reasoned that if the findings from their earlier interviews were representative of mathematical reading behaviour more generally, then they would expect to see large proportions of survey respondents stating that they (i) often appeal to the reputation of earlier readers of the proof (i.e., the journal's review process), (ii) often

[2] What we are here calling the vouching and clinching modes of H-D are discussed by Douglas Walton as the two argumentation schemes characteristic of Argument from Evidence to a Hypothesis: Argument from Verification and Argument from Falsification, respectively (Walton, 1996, pp. 67 ff.). For further discussion of these schemes in the context of mathematics, see Aberdein (2019, p. 831). The vouching H-D mode of reasoning is also familiar from the work of Charles Peirce as Abduction (for a direct comparison of these and other treatments of such reasoning, see Pease & Aberdein, 2011).

would check whether certain steps in the proof were valid, and (iii) try to understand the proof in terms of how its main ideas fit together. They found evidence that was consistent with all three predictions, allowing Mejía-Ramos and Weber to conclude that their hypotheses were more plausible after the study than they had been before.

The clinching mode of H-D reasoning has also been used to draw conclusions about mathematical practice. For example, Mejía-Ramos and Inglis (2011) were interested in the extent to which natural language meanings of words influence the way they are understood in mathematical contexts. By studying word frequencies obtained from corpora designed to be representative of day-to-day English in different contexts, they noticed that the verb form of proof ('prove') was more common in informal contexts than in specialist contexts, and that the noun form ('proof') was more common in specialist contexts than in informal contexts. Based on this, and on the theory that how mathematical words are understood is influenced by word meanings in natural language (the so-called semantic contamination hypothesis), they predicted that mathematicians would evaluate an ambiguous visual 'proof' differently depending on whether they were asked "is the argument a proof of the claim?" or "does the argument prove the claim?". Because visual proofs are less formal, Mejía-Ramos and Inglis suggested that more positive responses would be found in the verb condition than in the noun condition.

However, in line with Popper's clinching mode of H-D reasoning, Mejía-Ramos and Inglis (2011) set up their study by considering the negation of their hypothesis. Specifically, they examined the hypothesis that there would be no difference in mathematicians' interpretations of the verb and noun forms. If that hypothesis were correct, then no difference in the proportion of positive responses would be expected between the two conditions. But across two experiments Mejía-Ramos and Inglis found consistent differences: participants who were asked whether the argument proved the claim were more likely to say yes than

those who were asked whether the argument was a proof of the claim (this difference was significant at the 5% level, allowing the null hypothesis of no effect to be rejected). Mejía-Ramos and Inglis therefore concluded that the hypothesis that there is no difference between how mathematicians interpret the verb and noun forms of 'proof' was probably incorrect.

Note that although the rejection of this hypothesis follows deductively (assuming there has not been a Type 1 statistical error) from Popper's clinching version of H-D reasoning, we cannot conclude with certainty that the theory which led to the hypothesis must be correct. Specifically, there could be some other reason – other than the relative formality and informality of the verb and noun forms of 'proof' in natural language – that might explain why mathematicians interpret these words differently.[3] However, in line with the vouching mode of H-D, our confidence in this theory should be increased: the theory was used to deduce a risky prediction, and that prediction was observed across two experiments.

A further remark is in order about Mejía-Ramos and Inglis's (2011) study. The prediction they deduced from the semantic contamination hypothesis was somewhat artificial: clearly mathematicians are rarely presented with visual arguments and asked to make binary decisions about whether or not they are proofs, or whether or not they prove. In this sense, the hypothesis clinched by Mejía-Ramos and Inglis can be seen as lacking in external validity. It nevertheless successfully vouches for the wider semantic contamination hypothesis. As Mook (1983) argued, in many experimental situations the researcher does not aim to make predictions about the real world from the laboratory, but rather aims to test predictions (derived from theory) about what should happen in the lab. If the semantic contamination hypothesis were correct, we would expect to see its effects in the artificial setting constructed by Mejía-Ramos and Inglis. The fact that these effects were indeed observed, should increase our confidence in the hypothesis's adequacy.

[3] In other words, it is an auxiliary assumption that such other reasons can be ruled out.

Our goal in the remainder of the paper is to report a study that investigated mathematical practice using an explicitly H-D approach. Before discussing our main hypothesis, methods and results, we briefly situate the study's research question within the domain of mathematical aesthetics.

2. Mathematical Aesthetics

The notion of mathematical beauty is puzzling. The attribution of aesthetic properties to abstract mathematical objects – proofs, theorems, definitions, axiomatic systems – seems to be a ubiquitous part of mathematical practice: mathematicians regularly assess each others' proofs using aesthetic terms, and award each other prizes for work that is thought particularly deep or beautiful. But in what sense can mathematical proofs have aesthetic properties? Poincaré (1914) argued that mathematical beauty is a "real aesthetic feeling that all true mathematicians recognize" (p. 59), and many mathematicians have claimed that their research is driven by a pursuit of beautiful proofs (Engler, 1990). Given this, providing an account of how mathematical objects – especially proofs – acquire aesthetic properties seems to be a necessary feature of any adequate account of mathematical thinking and reasoning.

Philosophical accounts of mathematical beauty fall into two broad categories: *aesthetic realism* and *aesthetic anti-realism*. According to the realist position, mathematicians' aesthetic judgements successfully track intrinsic properties of the mathematics being assessed. In contrast, anti-realist accounts deny the existence of intrinsic properties, and instead suggest that aesthetic properties are projected onto the mathematics by the reader. Ernest (2016, p. 199) noted that "it remains an open question as to whether beauty is an objective or subjective mathematical value", and pointed out that this issue has been discussed since at least the time of Plato.

A second-order dispute concerns whether mathematicians' aesthetic judgements are genuinely aesthetic. Some theorists, such as Rota (1997) and Todd (2008), suggest that when mathematicians attribute aesthetic properties to mathematical proofs, they are actually assessing some non-aesthetic quality. These *reductive* accounts typically propose that the quality being assessed is some epistemic property, perhaps the extent to which the proof enlightens its reader. According to this view, when mathematicians use aesthetic adjectives to describe proofs, they are merely proxies for more appropriate epistemic adjectives. In contrast *literal* accounts, such as Hardy's (1940) or McAllister's (2005), take mathematicians' language at face value, and propose that apparently aesthetic adjectives reflect genuine aesthetic appraisals. Support for this position came from Zeki, Romaya, Benincasa and Atiyah's (2014) finding that the experience of mathematical beauty activates similar brain areas to those activated when experiencing beauty from other sources.

The reductivism/literalism distinction in mathematical aesthetics does not match exactly to the aesthetic realism/anti-realism distinction. Although there may be a greater overlap between literalists and realists than between literalists and anti-realists, and likewise a greater overlap between reductivists and anti-realists than between reductivists and realists, the two distinctions are conceptually independent and all four possible positions are in principle defensible.

Many mathematicians and philosophers are motivated to adopt a realist account because of the widespread observation that mathematicians' aesthetic judgements are largely homogeneous. Paul Dirac, for instance, remarked that, in literature, poetry and art, beauty may depend upon idiosyncratic factors such as culture and upbringing; in contrast, he claimed that mathematical beauty "is of a completely different kind and transcends these personal factors. It is the same in all countries and at all periods of time" (cited in Dyson, 1992, p. 305). Similarly, Rota (1997, p. 175) argued that "the beauty of a piece of

mathematics does not consist merely in subjective feelings experienced by an observing mathematician. The beauty of a theorem is a property of the theorem on a par with its truth or falsehood. […] Both the truth of a theorem and its beauty are equally objective qualities, equally observable characteristics of a piece of mathematics which are equally shared and agreed upon by the community of mathematicians." Views such as these appear to be widespread among practicing mathematicians (e.g., Bass, 2011; Sinclair, 2009).

While it is not possible to conclusively resolve the aesthetic realism debate via empirical investigation, empirical researchers can contribute to addressing the question of whether or not mathematicians do in fact exhibit a large degree of consensus when aesthetically appraising proofs. Note that these two questions are, strictly speaking, distinct. Under a realist account, it is possible that mathematicians could all agree that a given proof is beautiful when in fact it is ugly. Alternatively, mathematicians might disagree about a proof's beauty despite there being an objectively correct position. However, as we have previously argued (Inglis & Aberdein, 2016), it seems prima facie implausible that appraisals of mathematical beauty should motivate this sort of distinction: for instance, how might a proof be beautiful if no mathematician finds it so, or lack beauty although most mathematicians regard it so?[4]

The subjectivity or intersubjectivity of mathematical aesthetics bears on the question of whether aesthetics can be productively harnessed in educational contexts. Many mathematics education researchers have suggested that integrating aesthetic appreciation into the school curriculum would be valuable (e.g., Burton, 1995, 2001; Dreyfus & Eisenberg,

[4] In other words, we maintain that there are no unknowable truths of mathematical aesthetics, since this discourse exhibits what Crispin Wright calls "epistemic constraint" (Wright, 1992, p. 41). As Shapiro and Taschek (1996) observe of some other epistemically constrained discourses, "surely it would be bizarre to maintain that some things are genuinely funny, or delicious, although no one can ever know this" (p. 75). We express no opinion on the larger, independent, and genuinely vexed question of whether the discourse of mathematics itself is epistemically constrained.

1986; Sinclair, 2001, 2004). Some have suggested that doing so may have motivational benefits (e.g., Burton, 2001; Sinclair, 2004; 2009), that it might make mathematics more relevant for children (e.g., Sinclair, 2001), and also that it is inappropriate to deny learners experiences that are apparently so central to mathematicians' mathematical experiences (e.g., Burton, 2001; Dreyfus & Eisenberg, 1986). Dreyfus and Eisenberg went as far as to say that "something is terribly amiss in the mathematics curriculum" because of a lack of attention to aesthetics (p. 9).

Integrating aesthetic appreciation into mathematics education seems to presuppose a literal account (if aesthetic judgements actually concerned enlightenment, then it would surely be preferable to make pedagogical choices based on enlightenment directly). Moreover, the project of incorporating aesthetics into education would be dramatically simplified if aesthetic judgements were intersubjective. The possibility that such judgements are subjective raises several worries. One concerns elitism: if subjective aesthetic judgements influenced curriculum choices, then this would likely privilege the (arbitrary) views of elite mathematicians (cf. Sinclair, 2009). Another worry relates to what the intended learning outcomes of such a curriculum would be. While Crespo and Sinclair (2008) argued that mathematics students may be able to "learn to identify and even value the [aesthetic] criteria that guide the mathematical community" (p. 406), if aesthetic judgements were entirely subjective then no such criteria would exist.

Given the pedagogical importance some researchers attribute to aesthetic factors, and given that the manner in which aesthetics could or should be integrated into education is clearly affected by whether mathematical aesthetics are subjective or intersubjective, understanding this issue further seems important. But, despite the apparent importance of the topic, the subjectivity or intersubjectivity of mathematical aesthetics has not received a great deal of attention in the empirical research literature. An early effort to investigate the

question was made by Wells (1990), who conducted a study which he described as too "small" and "crude in construction" (p. 40) to permit strong conclusions. He invited readers of the *Mathematical Intelligencer* to rate the beauty of 24 theorems on a scale of 0 to 10. The lack of consensus apparent in the responses from 68 readers led Wells to suggest that "the idea that mathematicians largely agree in their aesthetic judgements is at best grossly oversimplified" (p. 40).

More recently two relevant studies have been published on the topic. These focused on the aesthetic judgements of laypeople and experts respectively. Johnson and Steinerberger (2019) reported an intriguing study in which 300 laypeople recruited from the Amazon MTurk platform were asked to read and reflect on four mathematical arguments. They were then asked to state the extent to which each argument was similar to four different landscape paintings, using a 0 to 10 scale (in a second study the paintings were replaced by clips of classical music). Participants' ratings were not random, indicating that there was at least some degree of consensus between participants' judgements (i.e. more participants thought that Gauss's demonstration of how to add the first n integers is more similar to a Constable painting of Suffolk than it is to a Bierstadt painting of Yosemite, or at least more participants thought that than would be expected by chance alone). A similar result was obtained when participants were asked to judge the similarity of the mathematical arguments and clips of classical music. These findings are consistent with what we would expect if aesthetics are at least somewhat intersubjective. If participants did not agree, to at least some degree, about the aesthetics of mathematical arguments, it is hard to see where these non-random similarity ratings would have come from. On the other hand, Johnson and Steinerberger's data do not support the kind of intersubjective account advanced by Dirac (Dyson, 1992) or Rota (1997). While participants' ratings exhibited more consensus than one would expect by chance, they were far from consistent. For instance, the highest degree of between-participant consensus

was only 37% (in the sense that the proportion of participants ranking a given artwork as most similar to a given argument never rose above 37%, compared to the chance level of 25%).

In our own work we have adopted a very different empirical approach. Building on our earlier investigation of the dimensionality of mathematical proof appraisal (Inglis & Aberdein, 2015), we constructed a short 'personality' scale which allows us to assess perceptions of a given proof on four dimensions: aesthetics, intricacy, precision and utility (Inglis & Aberdein, 2016). The short scale, based on the format of Saucier's (1994) scale for assessing human personalities, consists of four adjectives per dimension (shown in Table 1) that were chosen to give acceptable internal reliability coefficients. Such a scale can be used to identify where a mathematician positions a given proof on the four dimensions: participants can be asked to read the proof and then state the extent to which each of the adjectives in our short scale (presented in a random order) accurately characterises it, using a Likert scale. The sum of the responses for each dimension provides an estimate of where the mathematician positions the proof on that dimension. By asking research mathematicians to rate a specific proof using this scale we were able to demonstrate that there is, at least in some cases, substantial heterogeneity in mathematicians' proof appraisals. In other words, whereas some mathematicians rated our proof at the high end of the aesthetics dimension, others rated exactly the same proof at the low end (Inglis & Aberdein, 2016).

Table 1
*The short scale developed by Inglis & Aberdein (2016). All Cronbach's αs > .75. *reverse scored.*

Adjective	Dimension	Adjective	Dimension
ingenious	Aesthetics	flimsy	Non-Use
inspired	Aesthetics	shallow	Non-Use
profound	Aesthetics	careful	Precision
striking	Aesthetics	meticulous	Precision
dense	Intricacy	precise	Precision
difficult	Intricacy	rigorous	Precision
intricate	Intricacy	applicable	Utility
simple*	Intricacy	useful	Utility
careless	Non-Use	informative	Utility
crude	Non-Use	practical	Utility

Although these two studies appear to favour different answers to the question of whether mathematical aesthetics are subjective or intersubjective, Johnson and Steinerberger (2019) pointed out that this conflict may be artificial. For instance, one possibility is that both experts and laypeople exhibit low levels of agreement about mathematical aesthetics. If this were true then, although both groups exhibit similar underlying behaviour, our initial reactions in the two cases might be different because we expect experts to agree and laypeople to disagree.

Critically, none of the three studies we have discussed (Inglis & Aberdein, 2016; Johnson & Steinerberger, 2019; Wells, 1990) provide support for the level of intersubjectivity in aesthetic judgement asserted by Dirac (Dyson, 1992) or Rota (1997). Thus an important question arises: if mathematicians do not exhibit a high degree of consensus when making aesthetic judgements, why do mathematicians and philosophers commonly suppose the opposite? For instance, citations to mathematical prizes often laud the awardees for their profound and beautiful work. How is publicising such an evaluation tenable if there is only minimal consensus between mathematicians about such matters? If Johnson and Steinerberger (2019) and Inglis and Aberdein (2016) were correct to suppose only low levels of between-mathematician consensus, then a great many readers of such citations will disagree with these judgements. Why is there not more public dissent?

Here we explore one hypothesis designed to answer this question. In both previous investigations of the subjective versus intersubjective distinction in mathematical aesthetics, participants have been asked to reach their judgements in isolation from others. In particular neither the participants in Johnson and Steinerberger's (2019) study nor those in ours (Inglis & Aberdein, 2016) were able to observe how other participants were answering. In few real-world contexts where judgements of mathematical aesthetics are made is this the case. Perhaps one reason such judgements tend to cohere in real-world contexts is that

mathematicians are strongly influenced by the aesthetic judgements of their peers. In other words, perhaps aesthetic judgements in mathematics are partly social, and not purely aesthetic.

This phenomenon – where participants modify their judgements to better match the judgements of others – is referred to as *conformity* by social psychologists. Perhaps the most striking demonstration of the phenomenon came in Asch's (1956) classic series of experiments which showed that individuals' judgements of objective physical properties such as length are subject to social influence. Asch asked participants to judge which of three lines was the same length as a fourth. He found that if participants were asked to give their response after a series of confederates (experimenters posing as fellow participants) had each endorsed the wrong answer, then they too would typically give the wrong answer. Similarly, participants' judgements of works of art can be manipulated by informing them that the art was produced by an expert (e.g., Bernberg, 1953; Duerksen, 1972).

Given Asch's (1956) work, and the literature it spawned, it seems reasonable to suppose that the aesthetic judgements of mathematicians might also be subject to social conformity. The main purpose of the study reported in this paper was to test this hypothesis. We had two specific aims. First, we attempted to replicate our earlier finding that mathematicians often seem to disagree about whether a specific proof exhibits aesthetic properties (Inglis & Aberdein, 2016). Second, we asked whether or not aesthetic judgements about mathematical proofs are influenced by social conformity. To this end, we asked a large number of mathematicians to read the same proof used in our earlier study, and to rate its aesthetics, intricacy, precision and utility using the short scale described above. Half of our participants were told that the proof had appeared in *Proofs from THE BOOK* – a collection of beautiful proofs selected for their "brilliant ideas" – the other half were not. We paid

particular attention to aesthetics, and asked (i) if participants' judgements converged and (ii) whether knowing the origin of the proof influenced participants' appraisals.

Our approach directly followed Popper's clinching H-D method outlined above. Specifically, rather than directly investigate our hypothesis of interest, we examined its negation: that the presence of the information that the proof was published in *Proofs from THE BOOK* would *not* influence mathematicians' aesthetic judgements. If we were able to observe a causal relationship between the condition a mathematician was assigned to and their aesthetic judgement, then we would be able to reject the hypothesis that our *Proofs from THE BOOK* manipulation was irrelevant to aesthetic judgements. This, in turn, would raise our confidence in the underlying theoretical hypothesis from which the prediction was derived, namely that aesthetic judgement is influenced by social conformity.

3. Method

3.1 Participants

A power analysis indicated that we required 200 participants to have 80% power of detecting a between-conditions difference of $d = 0.4$, which would typically be regarded as a small-to-medium effect size. We therefore aimed to continue recruiting participants until we had exceeded $N = 200$. When we ceased data collection we had collected responses from 203 research-active mathematicians based at British or American universities.

Participants were recruited via an email sent by their departmental secretary. The email explained that the study aimed to investigate how mathematicians appraise the quality of proofs and invited the recipient to visit a website to take part. Once participants had given consent, they were asked two brief demographic questions. First, we asked them to state their broad research area (pure mathematics, applied mathematics, statistics) so we could control for this in our analysis, as prior research has found that an individual's research area is

sometimes predictive of their response to proof appraisal tasks (Inglis, Mejía-Ramos, Weber & Alcock, 2013). Second, we asked participants to state their current role to ensure that all participants were active researchers. Responses to these questions revealed that our sample consisted of 56 research students and 147 faculty (including postdoctoral researchers); and of 121 pure mathematicians, 58 applied mathematicians and 23 statisticians (1 participant declined to identify their primary research area).

3.2 Procedure

Once participants had completed the demographic questions, they clicked through to a website which asked them to read a proof of the Sylvester-Gallai Theorem, shown in Figure 1. Those who had been randomly assigned to the 'sourced' condition were additionally told that that the proof had been "taken from Aigner & Ziegler, *Proofs from THE BOOK*, 5th Edition, p. 73." (This text was placed in parentheses directly after the 'Proof' subheading.) Those who had been randomly assigned to the 'unsourced' condition read exactly the same theorem and proof, except that this line was missing (thus they saw exactly the same stimulus as used by Inglis & Aberdein, 2016).

Theorem. *In any configuration of n points in the plane, not all on a line, there is a line which contains exactly two of the points.*

Proof. Let $\mathcal{P}$ be the given set of points and consider the set $\mathcal{L}$ of all lines which pass through at least two points of $\mathcal{P}$. Among all pairs (P, ℓ) with P not on ℓ, choose a pair (P_0, ℓ_0) such that P_0 has the smallest distance to ℓ_0, with Q being the point on ℓ_0 closest to P_0 (that is, on the line through P_0 vertical to ℓ_0).

Claim: This line ℓ_0 does it!

If not, then ℓ_0 contains at least three points of $\mathcal{P}$, and thus two of them, say P_1 and P_2, lie on the same side of Q. Let us assume that P_1 lies between Q and P_2, where P_1 possibly coincides with Q. The figure below shows the configuration. It follows that the distance of P_1 to the line ℓ_1 determined by P_0 and P_2 is smaller than the distance of P_0 to ℓ_0, and this contradicts our choice for ℓ_0 and P_0. □

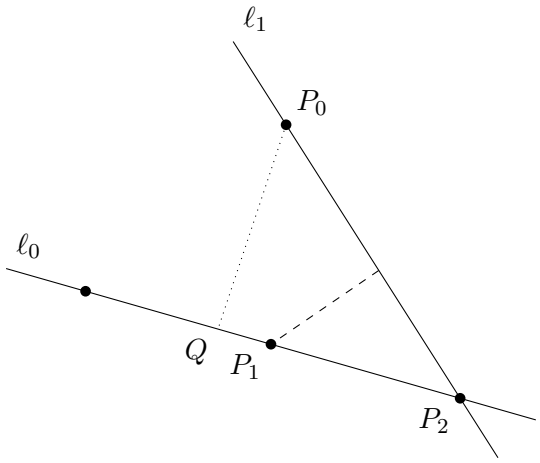

Figure 1. The theorem and proof seen by participants in the unsourced condition. Those in the sourced condition saw a version which also included "(Taken from Aigner & Ziegler, *Proofs from THE BOOK*, 5th Edition, p. 73.)" after the "Proof" subheading, but that was otherwise identical.

Paul Erdős (1913–1996) was one of the most celebrated pure mathematicians of the twentieth century. He is remembered for his extraordinary productivity – more than 1,500 papers with more than 500 co-authors – and the high calibre of his mathematical work, but also for his pithy anecdotes. *Proofs from THE BOOK* is a collection of supposedly highly aesthetic proofs inspired by one such anecdote: it is modelled on Erdős's whimsical suggestion that there is a book "in which God maintains the perfect proofs for mathematical theorems" (Aigner & Ziegler, 2014, p. V). Clearly God's version of the book is inaccessible, but there have been attempts to approximate it. Aigner and Ziegler's anthology is the best known: proofs that qualified for inclusion were said to contain "brilliant ideas, clever insights and wonderful observations" (p. V). This particular proof was described by Aigner and Ziegler as being "simply the best" (p. 73), and Bondy (1997) has suggested that it was Erdős's favourite proof.

On the same page participants were asked to "select how accurately each of the following words described **this proof**" (emphasis in the original), and were shown a randomly ordered list of the adjectives in Table 1. Note that we included four 'non-use' adjectives ('careless', 'crude', 'flimsy' and 'shallow') to ensure that at least some of our adjectives would receive low ratings. Responses to each adjective were made on a five-point Likert scale (very inaccurate, moderately inaccurate, neither inaccurate nor accurate, moderately accurate, very accurate). Finally, participants were thanked for their time and invited to contact the researchers if they had any questions.

4. Results

Nine participants (eight faculty) failed to respond to more than five of the twenty adjectives and were deleted from the analysis. Seven further participants failed to respond to

up to five adjectives (0.4% of the dataset). These missing data were imputed using item means. Thus the final dataset consisted of responses from 194 participants. First we calculated dimension scores by summing the responses to each adjective associated with each dimension (with the 'simple' response reverse scored). Judgements about the proof's aesthetics were therefore represented by a score between 4 (low) and 20 (high). Analogous scores were created for the other three dimensions.

The distribution of scores for each dimension are shown in Figure 2. For each of the four dimensions, some participants rated the proof highly, while others did not. These results therefore replicated our earlier finding (Inglis and Aberdein, 2016) that there is substantial heterogeneity in mathematicians' proof appraisals, at least for this specific proof.

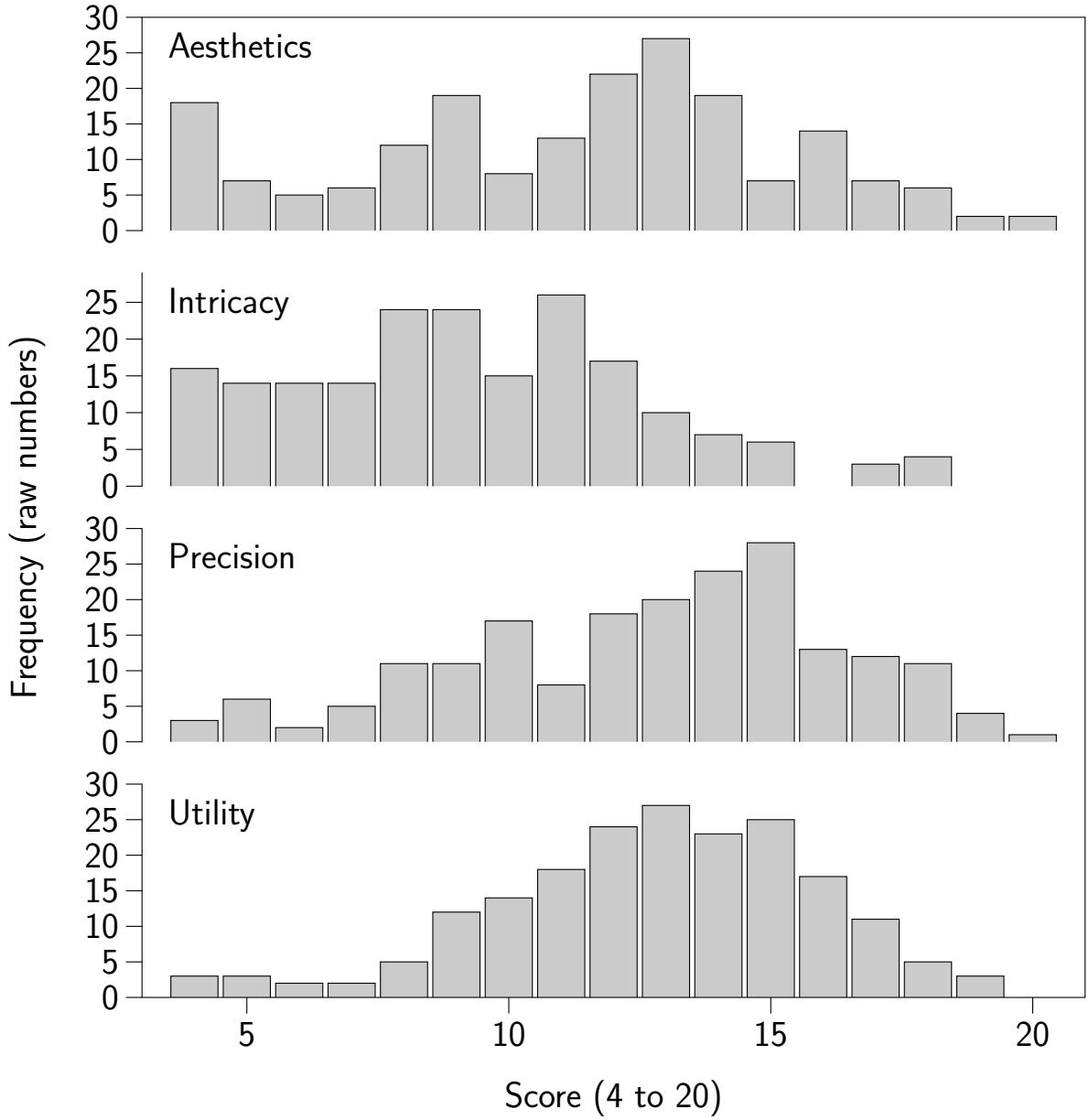

Figure 2. Histograms showing how participants rated the proof on each of the four dimensions.

Because we had relatively few statisticians in the sample, for the main analysis we collapsed the applied mathematicians and statisticians into a single group. Our primary hypothesis was related to the Aesthetics scores. These were subjected to a 2 (Condition: sourced, unsourced) × 2 (Group: student, faculty) × 2 (Research Area: pure, applied) between-subjects Analysis of Variance (ANOVA). This revealed a non-significant main effect of Group, $F(1,186) = 3.127$, $p = .079$, $\eta_p^2 = .017$, and a significant Condition × Research Area interaction, $F(1,186) = 4.647$, $p = .032$, $\eta_p^2 = .024$. This interaction is shown in Figure 3, and reflected that there was a significant effect of Condition for the pure mathematicians, $t(97.9) = 2.911$, $p = .004$, but not for the applied mathematicians, $t(75) = 0.425$, $p = .672$. Pure mathematicians who were told that the proof had been published in *Proofs from THE BOOK* rated it higher than those who did not (12.69 v 10.55, difference 2.15, 95% CI [0.71, 3.58]), whereas those in the applied group gave similar ratings across conditions (10.43 v 10.81, difference -0.39, 95% CI [-2.19, 1.42]). No other main effects or interactions in the ANOVA approached significance, ps > .1.

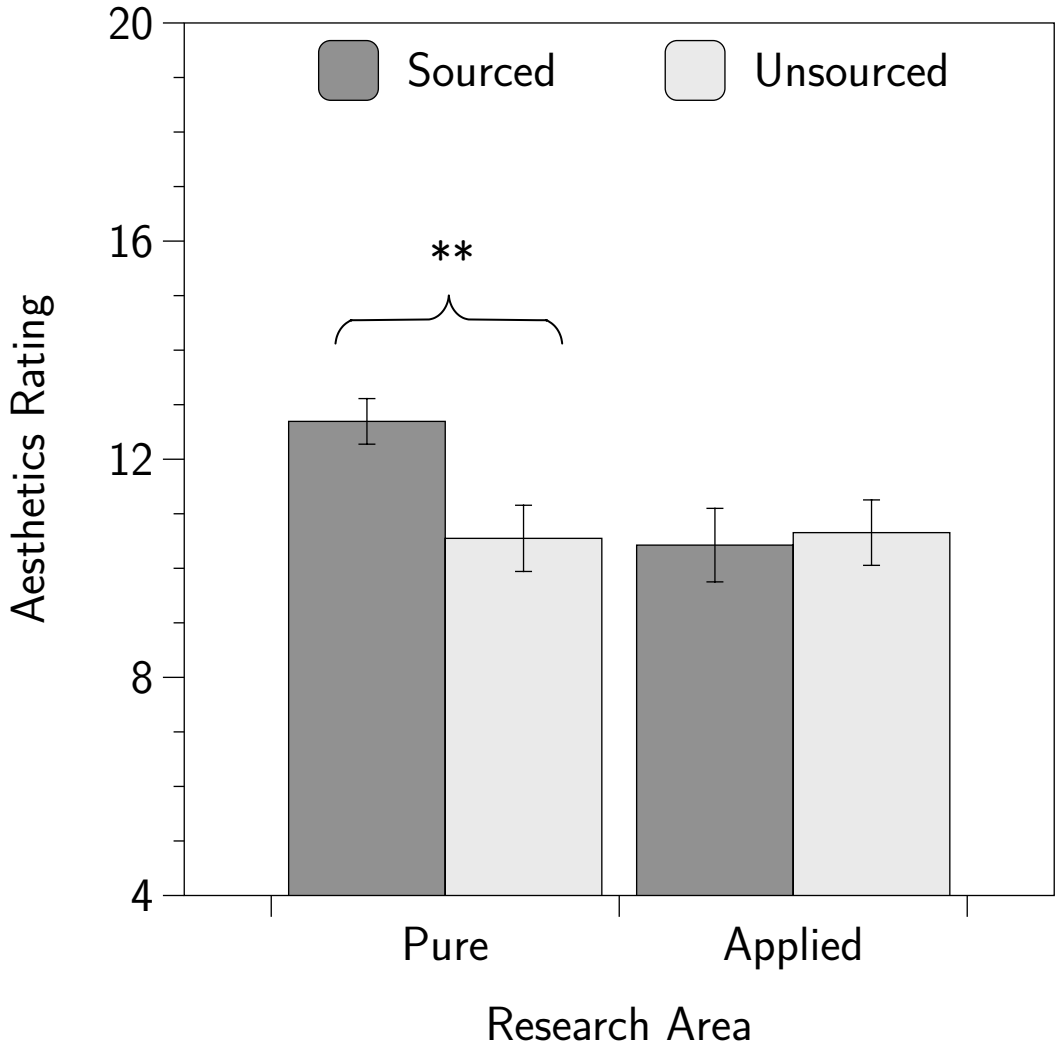

Figure 3. The mean Aesthetics rating given by pure and applied mathematicians in the two conditions. Error bars show ±1 SE of the mean. **$p < .01$.

Although our primary hypothesis concerned aesthetic judgements, we conducted similar analyses for the other three dimensions. These revealed that the proof was rated as being more intricate by the applied group than the pure group, 10.3 v 8.7, $t(192) = 3.309$, $p = .001$, difference 1.58, 95% CI [0.64, 2.52], and that the proof was rated as being more precise by the pure group than the applied group, 13.4 v 11.7, $t(192) = 3.376$, $p < .001$, difference 1.72, 95% CI [0.72, 2.73]. No other main effects or interactions approached significance.

5. Discussion

To summarise, we found two main results. First, as in our previous study (Inglis & Aberdein, 2016), we observed substantial heterogeneity in mathematicians' judgements about the aesthetic properties of this proof. Whereas some mathematicians gave the proof very high aesthetic ratings, others gave it very low aesthetic ratings. Second, we found evidence of social conformity: pure mathematicians gave the proof higher ratings if they had been told it was published in *Proofs from THE BOOK*, an anthology of proofs that God would supposedly deem to be perfect. In contrast, applied mathematicians showed no such effect.

Why did we find an effect for pure mathematicians but not applied mathematicians? We believe that this difference simply reflects that Erdős's notion of God's book is more familiar to pure mathematicians. Erdős was active in several domains of pure mathematics, but rarely ventured into applied domains. Informal discussions with mathematical colleagues back up this suggestion: The Book seems to be a more familiar notion among pure than applied mathematicians. Our experimental manipulation – we simply provided a citation for the proof – was extremely light touch. Clearly, we would not expect participants unfamiliar with the history and nature of The Book to be influenced by a manipulation that merely stated the title of a book based on it.

This interaction effect, along with the lack of between-conditions differences for the intricacy, precision and utility dimensions, renders implausible an alternative account which suggests that the differences observed were due to differences in perceptions of the overall quality of the proof, rather than just its aesthetics. Prior research has found that mathematical arguments are more likely to be rated as persuasive when they are authored by noted mathematicians (Inglis & Mejía-Ramos, 2009), so perhaps the addition of a reference to Aigner & Ziegler's (2014) text merely increased the proof's overall perceived quality. But if this were the case we would have expected to see effects on all four dimensions, not just the aesthetics dimension, and we would not have expected to see different effects for pure and applied mathematicians.

If our interpretation is correct, then in at least some cases social conformity can influence aesthetic judgements in mathematics. This result has implications for accounts of mathematical aesthetics. For example, one primary motivation for realist accounts of mathematical aesthetics is that mathematicians show a "remarkable degree of shared aesthetic sensibility" (Bass, 2011, p. 7). Our results suggest an alternative account. We have demonstrated that mathematicians show substantial heterogeneity in their aesthetic judgements, but it is plausible that this can only be observed in a context where social influence is minimised, such as the internet experiments conducted by Johnson and Steinerberger (2019) and Inglis and Aberdein (2016). If mathematicians' aesthetic judgements are influenced by the judgements of those around them, then it seems reasonable to suppose that – in day-to-day situations – mathematicians' aesthetic judgements do tend to converge, as claimed by Bass, Dirac, Rota, and others. But if such convergence can be successfully explained by conformity and social influence, then the existence of intrinsic

aesthetic properties, as asserted by literal realist accounts of aesthetic judgement, seems to be an unnecessary assumption.[5]

Of course, our results do not rule out realist accounts of mathematical aesthetics, they merely question a primary motivation for them. Just as Asch's (1956) experiments do not demonstrate that a line has no intrinsic length, only that human perception of length is subject to social influences, one could accept that mathematicians' perceptions are influenced by more than just the intrinsic aesthetic properties of the proof, without abandoning the assumption that these intrinsic properties do nevertheless exist. However, there is an important disanalogy between the two situations: in the case of line length there *is* an objective measuring device (a ruler), which gives one a good reason to wish to defend the assumption that length is an objective intrinsic property. There appears to be no analogous reason to defend the assumption that there are intrinsic aesthetic properties in the context of mathematics.

Regardless of its answer, the question of whether mathematical proofs (and other objects) have intrinsic aesthetic properties, the perceptions of which are influenced by social conformity, or whether social conformity is an adequate explanation for why mathematicians often assert that aesthetic judgements are consistent across mathematicians, has important implications for the role that aesthetics could have in the mathematics classroom. For example, the proposal that aesthetic appreciation should be an important goal of mathematics education seems difficult to reconcile with the suggestion that aesthetic realism is false and that aesthetic appreciation is nothing more than social conformity. If this account were correct it is hard to see how the proposal that "the aesthetic dimension plays a central role in

[5] A similar suggestion could plausibly account for the counter-intuitive finding that mathematicians' judgements of validity show considerable heterogeneity (Inglis, Mejía-Ramos, Weber, & Alcock, 2013; Weber, Inglis & Mejía-Ramos, 2014). Perhaps here too mathematicians tend to assume homogeneity because social influence and conformity operate in most real-world situations.

determining what mathematics proves personally or epistemologically relevant to children" (Sinclair, 2001, p. 25) could be true.

If, on the other hand, aesthetic realism is correct, and that social conformity moderates perceptions of intrinsic aesthetic properties of mathematical objects, then the suggestion that aesthetics could contribute to mathematics classrooms seems more plausible. Under this account, our finding that aesthetic judgements are subject to conformity would reinforce the elitism concern discussed by Sinclair (2009). As Sinclair noted, unlike the arts, mathematics has no tradition of public criticism in which aesthetic judgements are proposed and debated. Creating pedagogical materials and a classroom culture which permitted such a discussion, while minimising the risk that students would feel obliged to agree with the socially or epistemically privileged voices in the discussion, would seem to be a necessary but challenging task for those who advocate that aesthetics should play a greater role in the classroom.

More generally, our findings add to a growing body of research that demonstrates that mathematical practice seems to vary according to social and cultural factors (cf., Larvor, 2016), and that mathematicians sometimes evaluate mathematical statements using a variety of non-logical methods (e.g., Weber et al., 2014). Given the apparent centrality of aesthetics to the evaluation of mathematical quality at the research level, understanding how aesthetic judgements are reached seems to be necessary for a full understanding of advanced mathematical reasoning. Here we have demonstrated that any such account must include a role for social influence and conformity.

References

Aberdein, A. (2019). Evidence, proofs, and derivations, *ZDM Mathematics Education*, *51*, 825–834.

Aigner, M., & Ziegler, G. (2014). *Proofs from THE BOOK* (Fifth ed.). Berlin: Springer.

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied, 70,* 1–70.

Barany, M. J. (2016). Remunerative combinatorics: Mathematics and their sponsors in the mid-twentieth century. In B. Larvor (Ed.), *Mathematical cultures: The London meetings 2012–2014*, (pp. 329–346). Basel: Birkhäuser.

Bass, H. (2011). Vignette of doing mathematics: A meta-cognitive tour of the production of some elementary mathematics. *The Mathematics Enthusiast, 8,* 3–33.

Bernberg, R. E. (1953). Prestige suggestion in art as communication. *Journal of Social Psychology, 38*, 23–30.

Bondy, J. A. (1997). Paul Erdős et la combinatoire, *Gazette des Mathématiciens, 71,* 25–30.

Burton, L. (1995). Moving towards a feminist epistemology of mathematics. *Educational Studies in Mathematics, 28*, 275–291.

Burton, L. (2001). Research mathematicians as learners – and what mathematics education can learn from them. *British Educational Research Journal, 27*, 589–599.

Cartwright, N. (2007). *Hunting causes and using them: Approaches in philosophy and economics*, Cambridge: Cambridge University Press.

Cellucci, C. (2006). Introduction to *Filosofia e matematica*. In R. Hersh (Ed.), *18 unconventional essays about the nature of mathematics* (pp. 16–36). New York, NY: Springer.

Crespo, S., & Sinclair, N. (2008). What makes a problem mathematically interesting? Inviting prospective teachers to pose better problems. *Journal of Mathematics Teacher Education, 11*, 395–415.

De Toffoli, S., & Giardino, V. (2016). Envisioning transformations – The practice of topology. In B. Larvor (Ed.), *Mathematical cultures: The London meetings 2012–2014*, (pp. 25–50). Basel: Birkhäuser.

Dreyfus, T., & Eisenberg, T. (1986). On the aesthetics of mathematical thought. *For the Learning of Mathematics, 6*, 2–10.

Duerksen, G. L. (1972). Some effects of expectation on evaluation of recorded musical performance. *Journal of Research in Music Education, 20*, 268–272.

Duhem, P. (1904/1954). *The aim and structure of physical theory*. Princeton, NJ: Princeton University Press. Trans. P. P. Weiner.

Dyson, F. (1992). *From Eros to Gaia*. New York, NY: Pantheon Books.

Engler, G. (1990). Aesthetics in science and in art. *British Journal of Aesthetics, 30,* 24–34.

Ernest, P. (2016). Mathematics and values. In B. Larvor (Ed.), *Mathematical cultures: The London meetings 2012–2014*, (pp. 189–214). Basel: Birkhäuser.

Evans, J. St. B., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition, 11*, 295–306.

Franklin, J. (1987). Non-deductive logic in mathematics. *British Journal for the Philosophy of Science*, *38*, 1–18.

Gillies, D. (1993). *Philosophy of science in the twentieth century: Four central themes*. Oxford: Blackwell.

Hardy, G. H. (1940). *A mathematician's apology*. Cambridge: Cambridge University Press.

Hempel, C. (1966). *Philosophy of natural science*. Upper Saddle River, NJ: Prentice-Hall.

Inglis, M., & Aberdein, A. (2015). Beauty is not simplicity: An analysis of mathematicians' proof appraisals. *Philosophia Mathematica, 23,* 87–109.

Inglis, M., & Aberdein, A. (2016). Diversity in proof appraisal. In B. Larvor (Ed.), *Mathematical cultures: The London meetings 2012–2014*, (pp. 163–179). Basel: Birkhäuser.

Inglis, M., & Mejía-Ramos, J. P. (2009). The effect of authority on the persuasiveness of mathematical arguments. *Cognition and Instruction, 27*, 25–50.

Inglis, M., Mejía-Ramos, J. P., Weber, K., & Alcock, L. (2013). On mathematicians' different standards when evaluating elementary proofs. *Topics in Cognitive Science, 5*, 270–282.

Johansen, M. W., & Misfeldt, M. (2016). An empirical approach to the mathematical values of problem choice and argumentation. In B. Larvor (Ed.), *Mathematical cultures: The London meetings 2012–2014*, (pp. 259–270). Basel: Birkhäuser.

Johnson, S. G. B., & Steinerberger, S. (2019). Intuitions about mathematical beauty: A case study in the aesthetic experience of ideas. *Cognition 189*, 242–259.

Larvor, B. (2016) What are mathematical cultures? In S. Ju, B. Löwe, T. Müller, & Y. Xie (Eds.), *Cultures of mathematics and logic: Selected papers from the conference in Guangzhou, China, November 9–12, 2012* (pp. 1–22). Basel: Birkhäuser.

McAllister, J. W. (2005). Mathematical beauty and the evolution of the standards of mathematical proof. In M. Emmer (Ed.), *The visual mind II* (pp. 15–34). Cambridge, MA: MIT Press.

Mejía-Ramos, J. P., & Inglis, M. (2011). Semantic contamination and mathematical proof: Can a non-proof prove? *Journal of Mathematical Behavior 30*, 19–29.

Mejía-Ramos, J. P., & Weber, K. (2014). Why and how mathematicians read proofs: further evidence from a survey study. *Educational Studies in Mathematics, 85,* 161–173.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38*, 379–387.

Paseau, A. (2015). Knowledge of mathematics without proof. *British Journal for the Philosophy of Science*, *66*, 775–799.

Pantsar, M. (2016). The great gibberish – Mathematics in Western popular culture. In B. Larvor (Ed.), *Mathematical cultures: The London meetings 2012–2014*, (pp. 409–438). Basel: Birkhäuser.

Pease, A., & Aberdein, A. (2011). Five theories of reasoning: Interconnections and applications to mathematics. *Logic and Logical Philosophy*, *20*, 7–57.

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

Rota, G.-C. (1997). The phenomenology of mathematical beauty. *Synthese, 111,* 171–182.

Saucier, G. (1994). Mini-markers: A brief version of Goldberg's unipolar big-five markers. *Journal of Personality Assessment, 63,* 506–516.

Shapiro, S. and Taschek, W. W. (1996). Intuitionism, pluralism, and cognitive command. *Journal of Philosophy*, *93*, 74–88.

Sinclair, N. (2001). The aesthetic is relevant. *For the Learning of Mathematics, 21*, 25–32.

Sinclair, N. (2004). The roles of the aesthetic in mathematical inquiry. *Mathematical Thinking and Learning, 6*, 261–284.

Sinclair, N. (2009). Aesthetics as a liberating force in mathematics education? *ZDM–The International Journal on Mathematics Education, 41*, 45–60.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science, 22,* 259–264.

Todd, C. S. (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, 11,* 61–79.

Walton, D. N. (1996). *Argumentation schemes for presumptive reasoning.* Mahwah, NJ: Lawrence Erlbaum Associates.

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*, 36–58.

Weber, K., & Mejía-Ramos, J. P. (2011). Why and how mathematicians read proofs: An exploratory study. *Educational Studies in Mathematics, 76,* 329–344.

Wells, D. (1990). Are these the most beautiful? *The Mathematical Intelligencer, 12*, 37–41.

Worrall, J. (2003). Normal science and dogmatism, paradigms and progress: Kuhn 'versus' Popper and Lakatos. In T. Nickles (Ed.), *Thomas Kuhn*, (pp. 65–100). Cambridge: Cambridge University Press.

Wright, C. (1992). *Truth and objectivity*. Cambridge, MA: Harvard University Press.

Zeki, S., Romaya, J. P., Benincasa, D. M., & Atiyah, M. F. (2014). The experience of mathematical beauty and its neural correlates. *Frontiers in Human Neuroscience, 8*.