Cognition 189 (2019) 242–259

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original Articles

Intuitions about mathematical beauty: A case study in the aesthetic experience of ideas

Samuel G.B. Johnson[a,*], Stefan Steinerberger[b]

[a] Division of Marketing, Business, & Society, University of Bath School of Management, United Kingdom
[b] Department of Mathematics, Yale University, United States

ARTICLE INFO

Keywords:
Psychology of mathematics
Explanation
Aesthetics
Reasoning
STEM education

ABSTRACT

Can an idea be beautiful? Mathematicians often describe arguments as "beautiful" or "dull," and famous scientists have claimed that mathematical beauty is a guide toward the truth. Do laypeople, like mathematicians and scientists, experience mathematics aesthetically? Three studies suggest that they do. When people rated the similarity of simple mathematical arguments to landscape paintings (Study 1) or pieces of classical piano music (Study 2), their similarity rankings were internally consistent across participants. Moreover, when participants rated beauty and various other potentially aesthetic dimensions for artworks and mathematical arguments, they relied mainly on the same three dimensions for judging beauty—elegance, profundity, and clarity (Study 3). These aesthetic judgments, made separately for artworks and arguments, could be used to predict similarity judgments out-of-sample. These studies also suggest a role for expertise in sharpening aesthetic intuitions about mathematics. We argue that these results shed light on broader issues in how and why humans have aesthetic experiences of abstract ideas.

1. Introduction

> "Beauty is the first test. There is no permanent place in the world for ugly mathematics."
>
> G.H. Hardy

> "Mathematics is the music of reason."
>
> James Joseph Sylvester

Can an idea be beautiful? Scientists and mathematicians seem to think so—they often imbue explanations with aesthetic qualities. Proofs can be "elegant" or "beautiful"; they can be "dull" or "trivial." Albert Einstein "was quite convinced that beauty was a guiding principle in the search for important results in theoretical physics" (Zee, 1999), while physicist Paul Dirac (1963) even claimed that "it is more important to have beauty in one's equations than to have them fit experiment."

In this paper, we explore the aesthetics of ideas by testing laypeople's aesthetic experiences of mathematical arguments. In particular, we test whether laypeople have nonrandom degrees of consensus in their judgments of aesthetic similarity between mathematical proofs and art objects such as paintings and music (Studies 1 and 2), whether there is consensus about which proofs are most beautiful (Study 3), and what properties contribute to these judgments of aesthetic similarity and pleasure (Study 3). Overall, these findings reveal a degree of consensus about the aesthetics of even the most abstract kind of ideas. We argue that these findings contribute to our understanding of how and why ideas can evoke aesthetic experiences.

1.1. Aesthetics: the true, the adaptive, and the ugly

To judge by the practitioners quoted above, aesthetic experiences not only imbue their interactions with scientific and mathematical ideas, but guide them toward the truth. On the face of it, this is quite a strange (if not ugly) idea. When we seek knowledge, we strive to employ tools, such as deductive logic and the scientific method, that reliably lead to justified true beliefs. Aesthetic experiences, such as beauty, could hardly be more different. Indeed, beauty is commonly said to be in the eye of the beholder. David Hume (1985/1757) put this point (how else but) elegantly:

> Beauty is no quality in things themselves: It exists merely in the mind which contemplates them; and each mind perceives a different beauty. One person may even perceive deformity, where another is sensible of beauty; and every individual ought to acquiesce in his own sentiment, without pretending to regulate those of others.

* Corresponding author at: University of Bath School of Management, The Avenue, Bath BA2 7AY, United Kingdom.
E-mail address: sgbjohnson@gmail.com (S.G.B. Johnson).

https://doi.org/10.1016/j.cognition.2019.04.008
Received 14 December 2018; Received in revised form 3 April 2019; Accepted 10 April 2019

Some philosophers, while agreeing that beauty is subjective, have also claimed that people treat aesthetic judgments as universally true (e.g., Kant, 2002/1790). But recent experimental work casts doubt on this claim, suggesting that not only are aesthetic judgments subjective by nature, but also viewed as subjective by most laypeople (Cova, Garcia, & Liao, 2015; Cova & Pain, 2012).

This is a paradox. On the one hand, scientists and mathematicians experience ideas as beautiful or ugly and seem to use these experiences to guide them toward objective truth. On the other hand, aesthetic judgments seem to be inherently subjective, and most laypeople seem to believe that aesthetic judgments cannot be true or false.

Cognitive science provides one possible resolution to this paradox: An evolutionary account of aesthetics (e.g., Dutton, 2009; Hekkert, 2014; Pinker, 1997; Thornhill, 2003). On this view, aesthetic preferences (e.g., about art) are by-products of adaptations that serve survival and reproductive fitness. In some cases, these adaptations are directly linked to biological goals: Humans prefer symmetric over asymmetric faces, as the former tend to signal health (Rhodes et al., 1998; Thornhill & Gangestad, 1993; cf. Reber, Schwarz, & Winkielman, 2004), and prefer savannah-like landscapes, as they tend to boast water and shelter (Dutton, 2009; Orians & Heerwagen, 1992). In other cases, aesthetic preferences appear to be linked to more abstract stimulus processing goals, such as imposing order on visual stimuli as well as seeking an optimal degree of novelty (Berlyne, 1960; Boselie & Leeuwenberg, 1985; Hekkert, 2014). Approaching objects with these stimulus properties is adaptive for the same reason that vision is adaptive: Such approach patterns help the organism to make sense of and explore its environment. Aesthetic judgments are right or wrong, at most, in the same way that it is right or wrong to prefer ice cream over vegetables: We may be built to appreciate a fine gelato, but it is far from obvious that it is "wrong" to prefer broccoli, given the wild differences between our evolutionary and modern environments.

Thus, according to the evolutionary account, we tend to seek pleasure from (i.e., find beauty in) and therefore to approach objects in our environment that have a variety of adaptively desirable properties. Even though our aesthetic preferences evolved in response to natural objects such as faces and savannahs, they have been generalized to artifacts, such as paintings and designed gadgets, through cultural evolution acting on the raw psychological materials provided by biological evolution (Boyd & Richerson, 1985; Hekkert, 2014). Although this view is couched mainly in terms of visual aesthetics, similar mechanisms may well account for aesthetic preferences in other sensory modalities, such as music: Even non-human primates share some of the same musical preferences as infants and adults (Hauser & McDermott, 2003).

The discussion around the evolution of aesthetics has revolved almost exclusively around aesthetic *perception*. But what about the aesthetics of ideas? An emerging understanding of the cognitive science of explanation is consistent with the notion that our preferences for different explanations guide us toward explanations that are true (high in likelihood) and useful (high in their ability to guide understanding) (e.g., Douven & Schupbach, 2015; Johnson, 2017; Lipton, 2004; Lombrozo, 2016). For example, people often favor simple explanations over more complex ones, following Occam's razor. But in fact, this turns out to be one among several heuristic factors guiding people's explanatory preferences. People use simplicity as a proxy for an explanation's prior probability—for how likely it is to be true in the absence of data pointing either way (Lombrozo, 2007). But explanations that are too simple are often unable to account for all the data—think of the "hedgehog" political pundit who explains all geopolitical events in terms of her pet theory ("it's all about oil prices!") (Tetlock, 2005). Instead, people recognize that complex explanations have the virtue of more degrees of freedom for explaining complex data and therefore favor options with a moderate degree of complexity (Johnson, Valenti, & Keil, 2017). Comparable analyses have been given for a variety of "explanatory virtues," such as explanatory scope (i.e., the number of observations an explanations accounts for; Johnson, Johnston, Toig, & Keil, 2014; Johnson, Rajeev-Kumar, & Keil, 2016) and generality (i.e., level of abstraction; Johnston, Sheskin, Johnson, & Keil, 2018).

People *prefer* explanations that guide them toward truth and understanding, but it is less clear whether such preferences are experienced *aesthetically*. The poet John Keats claimed that "beauty is truth, truth beauty," but among philosophers there is disagreement. Some maintain that what makes something a good explanation is nothing more than its Bayesian probability, viewing explanatory reasoning as "inference to the likeliest explanation" (e.g., Van Fraassen, 1989). Others view broader considerations of explanatory "satisfyingness" as critical to an explanation's quality, viewing explanatory reasoning as "inference to the *loveliest* explanation" (e.g., Douven & Schupbach, 2015; Lipton, 2004), with the proviso that what humans find lovely often *is* likely. That sounds suspiciously similar to the account that evolutionary aesthetics gives for our approach to beautiful (adaptively beneficial) objects: The search for the loveliest explanation is a search, unconsciously, for explanations that lead to truth and understanding. Most directly of all, Gopnik (1998) has even likened the phenomenology of explanation to orgasm, arguing that the sense of pleasure we feel when we perceive an explanation as satisfying motivates us to acquire novel information, just as natural selection has endowed organisms with pleasure accompanying sexual orgasms to motivate reproduction.

The notion that people experience ideas aesthetically is thus consistent with anecdote, with phenomenology, and with evolutionary thought. But there is very thin empirical evidence for this proposition. The most direct evidence, to our knowledge, comes from studies of explanatory reasoning, where people seem to answer questions such as "how satisfying is this explanation?" and "how likely is this explanation to be true?" in similar ways, with the same factors influencing both judgments (e.g., Khemlani, Sussman, & Oppenheimer, 2011; Lombrozo, 2007). But this hardly makes an airtight case. We have suggested that people are heuristically substituting (Kahneman & Frederick, 2002) the difficult question of "how likely" with the subjectively easier question of "how lovely." But the logic works just as well the other way around. If people found the idea of explanatory loveliness bizarre, they could substitute the intelligible "how likely" question, producing the same pattern of results. To make a tighter case, other sources of evidence are required. Moreover, we need to study cases where all explanations are true to avoid the concern about reverse-directionality in heuristic substitution. A different approach is needed.

1.2. Intuitions about mathematical proof

For these reasons, we ask: Do laypeople have intuitions about the aesthetics of mathematical proofs? This operationalization solves the above problems—if we use indirect methods, we can avoid relying solely on direct questions, as well as avoiding the reverse-directionality problem by studying logically correct proofs (which are necessarily "true"). Moreover, this question is of independent interest to multiple constituencies. Most directly, it is of interest to mathematicians (as evidenced by articles in general-interest outlets for mathematicians; e.g., Krull, 1987; Wells, 1990). But it is also of interest to researchers in the psychology of mathematics and in math education by informing our understanding of how mathematical arguments are processed. Moreover, by comparing experts to laypeople, as we begin to do here, our work speaks to questions in the psychology of expertise, particularly the relationship between expertise and aesthetic judgment (e.g., Leder, Gerger, Dressler, & Shabmann, 2012; Margulies, 1977).

Is there a deep psychological reality underlying the perception that certain mathematical arguments are beautiful? Several previous studies, testing the judgments of professional mathematicians, suggest that there may be. For example, professional mathematicians seem to have a reasonable level of agreement about which theorems are most beautiful (Wells, 1990), though there is substantial disagreement about at least

some proofs (Inglis & Aberdein, 2016). One detailed study asked mathematicians to think of a recent proof they had read, rating it on a set of 80 dimensions, and factor-analyzed these judgments (Inglis & Aberdein, 2015). Mathematicians' judgments were captured by four components, including an aesthetic component (e.g., "beautiful," "sublime," "profound") as well as components reflecting intricacy ("dense", "difficult"), utility ("practical," "efficient"), and precision ("careful," "rigorous"). Thus, there is direct evidence that mathematicians' proof appraisals partly reflect aesthetic considerations. Even more directly, mathematicians' brains were scanned while contemplating either "beautiful" or "ugly" mathematical equations (Zeki, Romaya, Benincasa, & Atiyah, 2014). Judgments of mathematical beauty were correlated with activity in the same region of the medial orbito-frontal cortex that is known to track aesthetic judgments in other domains such as visual art and music (Ishizu & Zeki, 2011). Thus, the underlying experience of mathematical beauty, at least for mathematicians, seems to have a kinship with other forms of aesthetic experience.

However, it is not at all clear that these aesthetic experiences would extend to laypeople. Indeed, in Zeki et al.'s (2014) study, the effort to scan layperson participants' brains was abandoned after the researchers failed to find *any* equations that participants found beautiful. Perhaps, then, aesthetic experiences in mathematics are merely a by-product, not of human evolution, but of the social practices of mathematicians. At least, these data seemingly suggest, the aesthetics of mathematical arguments may require a great deal of training to appreciate.

Despite the discouraging direct evidence to date, might it nonetheless be possible that mathematical novices share the aesthetic perception of career mathematicians? Several lines of indirect evidence are suggestive.

Humans have a finely-tuned sense for numbers, which develops in infancy (Carey, 2009; Dehaene, 1996). Infants can distinguish among sets containing small numbers of objects with differing cardinality (Starkey & Cooper, 1980), can perform rudimentary addition and subtraction for such small sets (Wynn, 1992), and use an approximate number system (ANS) to discriminate among large arrays of objects (Xu & Spelke, 2000). Facility with numbers as measured in ANS tasks appears to undergird later mathematical ability, as this ability is associated with school math achievement (Halberda, Mazzocco, & Feigenson, 2008) and its impairment is associated with mathematical learning disabilities (Mazzocco, Feigenson, & Halberda, 2011). Just as people have a rudimentary ability to understand and use numbers, perhaps they possess deeper capacities for understanding more complex mathematical arguments. If so, perhaps this capacity might support later advanced mathematical abilities, such as those developed by professional mathematicians and taught in university courses in higher mathematics.

This may appear unlikely, given the famous disdain and hatred shown by high school students for geometry—most students' first introduction to rigorous mathematical proof (e.g., Senk, 1985). However, students' difficulties with mathematical proofs appear to be driven mainly by unfamiliarity with concepts and definitions, difficulty with mathematical notation, and uncertainty about how to start (Moore, 1994). That is, basic deductive logic is not necessarily unintuitive. This point is underscored by humans' basic capacity to perform simple forms of deductive reasoning, such as *modus ponens* (Johnson-Laird, 1999; Rips, 1994). This leaves open the possibility that people could *appreciate* the content of mathematical proofs even if they cannot *generate* them, so long as familiar concepts are used and mathematical notation is minimized.

Indeed, there is evidence that people automatically deploy deductive reasoning when contemplating very simple mathematical arguments. People have strong intuitions about the acceptability of simple mathematical explanations (Johnson, Johnston, Koven, & Keil, 2017) and such intuitions broadly track principles derived from philosophy of mathematics (e.g., Bolzano, 1817; Kitcher, 1975). For example, people evaluate "$4 - 2 = 2$ because $2 + 2 = 4$" as more acceptable than "$2 + 2 = 4$ because $4 - 2 = 2$," since the former explanation grounds a derivative operation (subtraction) in a more basic operation (addition), whereas the latter explanation does the converse (Johnson, Johnston, et al., 2017; Johnson, Valenti, et al., 2017). This is reminiscent of many mathematicians' preference to ground proofs of conceptually derivative domains in more conceptually primitive domains, such as geometry in analysis (Bolzano, 1817) or arithmetic in set theory (Frege, 1974/1884). Further, people appear to use chains of reasoning to evaluate more far-flung explanations such as "$3^2 = 9$ because $3 + 3 + 3 = 9$," since the acceptability of these explanations tracks the number of logical steps required to complete the derivation. The superiority of shorter proofs has been proclaimed by many, going back to Descartes (1954/1684) and Hume (1978/1738), on the basis that they are less likely to contain mistakes and more likely to confer understanding by holding the steps in mind simultaneously. Although such studies of simple arithmetic operations are a far cry from the sophisticated reasoning of professional mathematicians, they show that mathematical proof is not intrinsically unintuitive but in fact can be grasped automatically and intuitively, when complex concepts and notation are avoided.

1.3. The current studies

Our studies test whether laypeople do indeed have intuitions about the aesthetics of mathematical arguments. Unlike Johnson, Johnston, et al. (2017) and Johnson, Valenti, et al. (2017), we use sophisticated mathematical arguments familiar to practicing mathematicians; unlike Zeki et al. (2014), we avoid unfamiliar concepts and complex notation. These studies used two different approaches as converging evidence for aesthetic experiences in apprehending mathematical proofs.

First, Studies 1 and 2 tested aesthetic categorizations, examining whether people can systematically match artworks with mathematical arguments. These studies ask participants to read a series of mathematical arguments, and to rate their similarity to a set of landscape paintings (Study 1) or classical music performances (Study 2). Similar approaches have been used in empirical aesthetics and vision science to test related questions about aesthetic experiences and cross-modal associations. For example, one study examined correspondences between classical music and colors, asking participants to match which colors were most and least consistent with each piece of music (Palmer, Schloss, Xu, & Prado-León, 2013). There was a high degree of consensus in these cross-modal correspondences, both within and across cultures, which were mediated by emotional associations. Likewise, people make consistent cross-modal correspondences between pitch and color (Hubbard, 1996), pitch and size (Evans & Treisman, 2010), loudness and brightness (Marks, 1987), timbre and color saturation (Caivano, 1994), taste and shape (Velasco, Woods, Deroy, & Spence, 2015), color and odor (Levitan et al., 2014), and even taste and typeface (Velasco, Woods, Hyndman, & Spence, 2015) (see Spence, 2011 for a review). If we could likewise discover consistent correspondences between mathematical arguments and artworks, this would provide prima facie evidence that people have intersubjectively consistent aesthetic experiences of mathematical arguments.

Second, Study 3 asks *how* people are able to produce such consistent pairings between artworks and arguments. What dimensions do people use to evaluate the aesthetics of mathematical arguments and of artworks? Are these the same dimensions? Similarity judgments have been analyzed by many cognitive scientists (Gärdenfors, 2000; Goodman, 1972; Tversky, 1977), and a particularly difficult problem is the question of what dimensions people use for computing similarity (e.g., Medin, Goldstone, & Gentner, 1993). Study 3 begins to address this problem for perceptions of similarity between mathematics and art by measuring judgments of aesthetic pleasure (i.e., beauty) and several other potentially aesthetic qualities that could be shared across these domains (e.g., simplicity, profundity, intricacy). We then use these

judgments to examine how people assess beauty across domains and to model the similarity judgments from Study 1. Although these two lines of inquiry are individually susceptible to critique, they provide a converging case when taken together.

Throughout these studies, we also test the possible role of expertise in sharpening these aesthetic intuitions. Expertise shapes categorization schemes and perceptions of similarity. For example, physics experts use deeper physical principles to sort physics problems, whereas laypeople sort problems based on superficial features (Chi, Feltovich, & Glaser, 1981). Even among experts, categorization schemes can differ based on the nature of their expertise: Landscape workers sort trees based on goal-derived categories, whereas taxonomists place a greater weight on morphological features (Medin, Lynch, & Coley, 1997). Moreover, aesthetic expertise is known to shape aesthetic experiences (Leder et al., 2012). For example, art experts tend to have more nuanced aesthetic experiences (Fayn, Silvia, Erbas, Tiliopoulos, & Kuppens, 2018), dampened affective reactions to negative or unpleasant artworks (Leder, Gerger, Brieber, & Schwarz, 2014), and a tendency to experience interest rather than confusion in response to artworks (Silvia, 2013). Based on these lines of research, math experts may rely on different dimensions for assessing similarity and beauty; for example, professional mathematicians may place greater weight on "deeper" features of arguments. We begin to examine these issues by recruiting an additional, albeit small, sample of professional mathematicians in Study 1 and a larger sample of undergraduate math students in Study 2, as well as testing for differences in expertise in our larger layperson sample. We argue, overall, that even laypeople people share an intuitive sense of mathematical aesthetics, but that this sense sharpens with mathematics training.

2. Study 1

Our first two studies operationalized aesthetic experience in terms of aesthetic categorization, as measured through similarity judgments (e.g., Palmer et al., 2013). Specifically, we asked participants to rate the similarity of various pieces of art to a set of classic mathematical arguments. To the extent that people consistently categorize arguments with artworks, this suggests common aesthetic intuitions underlying participants' experiences of mathematics. Study 1 focused on comparisons of mathematical arguments to visual art—specifically, landscape paintings. We chose landscape paintings as an aesthetic domain that is easy to process quickly, yet relatively devoid of objects (e.g., people) which people might associate with the surface details of mathematical arguments. Moreover, evolutionary accounts of human aesthetics have proposed that human aesthetics evolved in part from an intuitive capacity to detect and approach resource-rich environments, such as those depicted in appealing landscape paintings (Dutton, 2009; Orians & Heerwagen, 1992; Pinker, 1997). As a potentially foundational aspect of human aesthetics, landscape depictions make a good test case for aesthetic correspondences with other domains.

2.1. Methods

Participants were recruited from the online crowdsourcing platform Amazon Mechanical Turk (N = 300, 46% female) and were from the United States; the median participant had completed a 4-year college degree. The same sample size was used in subsequent studies to avoid experimenter degrees of freedom. The research was deemed exempt from review by the Yale University Human Subjects Committee.

A subset of these participants (N = 99) had taken a university-level math course above the level of calculus, while a larger subset had not (N = 201). We refer to the former subsample as *experienced* and the latter as *inexperienced*.

Separately from the Mechanical Turk sample, a sample of professional mathematicians (N = 8) was recruited from the second author's professional network. These participants were contacted by email and we included all participants who completed the principal measures.

Participants each read four mathematical arguments (see Appendix A):

Geometric	Sum of an infinite geometric series
Gauss	Gauss's summation trick for positive integers
Pigeonhole	Pigeonhole principle
Faulhaber	Geometric proof of a Faulhaber formula

For each argument, they were asked to first read and reflect on the argument. Then, on subsequent pages, participants rated the similarity of the argument to four different landscape paintings on a scale from 0 ("Not at all similar") to 10 ("Very similar"), with these ratings converted to ranks for some analyses. Painting were presented on their own, without title or artist information. The paintings used were:

Yosemite	Looking Down Yosemite Valley, California [by Albert Bierstadt]
Rockies	A Storm in the Rocky Mountains, Mt. Rosalie [by Albert Bierstadt]
Suffolk	The Hay Wain [by John Constable]
Andes	The Heart of the Andes [by Frederic Edwin Church]

The arguments were presented in a random order. Paintings were blocked within each argument, presented on separate pages, and in a separate random order for each argument.

After the main task, a series of memory check questions was included to monitor whether participants had achieved basic comprehension of the materials (e.g., identifying that one of the arguments concerned the sum of numbers 1 through 100 but that none of the arguments were about properties of prime numbers). Participants were excluded from analysis if they incorrectly answered one-fourth or more of these questions (N = 67) or did not produce a complete set of ratings (N = 1). The same exclusion criterion was used in all studies, and the inattentive participants in all cases were distributed between the experienced and inexperienced groups in rough proportion to their size (for Studies 1–3 respectively, 57%, 71%, and 68% of excluded participants were inexperienced).

2.2. Results

Data for all studies can be accessed from the Open Science Framework at http://bit.ly/2GLwhEZ. The results of Study 1 are given in Table 1, computed for the Mechanical Turk sample as a whole. Table 1 presents (a) the average raw similarity scores for each painting–argument pair; (b) the ranking of these raw scores; (c) the average z-scored similarity score for each painting–argument pair, with the z-scores computed among each participant's 16 judgments to account for variability in how the scale is used; and (d) the frequency with which participants ranked each painting as the most similar to a given argument. The corresponding tables for professional mathematicians and for the Mechanical Turk subgroups with and without higher mathematics training are given in Appendix B.

Participants' similarity judgments were not random: They reflected some consensus about the similarity of mathematical arguments and artworks. The difference in z-scores between the highest and lowest ranked pairings was 0.85 standard deviations, which corresponds to a "large" effect using the standard criteria (Cohen, 1988). A second way of estimating the effect size is to calculate inter-rater reliability using Cronbach's alpha, treating each of the 16 ratings (i.e., 4 arguments × 4 artworks) as an observation and each participant as a scale component. These scores indicate a very high degree of consistency across participants [α = 0.93]. However, since these approaches do not allow us to calculate inferential statistics, it does not show conclusively that participants' internal consistency was above chance levels. We use three approaches to show that this consistency was nonrandom.

First, to test whether the distribution of similarity scores was nonrandom, we calculated for each participant the correlation between that

Table 1
Similarity judgments in Study 1 – Mechanical Turk sample.

	Yosemite	Rockies	Suffolk	Andes
Raw means				
Geometric	3.51	2.99	3.30	3.05
Gauss	2.38	2.23	2.43	1.96
Pigeonhole	2.42	2.21	2.25	2.49
Faulhaber	2.97	2.75	3.21	2.44
Ranks				
Geometric	16	12	15	13
Gauss	5	3	7	1
Pigeonhole	6	2	4	9
Faulhaber	11	10	14	8
Z-scored means				
Geometric	0.45	0.20	0.34	0.20
Gauss	−0.13	−0.26	−0.13	−0.40
Pigeonhole	–0.09	−0.29	−0.26	0.00
Faulhaber	0.23	0.04	0.29	−0.19
Frequency top-ranked				
Geometric	37%	19%	22%	22%
Gauss	27%	21%	33%	19%
Pigeonhole	28%	15%	20%	37%
Faulhaber	29%	21%	36%	14%

Note. Top panel: Raw mean similarity judgments. Second panel: Ranked similarity of each pair out of all 16 pairs (high numbers = more similar). Third panel: Mean similarity judgments, *z*-scored for each participant and then averaged (ignoring 9% of participants who entered identical scores for all items). Bottom panel: Proportion of participants ranking each artwork highest for a given argument (ignoring ties).

participant's 16 ranked similarity judgments and the ranked judgments averaging across the other participants (leaving out that participant from the sample).[1] These correlations were systematically positive [$p_{1t} < 0.001$, Wilcoxon test], reflecting the fact that positive correlations [$N = 156$] were much more common than negative correlations [$N = 77$], although the median correlation was modest in magnitude [$r_{med} = 0.22$].[2] This approach demonstrates that participants agreed with one another about the rankings of the 16 pairs. However, one may be concerned that this agreement might have been generated by some arguments appearing more "painting-like" than others; for example, the geometric series and Faulhaber arguments both present geometric proofs and might be seen as more similar to paintings simply because they are pictorial.

Second, to test whether the paintings were ranked nonrandomly even within each argument, we can look at the top-ranked artwork for each argument (ignoring ties), and test whether these top ranks were distributed randomly across artworks (see Table 1). They were not [$\chi^2(9) = 35.59$, $p_{2t} < 0.001$, $w = 0.24$]. Participants reliably believed the Yosemite painting to be most similar to the geometric series argument, the Suffolk painting most similar to the Faulhaber and Gauss's summation trick arguments, and the Andes painting most similar to the pigeonhole argument. It is not just that some arguments are more painting-like or some paintings are more math-like: People have specific intuitions about the match of specific mathematical arguments to specific artworks.

Third, we compared similarity rankings across participant groups to provide evidence of consensus across groups, as well as to check for any role of expertise. The experienced and inexperienced Mechanical Turk samples ranked the similarity of the 16 pairs in a remarkably consistent way [$r_s(14) = 0.89$, $p_{1t} < 0.001$], providing even further evidence for the consistency of these judgments (see Tables B1–2 in Appendix B). However, these judgments did not correlate significantly with the experts' (Table B3) when using Spearman correlations [$r_s(14) = 0.32$, $p_{1t} = 0.12$]. One possibility is that the experts' rankings were somewhat unstable due to the small sample size. This suspicion is supported by the significant *Pearson* correlations between these two groups [$r(14) = 0.51$, $p_{1t} = 0.023$], which looks at the mean judgments rather than ranks. Overall, the comparison to experts suggests some consistency between professional mathematicians and laypeople, although it seems likely that mathematical experience influences one's aesthetic judgments given the modest correlations across samples. In Study 2, we use a less-expert but more-accessible population of undergraduate math students to further test this issue of expertise with improved precision.

2.3. Discussion

These results show that laypeople come to considerable consensus in comparing mathematical arguments to artworks. This consensus was not driven merely by some arguments seeming particularly painting-like or some paintings seeming particularly mathematical, since different paintings were deemed most fitting for different arguments.

The question naturally arises of how people are making these judgments: What aesthetic criteria do they use, and indeed are they aesthetic at all? We turn to this question in Study 3. Before doing so, however, we seek to replicate this basic effect within another domain—music.

3. Study 2

Commentators have long noted an affinity between mathematics and music (Fauvel, Flood, & Wilson, 2006). Music has a mathematical structure, and the Greek mathematician Pythagoras worked out aspects of Western music theory that persist to this day. Study 2 tests whether the aesthetics of specific mathematical arguments intuitively correspond to different pieces of music, in keeping with this historically noted affinity and with the consensus observed for paintings in Study 1.

3.1. Methods

Participants were recruited from Mechanical Turk ($N = 299$, 51% female). Similar to Study 1, a subset of *experienced* participants ($N = 90$) had taken a higher mathematics course, while a larger subset of *inexperienced* participants had not ($N = 207$). Participants were excluded if they failed the same check questions used in Study 1 ($N = 73$).

Separate from the Mechanical Turk sample, a sample of Yale undergraduates ($N = 28$) was recruited from Mathematics and Applied Mathematics courses, including Calculus, Linear Algebra, Abstract Algebra, and Probability Theory. Students were contacted by email and we included all participants who completed the principal measures.

On the same 0–10 scale used in Study 1, participants rated the similarity of each argument to four 20-second clips of classical music (aside from sound itself, no information about the music was provided):

Schubert	Moment Musical No. 4, D 780 (Op. 94) [perf. David Fray]
Bach	Fugue from Toccata in E Minor (BWV 914) [perf. Glenn Gould]
Beethoven	Diabelli Variations (Op. 120) [perf. Grigory Sokolov]
Shostakovich	Prelude in D-flat major (Op.87 No. 15) [perf. Adrian Brendle]

The procedure was otherwise identical to Study 1.

[1] Where possible, we rely on nonparametric statistics to reflect the fact that participants may use the similarity scale in nonuniform and nonlinear ways. Therefore, we use Spearman correlations (computing the correlations between the ranks of two variables), one-sample Wilcoxon tests (comparing the distribution of ranks to chance), and chi-squared tests (comparing the distribution of top-ranked artworks to chance).

[2] We use one-tailed *p*-values for tests where consensus is a directional hypothesis, because our predictions concern consensus within groups, across groups, and across domains; the alternative to this prediction is random ranks, not systematically opposite ranks across individuals. (For example, it is implausible that the average person in a sample would produce a pattern of rankings that is opposite to the remainder of the sample.) We therefore indicate for each *p*-value whether it is two-tailed (p_{2t}) or one-tailed (p_{1t}), with this decision determined by the criterion above.

Table 2
Similarity judgments in Study 2 – Mechanical Turk.

	Schubert	Bach	Beethoven	Shostakovich
Raw means				
Geometric	4.76	4.39	4.36	4.62
Gauss	4.61	5.11	4.67	4.31
Pigeonhole	4.52	4.42	4.89	4.83
Faulhaber	4.32	4.59	5.04	5.06
Ranks				
Geometric	11	4	3	9
Gauss	8	16	10	1
Pigeonhole	6	5	13	12
Faulhaber	2	7	14	15
Z-scored means				
Geometric	0.03	−0.14	−0.12	−0.01
Gauss	−0.08	0.18	−0.03	−0.15
Pigeonhole	−0.06	−0.14	0.12	0.08
Faulhaber	−0.13	0.00	0.20	0.24
Frequency top-ranked				
Geometric	28%	22%	22%	29%
Gauss	17%	33%	27%	24%
Pigeonhole	23%	21%	31%	25%
Faulhaber	18%	25%	31%	26%

Note. Top panel: Raw mean similarity judgments. Second panel: Ranked similarity of each pair out of all 16 pairs (high numbers = more similar). Third panel: Mean similarity judgments, *z*-scored for each participant and then averaged (ignoring 3% of participants who entered identical scores for all items). Bottom panel: Proportion of participants ranking each artwork highest for a given argument (ignoring ties).

3.2. Results

The similarity ratings were much higher overall in Study 2 [$M = 4.66$] than in Study 1 [$M = 2.66$], consistent with the idea that music is imbued with a more mathematical character. Moreover, participants continued to associate different pieces of music with different arguments, albeit less robustly than they did with paintings in Study 1. Table 2 provides the descriptive statistics for Study 2, including raw means, ranked means, *z*-scored means, and the distribution of top ranks. Whereas Table 2 presents these statistics for the Mechanical Turk sample as a whole, Appendix C presents the comparable table for the Mechanical Turk subsamples with and without higher math training and for the Yale undergraduates.

Overall, the degree of consistency was more modest in Study 2, compared to Study 1. Whereas the largest difference in *z*-scores across pairings was 0.85 in Study 1, it was only 0.39 in Study 2, which corresponds traditionally to a small- to medium-sized effect (Cohen, 1988). Likewise, the inter-rater reliability was reasonably high [$\alpha = 0.72$], though less impressive compared to Study 1 [$\alpha = 0.93$].

We used the same three methods as in Study 1 to test whether this degree of consensus could be random; all three methods reject this null hypothesis. First, participants' ranking of the pairs correlated with the ranking produced by the other participants more often than chance [$p_{1t} < 0.001$, Wilcoxon test], although the median correlation was small in magnitude [$r_{med} = 0.09$].

Second, the distribution of top-ranked pieces of music differed systematically from chance [$\chi^2(9) = 18.66$, $p_{2t} = 0.028$, $w = 0.15$]. Analogous to Study 1, this result did not occur because participants uniformly favored certain pieces of music for all arguments. Instead, participants felt that the Beethoven pieces best fit the pigeonhole and Faulhaber arguments, that the Bach piece best fit Gauss's summation trick, and that Schubert and Shostakovich best fit the geometric series argument. Participants did not just see some music as more mathematical or some mathematics as more musical, but ascribed a unique aesthetic to each argument.

Third, the experienced and inexperienced Mechanical Turk subsamples were marginally correlated [$r_s(14) = 0.41$, $p_{1t} = 0.059$], as were the student sample and Mechanical Turk sample as a whole [$r_s(14) = 0.34$, $p_{1t} = 0.096$]. However, the students' rankings were more similar to the experienced than to the inexperienced Mechanical Turk subsample. Whereas the former subsample correlated significantly and strongly with students' ranks [$r_s(14) = 0.60$, $p_{1t} = 0.007$], the latter did not [$r_s(14) = 0.13$, $p_{1t} = 0.32$]. Thus, aesthetic intuitions about mathematical arguments may develop with training in higher mathematics.

3.3. Discussion

These results show that, just as people can form consistent intuitions about the aesthetic correspondences between visual art and mathematical arguments, they can also do so for pieces of classical piano music. Using four different analytical methods, participants' similarity judgments for pairs of music and mathematics exhibited some degree of consensus.

Mathematical arguments are more often associated with musical rather than visual aesthetics, and this was reflected in the higher overall similarity judgments in Study 2. Yet, the consensus about correspondences between particular musical pieces and mathematical arguments was less robust compared to the results of Study 1 for landscape paintings. Though somewhat counterintuitive, this is consistent with previous findings in the empirical aesthetics literature. Aesthetic preferences are known to be more consistent for more concrete rather than abstract stimuli (Vessel & Rubin, 2010), meaning that participants may have felt less confident in their aesthetic judgments of classical music rather than paintings. Exacerbating this is the fact that our American participants were likely relatively unfamiliar with classical music, along with the methodological limitation that paintings can be displayed all at once, while musical clips unfold over time. Participants may have not only had duller intuitions about the musical pieces, since they cannot be perceived at once, but also may not have listened to the entire clip on each trial. It is impressive that, despite these conceptual and methodological limitations, participants still evinced a statistically robust consensus.

Study 2 also provided some evidence that this consensus develops with mathematics expertise: The experienced Mechanical Turk subsample was more similar in their judgments to the undergraduate math students than either group was to the inexperienced Mechanical Turk subsample. The notion that expertise can influence aesthetic judgments is further bolstered by previous work in aesthetics, finding, for example, that photo professionals have higher ability to process photographs and consequently prefer more novel and uncertain photographs (Axelsson, 2007) and that art experts are likelier to experience interest and less likely to experience confusion when contemplating artworks (Silvia, 2013). Perhaps mathematical expertise similarly deepens aesthetic preferences in the mathematical domain, leading to more consistent responses.

4. Study 3

In Study 3, we use a different operationalization of aesthetic experience—ratings of beauty and other potentially aesthetic properties—to provide converging evidence for consistency in aesthetic intuitions about mathematics. We have so far shown systematic consistency in judgments of similarity between artworks and mathematical arguments. Such aesthetic categorizations support the idea that mathematical arguments are experienced in some dimensions analogous to artworks—that they are experienced aesthetically. However, these studies have not shown *how* people make these evaluations. In the empirical aesthetics literature, a prominent view holds that people make aesthetic judgments based on an artwork's *collative properties*, or higher-order properties relating to its arousal potential (e.g., complexity and novelty; Berlyne, 1960). How do such properties contribute to the aesthetics of mathematical arguments, and is their role in the mathematical domain similar to their role in the visual domain? Would

participants report consistent experiences of aesthetic pleasure (beauty) in mathematical arguments, as they make consistent aesthetic categorizations?

Study 3 also seeks to address an important concern about Studies 1 and 2: That the similarity judgments in those studies may not reflect specifically aesthetic correspondences between the artworks and arguments, but rather more superficial similarities. For example, in a pilot version of Study 1, one of the paintings we had chosen included a group of bears. Some participants reported in their written comments that they had paired that painting with the argument about the pigeonhole principle because that argument uses an analogy to a group of friends, which would create consistency across responses due to a highly superficial cue. We replaced this painting in the full version of Study 1, restricting ourselves to pure landscape paintings, and carefully scoured participant's comments for any further evidence of such thinking at an explicit level. Nonetheless, such superficial correspondences are difficult to rule out definitively on the basis of Studies 1 and 2 alone.

To address this problem as well as the question of what properties guide aesthetic judgments across the visual and mathematical domains, Study 3 asked participants to rate a variety of potentially aesthetic dimensions for both the paintings and the arguments used in Study 1, using a method similar to that of Inglis and Aberdein's (2015) study of professional mathematicians. In choosing these dimensions, we faced several constraints: They had to be appropriate as descriptions of artworks and of mathematical arguments, they had to be plausibly related to aesthetics, and they needed to be intuitive and evaluable by laypeople. One possibility would be to use collative properties measured in the empirical aesthetics literature (e.g., complexity). However, as descriptive terms, some important collative properties are either unclear in context to laypeople (e.g., unity) or awkward as descriptions of proofs (e.g., variety). Moreover, some higher-order dimensions considered anecdotally important to practicing mathematicians (e.g., profundity) would be overlooked by this approach. A second possibility would be to start with the dimensions previously tested in studies of mathematicians' proof appraisals (Inglis & Aberdein, 2015, 2016). However, this leads to the opposite problem: Some important dimensions for appraising proofs are non-aesthetic and do not make sense as appraisals of paintings (e.g., useful, applicable) and others would be difficult for laypeople to evaluate for proofs (e.g., rigorous).

Although one solution would be to mix and match dimensions from each of these two approaches, we instead chose an approach based on first principles: Mathematicians' own anecdotal reports of aesthetic pleasure from proofs. We derived 10 dimensions from the mathematician G. H. Hardy's (1940) famous discussion of mathematical beauty in *A Mathematician's Apology*. Hardy discusses six dimensions as essential to mathematical beauty—seriousness, generality, depth, unexpectedness, inevitability, and economy. For the first five dimensions, we used near-synonyms for some that would be more appropriate for artworks: *serious*, *universal*, *profound*, *novel*, and *clear*. For economy, we used four terms capturing different aspects: *simple*, *elegant*, *intricate*, and *sophisticated*.

4.1. Methods

Participants were recruited from Mechanical Turk (N = 300, 51% female), with a larger subsample that was inexperienced in higher mathematics (N = 194) and a smaller subset that was experienced (N = 104). Participants were excluded based on the same criteria used in previous studies (N = 94).

The main task had two parts, completed in a counterbalanced order. One part asked participants to consider each of the four paintings used in Study 1 and to rate these paintings on ten dimensions ("In your judgment, to what extent do the following descriptions apply to this painting?"): *beautiful*, *serious*, *universal*, *profound*, *novel*, *clear*, *simple*, *elegant*, *intricate*, *sophisticated*. These ratings were made in a random order for each painting and the paintings were presented in a random order. The other part asked participants to make these same ratings for the four arguments used in Studies 1 and 2.

4.2. Results

The mean judgments on each dimension are given in Table 3. The breakdown by expertise group (the subsamples with and without higher math experience) in Appendix D reveals that most aesthetic judgments—including beauty—were higher among the more expert subsample.

Overall, participants relied on similar aesthetic criteria for assessing beauty in both the artistic and mathematical domains. Moreover, these dimensions predicted the similarity ratings in Study 1.

Consistency of aesthetic intuitions. Participants made consistent judgments of each aesthetic judgment, both within the artistic and mathematical domains. To quantify consistency, we calculated interrater reliability using an analogous procedure to Studies 1 and 2, separately for the artworks and arguments. First, in each case we treated all forty judgments (10 judgments for each of the 4 paintings or arguments) as observations and each participant as a scale component. This revealed highly reliable judgments, both for artworks [α = 0.98] and arguments [α = 0.97]. Second, we investigated each aesthetic judgment separately, now treating each of the 4 paintings or arguments as observations. We find reliable judgments on most measures, for both paintings and arguments. In all cases, $\alpha > 0.70$, and usually much higher: beauty [α = 0.96 and α = 0.97, respectively], seriousness [α = 0.91 and 0.89], universality [α = 0.89 and 0.97], profundity [α = 0.98 and 0.93], novelty [α = 0.85 and 0.82], clarity [α = 0.94 and 0.99], simplicity [α = 0.92 and 0.98], elegance [α = 0.94 and 0.96], intricacy [α = 0.73 and 0.90], and sophistication [α = 0.95 and 0.81]. Thus, not only were beauty judgments highly reliable across participants, but judgments of many other potentially aesthetic properties were reliable too.

Aesthetic intuitions across domains and expertise. To test how participants judged beauty in each of these domains, we fit two hierarchical regression models—one for paintings and one for arguments—using judgments of beauty as the dependent variable and the nine aesthetic dimensions as predictors. Random intercepts were included for item and for each participant, to account for the repeated measures design. The coefficients are displayed in Table 4. The

Table 3
Aesthetic judgments in Study 3 – all participants.

	Yosemite	Rockies	Suffolk	Andes
Beauty	8.49	8.05	8.24	7.25
Seriousness	6.49	6.13	7.07	6.56
Universality	6.41	6.56	6.20	5.76
Profundity	6.83	6.24	6.93	5.25
Novelty	5.05	4.83	5.44	4.81
Clarity	7.27	7.23	6.97	6.26
Simplicity	4.03	3.97	3.20	4.27
Elegance	7.07	6.84	7.05	6.08
Intricacy	7.00	7.13	7.32	6.83
Sophistication	6.60	6.46	6.88	5.66
	Geometric	Gauss	Pigeonhole	Faulhaber
Beauty	3.95	3.54	2.56	4.27
Seriousness	5.88	5.96	5.14	6.05
Universality	6.88	5.56	5.10	6.72
Profundity	4.58	4.97	3.99	5.24
Novelty	4.36	4.95	4.24	4.85
Clarity	7.11	4.91	4.14	6.58
Simplicity	6.08	3.83	3.46	5.43
Elegance	4.04	3.72	3.02	4.68
Intricacy	4.80	5.87	5.48	5.05
Sophistication	4.74	5.05	4.68	5.44

Note. Entries are mean aesthetic judgments.

Table 4
Predictors of beauty judgments in Study 3 – all participants.

	Paintings	Arguments
Seriousness	−0.068 (0.025)**	0.027 (0.024)
Universality	0.073 (0.025)**	−0.009 (0.025)
Profundity	0.178 (0.028)***	0.104 (0.026)***
Novelty	−0.001 (0.026)	0.043 (0.024)°
Clarity	0.176 (0.027)***	0.063 (0.029)*
Simplicity	−0.062 (0.022)**	0.050 (0.027)°
Elegance	0.316 (0.033)***	0.679 (0.026)***
Intricacy	0.055 (0.030)°	0.024 (0.026)
Sophistication	0.039 (0.031)	−0.009 (0.027)°

Note. Entries are the coefficients (SEs) in a hierarchical regression predicting ratings of "beautiful" from ratings of all other attributes.
° $p_{2t} < 0.10$.
* $p_{2t} < 0.05$.
** $p_{2t} < 0.01$.
*** $p_{2t} < 0.001$.

corresponding analyses for the experienced and inexperienced Mechanical Turk subsamples are presented in Appendix D.

Overall, the three most important dimensions (adjusting for the others) for both artworks and arguments were elegance, followed by profundity, followed by clarity. (Elegance was the most important factor by far for arguments, whereas these factors were more equal for paintings.) Three other factors had strong relationships for paintings, but not arguments: Simplicity (negative for paintings, positive for arguments), universality (positive for paintings only), and seriousness (negative for paintings only). The other three factors (sophistication, intricacy, and novelty) played little role within either domain.

We can test whether participants used similar dimensions across domains by calculating the correlation between the two sets of coefficients. The Pearson correlation coefficient was significant [$r(7) = 0.75$, $p_{1t} = 0.010$]. However, given the very strong influence of elegance, it is useful to look at Spearman correlations as well, which consider only the rank-ordering of the coefficients. This correlation reaches marginal significance [$r_s(7) = 0.48$, $p_{1t} = 0.097$]. This similarity in rank-ordering between the two domains was driven by the experienced subsample [$r_s(7) = 0.53$, $p_{1t} = 0.074$], but not the inexperienced subsample [$r_s(7) = 0.23$, $p_{1t} = 0.28$]. This further bolsters the case that intuitions about mathematical beauty sharpen with experience in higher mathematics.

To look into the expertise issue in more detail, we can compare the coefficients across expertise groups. For paintings, the groups with and without higher mathematics background used remarkably similar weights, with simplicity, profundity, and clarity the most important predictors for both groups. Indeed, the coefficients were very highly correlated [$r_s(7) = 0.92$, $p_{1t} < 0.001$]. However, for the mathematical arguments, the two groups used quite different criteria. Whereas the group with higher mathematics background relied mainly on profundity, clarity, and elegance (as they did for paintings), the group without such background relied almost exclusively on elegance (as well as a smaller weight on sophistication), leading to a nonsignificant correlation between the two groups' coefficient rankings [$r_s(7) = 0.07$, $p_{1t} = 0.44$]. Thus, higher mathematics training sharpens aesthetic intuitions in a domain-specific way, with more consistent intuitions about mathematical beauty but intuitions about artistic beauty similar to those of untrained novices.

Using aesthetic intuitions to predict similarity judgments. Since participants in Study 1 gave similarity judgments for each pair of painting and argument, we can use participant's judgments about the aesthetic qualities of each painting and argument from Study 3 to model these judgments. To do so, we computed the pairwise similarities using the Study 3 data, in three different ways, in all cases using Euclidean distance to calculate (dis)similarity.

A simple first approach was to calculate the Euclidean distance between the beauty judgments for each pair:

$$\sqrt{(B_{art} - B_{arg})^2} = |B_{art} - B_{arg}|$$

where B_{art} and B_{arg} represent beauty judgments for artworks and arguments, respectively. We then converted these to ranks, which are displayed in Table 5. These ranks were marginally correlated with the similarity rankings from Study 1 [$r_s(14) = 0.36$, $p_{1t} = 0.088$], suggesting that the Study 3 data have some out-of-sample predictive power for the Study 1 similarity scores. Although this measure has the advantage of being straightforward and capturing aesthetic pleasure in its purest form, it collapses all of aesthetics into a single dimension. This is limiting, since something can be beautiful in one sense but not another, and since aesthetic experience is broader than aesthetic pleasure. For example, Gödel's incompleteness theorems might be considered very high on profundity (indeed, they fundamentally challenged the philosophical foundations of mathematics) but very low on simplicity (their proofs are extremely complicated).

Table 5
Similarity rankings computed from Study 3 data.

	Yosemite	Rockies	Suffolk	Andes
Similarity ranking from beauty judgments				
Geometric	10	6	8	2
Gauss	13	9	12	3
Pigeonhole	16	14	15	11
Faulhaber	7	4	5	1
Similarity ranking from elegance, profundity, and clarity				
Geometric	8	7	9	2
Gauss	13	11	12	4
Pigeonhole	16	14	15	10
Faulhaber	6	3	5	1
Similarity ranking weighted by coefficients				
Geometric	9	7	8	2
Gauss	13	11	12	4
Pigeonhole	16	14	15	10
Faulhaber	6	3	5	1

Note. Entries are the ranked similarity scores using the three similarity metrics described in the main text, computed using Study 3 data. Higher numbers indicate higher similarity.

A second approach which allows for similarity scores that account for multiple dimensions was to rely on the three dimensions that were important predictors of beauty judgments, both for artworks and for arguments: elegance, profundity, and clarity. For each pair, we computed:

$$\sqrt{(E_{art} - E_{arg})^2 + (P_{art} - P_{arg})^2 + (C_{art} - C_{arg})^2}$$

where E_{art} and E_{arg} represent elegance judgments, P_{art} and P_{arg} represent profundity judgments, and C_{art} and C_{arg} represent clarity judgments. The resulting similarity rankings (see Table 5) correlated significantly with participants' similarity judgments from Study 1 [$r_s(14) = 0.44$, $p_{1t} = 0.046$]. This ranking has the benefit of taking account of multiple dimensions, but arbitrarily applies equal weight to all dimensions.

A third approach which allows for the weight to vary across dimensions is to compute a vector of weights **w**, where each entry in the vector w_i is a weight to be applied to each of the nine aesthetic dimensions, by multiplying the two coefficients for that dimension, for artworks and arguments [$b_{i,art} * b_{i,arg}$], found in the top panel of Table 5. This approach weights dimensions to the extent that they have predictive power (in the same direction) for both artworks and arguments. These weights were then applied to the nine judgment dimensions:

$$\sqrt{\sum_{i=1}^{9} (D_{i,art} - D_{i,arg})^2 \cdot w_i}$$

where $D_{i,art}$ and $D_{i,arg}$ represent the ith of the nine dimensions,

respectively for artworks and arguments. Once again, these similarity rankings (shown in Table 5) were significantly correlated with the similarity rankings made out-of-sample in Study 2 [$r_s(14) = 0.44$, $p_{1t} = 0.047$].

4.3. Discussion

Study 3 builds on the findings of Studies 1 and 2 in several respects. First, participants relied principally on the same three dimensions for evaluating artistic and mathematical beauty—elegance, profundity, and clarity. This further supports the notion that people have aesthetic intuitions about mathematics which are related to our broader aesthetic sensibilities. Second, intuitions about these aesthetic dimensions sharpened with mathematics training, consistent with evidence from Studies 1 and 2 that people develop expertise in these aesthetic judgments. Third, the individual aesthetic judgments of artworks and arguments in Study 3 could be used to model the similarity judgments made out-of-sample in Study 1, lending further support to the claim that the Study 1 judgments were indeed judgments of aesthetic similarity, rather than more superficial features.

Although we did not test professional mathematicians in Study 3, we can compare our results to Inglis and Aberdein (2015), who asked professionals to assess mathematical arguments on 80 different dimensions. In fact, of the remaining 79 dimensions, elegance was the strongest predictor of beauty within their sample too, with judgments of profundity and clarity also reaching significance. Simplicity and intricacy were not significantly associated with beauty among the professionals in their study, consistent with our finding of at most a weak relationship between beauty and these dimensions. Once again, this suggests some commonality among the dimensions used by (experienced) novices and experts, with the qualification that our inexperienced subsample relied only on elegance among these dimensions. In fact, this is the same pattern we found in Study 2, where student "experts" had similar judgments to the experienced subsample but not the inexperienced subsample. Even a modest dose of higher math training appears to develop aesthetic intuitions in the direction of professional mathematicians, consistent with research in categorization, which finds that expertise alters the dimensions we use in categorization (Chi et al., 1981; Medin et al., 1997), and in empirical aesthetics, which finds that experts experience more nuanced aesthetic emotions (Fayn et al., 2018).

5. General discussion

Paul Erdős used to say that a mathematician did not need to believe in God, but *did* need to believe in The Book—a platonic collection of mathematical statements with their most beautiful proofs. Mathematicians to this day will colloquially refer to particularly beautiful proofs as "being form The Book." When two mathematicians, Aigner and Ziegler (1998), published a "first approximation" to The Book, containing a collection of beautiful statements with particularly beautiful proofs, the result was a bestseller in the mathematical community.

Can laypeople, like mathematicians, discern which proofs belong in The Book and which do not? Our studies demonstrate that, when mathematical arguments are stated in a simple form, laypeople have consistent intuitions about the aesthetics of mathematics. Participants' ratings of how mathematical arguments corresponded to different artworks had strong internal consistency, as well as some relationship to those of experts. This was true for landscape paintings (Study 1) and classical music (Study 2). Moreover, people relied on several of the same dimensions for evaluating mathematical and artistic beauty (Study 3), with the pairwise similarity between aesthetic ratings predicting judgments of similarity in a separate sample. Mathematical beauty, then, does not appear to be solely in the eye of the beholder but appears to have deep psychological roots.

Several results spoke to the role of expertise in aesthetic judgments. In Study 1, lay participants' judgments were weakly correlated with expert mathematicians' judgments (only reaching significance using Pearson correlations). This may suggest a shift in aesthetic beliefs associated with expertise, although caution is warranted given the small sample of professionals. In Study 2, laypeople with experience in higher mathematics were more similar to math undergraduates than either group was to laypeople without such expertise, suggesting a shift due to expertise in a more convincing way. In Study 3, laypeople with and without higher mathematics experience used very similar dimensions for judging the aesthetics of paintings, but differed wildly in how they judged mathematical arguments, once again speaking to a shift with expertise. Overall, expertise appears to shift the dimensions along which mathematical beauty is computed, much as expertise shapes our categorization schemes and similarity judgments in other domains (Chi et al., 1981; Medin et al., 1997) and aesthetic expertise specifically shapes aesthetic experience (Fayn et al., 2018; Leder et al., 2012, 2014).

5.1. Relation to cognitive science broadly

The results of these studies inform several areas of inquiry in cognitive science, including empirical aesthetics, expertise, education, reasoning, and the philosophy of mathematics.

First, within empirical aesthetics, these results contribute to the research tradition documenting aesthetic commonalities across individuals. For instance, people share aesthetic knowledge within a culture for geometric figures (Westphal-Fitch & Fitch, 2017). Moreover, these interpersonally shared intuitions are largely automatic: People can form stable aesthetic judgments about paintings on time-scales as short as 50 ms (Verhavert, Wagemans, & Augustin, 2018). Evolutionary accounts of aesthetics (e.g., Dutton, 2009; Hekkert, 2014; Pinker, 1997) have typically focused on aesthetic perception, grounded in the idea that aesthetic pleasure draws us toward adaptively beneficial objects, such as resource-rich savannahs, reproductively fit mates, and readily processed visual stimuli. The current results suggest that we need a broader conceptualization to capture the full range of human aesthetic experiences, including the aesthetics of ideas. We think an enriched evolutionary approach, accounting for the adaptiveness of aesthetically appealing ideas, is a promising avenue for future research (see Hekkert, 2014 and "Future Research" below).

Second, within the psychology of expertise, these results contribute to the longstanding question of what distinguishes experts from laypeople within a domain. For example, chess experts are not distinguished from non-experts so much in terms of their calculating speed or ability to look ahead large numbers of moves, but instead in their ability to recall and reason about a large number of known chess positions (Chase & Simon, 1973). In fact, chess experts even use the fusiform face area to recognize naturalistic (but not scrambled) chess positions (Bilalic, Langner, Ulrich, & Grodd, 2011). The current results suggest that expertise is not only associated with different quantities and types of knowledge, but also with different aesthetic preferences. That is, while the judgments of experts usually bore some similarity to those of laypeople, the balance of the evidence suggests a shift in what criteria are used to evaluate the beauty of mathematical arguments.

The role of aesthetic judgments and domain expertise would be interesting to investigate in other domains, such as chess, where other aspects of expert development are better-understood. Indeed, aesthetic judgments about chess are common among chess experts (Margulies, 1977). The writer Vladimir Nabokov even published *Poems and Problems*—a book containing 53 poems and 18 chess problems—stating that "Chess problems demand from the composer the same virtues that characterize all worthwhile art: originality, invention, conciseness, harmony, complexity and splendid insincerity" (Nabokov, 1969). Perhaps aesthetic intuitions play an aesthetic role in chess play, as some mathematicians have reported that they do in guiding their

mathematics.

Third, within educational psychology, these results might inform debates about the teaching of mathematics and the introduction of rigorous proof. Students appear to struggle with proof because of unfamiliarity with concepts, the difficulty of mathematical notation and language, and uncertainty about where to start (Moore, 1994). It is not unreasonable to introduce rigorous proof in the context of geometry, which is laden with less mathematical notation than other areas of mathematics. But if adults' shared sense of mathematical aesthetics is shared in adolescence or earlier (an open question), educators may be missing opportunities to capitalize on students' aesthetic sensibilities. Perhaps students introduced to formal proof with aesthetically pleasing arguments would not only be likelier to understand them, but also develop an affection for mathematics (see Sinclair, 2004).

Fourth, within the psychology of reasoning, these results speak to the role of aesthetic considerations in guiding inference. Some have argued that people infer that an explanation is true when it strikes them as elegant, beautiful, or satisfying—that is, aesthetically pleasing (Johnson, 2017; Lipton, 2004). This fits anecdotal evidence from scientists and mathematicians who claim to use beauty as a guide to the truth (e.g., Dirac, 1963). And indeed, previous studies do suggest that people find explanations likely to the extent that they are satisfying (Khemlani et al., 2011; Lombrozo, 2007) and that satisfying explanations are often "truth-tracking" in conforming to the laws of probability (Johnson et al., 2016; Johnson, Johnston, et al., 2017; Johnson, Valenti, et al., 2017; Johnston, Johnson, Koven, & Keil, 2017). However, the current studies are the first, to our knowledge, to directly demonstrate an aesthetic component to lay explanation by measuring beauty directly or correspondences to aesthetic objects.

Finally, and most directly, these results inform the debate about the aesthetics of mathematical arguments, which has long raged among philosophers and mathematicians. Many mathematicians have claimed that aesthetic beauty is an objective property of mathematical arguments (e.g., Hardy, 1940; Poincaré, 2007/1914; Tao, 2007). On the other hand, some writers have argued that statements made about mathematical beauty are really about the truth or utility of the underlying argument (e.g., Harré, 1958; Rota, 1997; Todd, 2008). One piece of evidence against the latter view is that professional mathematicians' appraisals of proofs are characterized by several dimensions, including aesthetics, intricacy, utility, and precision, with aesthetic components strongly affecting judgments of beauty (Inglis & Aberdein, 2015). The current evidence suggests that even among laypeople, proof appraisals have an aesthetic component akin to the aesthetic evaluation of artworks, almost as though they treat proofs *as* artworks, and indeed Study 3 found that mathematically experienced laypeople relied on many of the same dimensions as the experts in Inglis and Aberdein (2015).

5.2. Future research

Cognitive mechanisms. Although the current results suggest that people can experience mathematical arguments aesthetically, there are lingering questions about the extent and cognitive underpinnings of this ability. For example, all of our arguments were conventionally "beautiful" proofs. Would laypeople also be able to distinguish between beautiful and ugly proofs, as mathematicians seem to (Wells, 1990)?

Moreover, these studies do not tell us what cues people rely on to judge the underlying dimensions of their proof appraisals, such as elegance. There is research on how simplicity guides explanatory judgments broadly (Bonawitz & Lombrozo, 2012; Johnson, Johnston, et al., 2017; Johnson, Valenti, et al., 2017; Lombrozo, 2007), though much less in the mathematical domain. But very little is known about how people judge aesthetic qualities such as elegance or profundity. What drives such judgments? Both emotions versus semantic content seem to play roles in guiding aesthetic judgments in different domains (Brielmann & Pelli, 2017; Palmer et al., 2013; Vessel & Rubin, 2010). In other domains, such as photography, semantic content is known to be an important driver of aesthetic preferences. Eliciting more multi-dimensional ratings from participants (e.g., Blijlevens et al., 2017; Inglis & Aberdein, 2015) would help to uncover the finer basis of these judgments in affective or cognitive processes.

Other kinds of data could also help to illuminate the cognitive underpinnings of both mathematical aesthetic experiences and aesthetic experiences of ideas more broadly. First, developmental data could be useful in pinpointing which cognitive abilities are required for aesthetic experiences of mathematics, such as explanatory reasoning abilities (Bonawitz & Lombrozo, 2012; Johnston et al., 2017). More provocatively, perhaps the developmental trajectory is reversed: Maybe broader explanatory reasoning abilities are dependent on aesthetic experiences. In that case, perhaps aesthetic preferences can be harnessed to facilitate math education, as suggested by Sinclair (2004). Second, brain data could useful for further understanding the relationship between novices' and experts' aesthetic experiences of mathematics, following on studies of the neural correlates of mathematicians' aesthetic judgments of proofs (Zeki et al., 2014) and developmental changes in the neural correlates of music appreciation (Nieminen, Istók, Brattico, Tervaniemi, & Huotilainen, 2011).

Consensus and expertise. These studies do not tell us whether the moderate degree of consensus we see among novices reflects an underlying Platonic sense of beauty. While this is ultimately a philosophical question, further empirical evidence could be relevant to the normative debate. For example, Inglis and Aberdein (2016) find that professional mathematicians experience a surprisingly high degree of *disagreement* in their appraisals of at least one proof, in apparent contrast to our finding that laypeople experience a surprisingly high degree of *agreement*. This paradox may be more apparent than real, if we expect a very high degree of consensus from mathematicians and low degree of consensus from laypeople, but in fact find a moderate degree of consensus among both groups. It is also possible that some proofs are divisive for mathematicians (e.g., "clever" proofs that elegantly prove some theorem while providing relatively little illumination), whereas others are widely agreed as beautiful or ugly. Both of these possibilities are not testable with our current data, since our sample of professional mathematicians was small and Inglis and Aberdein (2016) tested only one argument. Further studies testing a larger sample of mathematicians and laypeople on an expanded set of arguments could help to resolve this issue.

Cross-cultural data could illuminate the related issue of (dis)agreement across cultures. In some cases, aesthetic preferences do appear to be similar across cultures (e.g., Palmer et al., 2013). However, given that culture shapes approaches to logical reasoning (Peng & Nisbett, 1999) and categorization (Medin & Atran, 2004), might people from different cultures have different aesthetic sensibilities about mathematical arguments? A negative answer would tend to support a more Platonic view of mathematical aesthetics, whereas a positive answer would support a more subjectivist view. Our suspicion is that there is substantial cross-cultural consensus, in keeping with the idea that aesthetic preferences in mathematics are rooted in broader evolved capacities. This would also be in keeping with anecdotes about some mathematicians with little formal training in Western mathematics, such as Srinivasa Ramanujan (Kanigal, 2016). Ramanujan discovered mathematics widely perceived as both strikingly original and deeply beautiful by the world mathematical community, some of these discoveries pointing toward deep and mysterious structures that are still not fully understood. Still, it is entirely possible that such aesthetic judgments manifest themselves differently across cultures, just as there is variability in musical preferences across cultures despite the apparent existence of some evolved aspects of our musical faculties (Hauser & McDermott, 2003).

The aesthetics of ideas. Finally, we hope that this work serves as a broad call to action in investigating the aesthetics of abstract ideas. If people experience aesthetic pleasure and make reliable aesthetic

judgments about a domain as abstract as mathematics, it is likely that aesthetic experiences in less abstract domains of ideas would be even more powerful. Scientists often invoke aesthetic considerations in justifying their preference for one theory over another, which perhaps helps to guide scientists toward truth (if we believe Dirac, Einstein, and Keats). Social engineers may experience aesthetic pleasure at contemplating particularly beautiful social arrangements, and such considerations may play a role in political decision-making, particularly in regimes unconstrained by democratically expressed preferences. Consumers may be attracted to particular ideas expressed in the marketplace, especially in the less tangible domain of services, complementing insights about physical products in consumer behavior and product design (e.g., Desmet & Hekkert, 2007; Hoyer & Stokburger-Sauer, 2012). Understanding the aesthetics of ideas could prove to be a critical link between the affective and cognitive sciences.

Acknowledgements

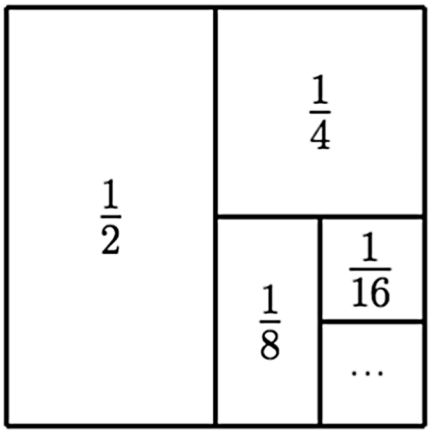
We thank Woo-kyoung Ahn, Lance Rips, and Nadya Vasilyeva for discussion, and Asher Auel, Miki Havlickova, John Hall, Brett Smith, and Sarah Vigliotta for assistance with data collection. These studies were presented at the 40th Conference of the Cognitive Science Society and a subset of the findings were discussed in an informal commentary in the *Mathematical Intelligencer* magazine; we thank the conference attendees and *Intelligencer* editor for their interest (Johnson & Steinerberger, 2018, 2019). This work was supported by a grant from the American Psychological Association awarded to the first author.

Appendix A. Text of arguments

Geometric: Sum of an infinite geometric series

$$\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+\frac{1}{32}+\cdots=1.$$

We can see this by cutting a square with total area 1 into little pieces.



Gauss: Gauss's summation trick for positive integers[3]

A quick way of computing

$$1+2+3+4+\cdots+98+99+100=5050$$

is as follows: write the total sum twice and add the columns

$$\begin{array}{rrrrrrrrrr} 1+ & 2+ & 3+ & 4+\cdots+ & 98+ & 99+ & 100 \\ 100+ & 99+ & 98+ & 98+\cdots+ & 3+ & 2+ & 1 \\ \hline 101 & 101 & 101 & 101 & \ldots & 101 & 101 & 101 \end{array}$$

This yields a total of 100 times 101 (giving 10100) and half of that is exactly 5050.

[3] The minor typo in this item (can you spot it?) was present in the experimental materials. However, it is unlikely that this substantially influenced the results of the study, as (a) it scores as well as or better than the other proofs on the dimensions measured in Study 3, and (b) Study 1 participants had the opportunity to write comments at debriefing, and no participant mentioned the typo.

Pigeonhole: Pigeonhole principle

In any group of 5 people, there are two who have the same number of friends within the group.

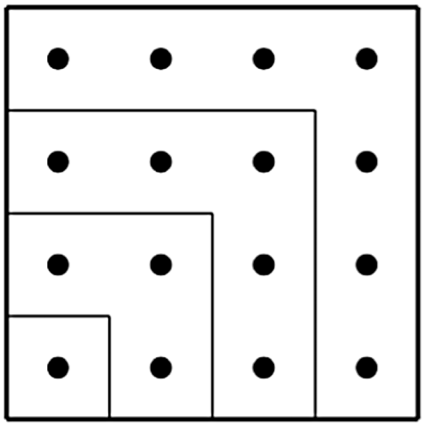
We can see this as follows: suppose there exist somebody who is friends with everybody else. Then every person in the group has either 1,2,3 or 4 friends. Since there are 5 people, one number has to appear twice. If nobody is friends with everybody else, then everybody has either 0,1,2 or 3 friends; again, since there are 5 people, one number has to appear twice.

Faulhaber: Geometric proof of a Faulhaber formula

The sum of consecutive odd numbers always adds up to a square number:

$$1 = 1^2$$
$$1 + 3 = 2^2$$
$$1 + 3 + 5 = 3^2$$
$$1 + 3 + 5 + 7 = 4^2$$

The reason is explained in the picture below: adding the next odd number creates a suitable layer for the next square.



Appendix B. Study 1 results by subgroup

See Tables B1–B3.

Table B1
Similarity judgments in Study 1 – Mechanical Turk sample *without* higher mathematics training.

	Yosemite	Rockies	Suffolk	Andes
Raw means				
Geometric	3.41	2.88	3.08	2.88
Gauss	2.31	2.15	2.28	1.88
Pigeonhole	2.23	1.99	2.09	2.37
Faulhaber	2.77	2.64	3.06	2.35
Ranks				
Geometric	16	12.5	15	12.5
Gauss	7	4	6	1
Pigeonhole	5	2	3	9
Faulhaber	11	10	14	8
Z-scored means				
Geometric	0.48	0.23	0.33	0.20
Gauss	−0.12	−0.24	−0.12	−0.38
Pigeonhole	−0.10	−0.38	−0.31	−0.02
Faulhaber	0.19	0.09	0.33	−0.17
Frequency top-ranked				
Geometric	39%	18%	23%	20%

(*continued on next page*)

Table B1 (*continued*)

	Yosemite	Rockies	Suffolk	Andes
Gauss	32%	18%	31%	19%
Pigeonhole	31%	11%	19%	39%
Faulhaber	28%	23%	36%	14%

Note. Top panel: Raw mean similarity judgments. Second panel: Ranked similarity of each pair out of all 16 pairs (high numbers = more similar). Third panel: Mean similarity judgments, *z*-scored for each participant and then averaged (ignoring participants who entered identical scores for all items). Bottom panel: Proportion of participants ranking each artwork highest for a given argument (ignoring ties).

Table B2
Similarity judgments in Study 1 – Mechanical Turk sample *with* higher mathematics training.

	Yosemite	Rockies	Suffolk	Andes
Raw means				
Geometric	3.76	3.27	3.79	3.44
Gauss	2.55	2.40	2.77	2.13
Pigeonhole	2.85	2.71	2.62	2.75
Faulhaber	3.45	3.01	3.56	2.66
Ranks				
Geometric	15	11	16	12
Gauss	3	2	8	1
Pigeonhole	9	6	4	7
Faulhaber	13	10	14	5
Z-scored means				
Geometric	0.36	0.14	0.35	0.20
Gauss	−0.15	−0.31	−0.13	−0.43
Pigeonhole	−0.06	−0.11	−0.16	0.04
Faulhaber	0.32	−0.06	0.20	−0.24
Frequency top-ranked				
Geometric	31%	24%	18%	27%
Gauss	19%	25%	38%	19%
Pigeonhole	23%	21%	21%	35%
Faulhaber	31%	19%	35%	15%

Note. Top panel: Raw mean similarity judgments. Second panel: Ranked similarity of each pair out of all 16 pairs (high numbers = more similar). Third panel: Mean similarity judgments, *z*-scored for each participant and then averaged (ignoring participants who entered identical scores for all items). Bottom panel: Proportion of participants ranking each artwork highest for a given argument (ignoring ties).

Table B3
Similarity judgments in Study 1 – mathematicians.

	Yosemite	Rockies	Suffolk	Andes
Raw means				
Geometric	4.19	3.30	3.20	2.17
Gauss	3.40	3.46	2.84	2.34
Pigeonhole	2.27	2.59	2.31	2.08
Faulhaber	4.19	3.61	2.95	2.81
Ranks				
Geometric	15.5	11	10	2
Gauss	12	13	8	5
Pigeonhole	3	6	4	1
Faulhaber	15.5	14	9	7
Z-scored means				
Geometric	0.67	0.06	−0.02	0.02
Gauss	0.19	0.31	−0.16	−0.38
Pigeonhole	−0.55	−0.14	−0.35	−0.57
Faulhaber	0.56	0.48	0.02	−0.14
Frequency top-ranked				
Geometric	29%	14%	14%	43%
Gauss	50%	25%	12%	12%
Pigeonhole	25%	12%	25%	38%
Faulhaber	38%	12%	12%	38%

Note. Top panel: Raw mean similarity judgments. Second panel: Ranked similarity of each pair out of all 16 pairs (high numbers = more similar). Third panel: Mean similarity judgments, *z*-scored for each participant and then averaged (ignoring participants who entered identical scores for all items). Bottom panel: Proportion of participants ranking each artwork highest for a given argument (ignoring ties).

Appendix C. Study 2 results by subgroup

See Tables C1–C3.

Table C1
Similarity judgments in Study 2 – Mechanical Turk sample *without* higher mathematics training.

	Schubert	Bach	Beethoven	Shostakovich
Raw means				
Geometric	4.66	4.38	4.49	4.79
Gauss	4.47	5.01	4.63	4.29
Pigeonhole	4.33	4.35	4.98	4.96
Faulhaber	4.53	4.58	5.07	5.19
Ranks				
Geometric	10	4	6	11
Gauss	5	14	9	1
Pigeonhole	2	3	13	12
Faulhaber	7	8	15	16
Z-scored means				
Geometric	−0.03	−0.17	−0.07	0.06
Gauss	−0.14	0.12	−0.03	−0.16
Pigeonhole	−0.13	−0.20	0.18	0.12
Faulhaber	−0.07	0.00	0.22	0.29
Frequency top-ranked				
Geometric	24%	21%	25%	30%
Gauss	15%	32%	27%	26%
Pigeonhole	18%	19%	36%	27%
Faulhaber	20%	23%	29%	28%

Note. Top panel: Raw mean similarity judgments. Second panel: Ranked similarity of each pair out of all 16 pairs (high numbers = more similar). Third panel: Mean similarity judgments, *z*-scored for each participant and then averaged (ignoring participants who entered identical scores for all items). Bottom panel: Proportion of participants ranking each artwork highest for a given argument (ignoring ties).

Table C2
Similarity judgments in Study 2 – Mechanical Turk sample *with* higher mathematics training.

	Schubert	Bach	Beethoven	Shostakovich
Raw means				
Geometric	5.04	4.43	4.11	4.30
Gauss	4.91	5.32	4.71	4.32
Pigeonhole	4.97	4.60	4.73	4.58
Faulhaber	3.91	4.61	5.06	4.78
Ranks				
Geometric	14	5	2	3
Gauss	12	16	9	4
Pigeonhole	13	7	10	6
Faulhaber	1	8	15	11
Z-scored means				
Geometric	0.18	−0.08	−0.23	−0.14
Gauss	0.05	0.29	−0.04	−0.14
Pigeonhole	0.07	−0.01	0.01	−0.01
Faulhaber	−0.26	−0.01	0.19	0.12
Frequency top-ranked				
Geometric	35%	22%	17%	25%
Gauss	21%	33%	25%	21%
Pigeonhole	31%	25%	22%	22%
Faulhaber	16%	27%	36%	21%

Note. Top panel: Raw mean similarity judgments. Second panel: Ranked similarity of each pair out of all 16 pairs (high numbers = more similar). Third panel: Mean similarity judgments, *z*-scored for each participant and then averaged (ignoring participants who entered identical scores for all items). Bottom panel: Proportion of participants ranking each artwork highest for a given argument (ignoring ties).

Table C3
Similarity judgments in Study 2 – yale math undergraduates.

	Schubert	Bach	Beethoven	Shostakovich
Raw means				
Geometric	5.17	5.86	4.72	4.78
Gauss	5.21	5.51	4.76	5.20
Pigeonhole	4.99	4.96	5.09	4.76
Faulhaber	4.04	4.34	6.01	5.11
Ranks				
Geometric	11	15	3	6
Gauss	13	14	4.5	12
Pigeonhole	8	7	9	4.5
Faulhaber	1	2	16	10
Z-scored means				
Geometric	0.08	0.38	−0.13	−0.13
Gauss	0.07	0.20	−0.10	0.07
Pigeonhole	0.01	0.15	0.10	−0.13
Faulhaber	−0.37	−0.42	0.51	0.01
Frequency top-ranked				
Geometric	18%	39%	18%	25%
Gauss	26%	26%	30%	19%
Pigeonhole	12%	28%	44%	16%
Faulhaber	11%	15%	41%	33%

Note. Top panel: Raw mean similarity judgments. Second panel: Ranked similarity of each pair out of all 16 pairs (high numbers = more similar). Third panel: Mean similarity judgments, z-scored for each participant and then averaged (ignoring participants who entered identical scores for all items). Bottom panel: Proportion of participants ranking each artwork highest for a given argument (ignoring ties).

Appendix D. Study 3 results by subgroup

See Tables D1–D3.

Table D1
Aesthetic judgments in Study 3 – Mechanical Turk sample *without* higher mathematics training.

	Yosemite	Rockies	Suffolk	Andes
Beauty	8.46	7.97	8.27	7.06
Seriousness	6.37	6.10	7.08	6.64
Universality	6.22	6.40	6.05	5.59
Profundity	6.65	6.10	6.81	4.96
Novelty	4.67	4.53	5.24	4.51
Clarity	7.38	7.35	6.98	6.04
Simplicity	3.88	3.94	2.97	4.23
Elegance	6.98	6.65	6.94	5.71
Intricacy	6.94	7.06	7.43	6.79
Sophistication	6.46	6.27	6.64	5.26
	Geometric	Gauss	Pigeonhole	Faulhaber
Beauty	3.52	3.12	2.05	3.93
Seriousness	5.79	5.89	5.00	6.04
Universality	6.70	5.43	4.72	6.50
Profundity	4.22	4.78	3.83	5.11
Novelty	4.07	4.67	3.94	4.60
Clarity	6.93	4.56	3.54	6.40
Simplicity	5.87	3.41	2.88	5.26
Elegance	3.52	3.26	2.42	4.32
Intricacy	4.51	5.97	5.58	5.00
Sophistication	4.31	4.89	4.38	5.25

Note. Entries are mean aesthetic judgments.

Table D2
Aesthetic judgments in Study 3 – Mechanical Turk sample *with* higher mathematics training.

	Yosemite	Rockies	Suffolk	Andes
Beauty	8.52	8.17	8.17	7.61
Seriousness	6.72	6.15	7.01	6.35
Universality	6.73	6.84	6.46	6.06
Profundity	7.14	6.51	7.14	5.80
Novelty	5.78	5.45	5.83	5.40
Clarity	7.06	7.01	6.98	6.73
Simplicity	4.27	4.00	3.61	4.34
Elegance	7.26	7.27	7.31	6.81
Intricacy	7.14	7.33	7.16	6.90
Sophistication	6.91	6.89	7.43	6.47
	Geometric	Gauss	Pigeonhole	Faulhaber
Beauty	4.82	4.38	3.57	4.99
Seriousness	6.03	6.03	5.36	6.02
Universality	7.18	5.76	5.77	7.10
Profundity	5.23	5.34	4.26	5.45
Novelty	4.93	5.46	4.75	5.37
Clarity	7.44	5.54	5.25	6.89
Simplicity	6.48	4.59	4.53	5.73
Elegance	5.10	4.67	4.20	5.45
Intricacy	5.35	5.67	5.24	5.17
Sophistication	5.52	5.38	5.30	5.82

Note. Entries are mean aesthetic judgments.

Table D3
Predictors of beauty in Study 3 by subgroup.

	Paintings	Arguments
Mechanical Turk (without higher math)		
Seriousness	−0.027 (0.044)	0.015 (0.047)
Universality	0.123 (0.042)**	0.007 (0.052)
Profundity	0.145 (0.045)**	0.027 (0.053)
Novelty	−0.027 (0.042)	0.072 (0.043)°
Clarity	0.187 (0.045)***	0.041 (0.059)
Simplicity	−0.048 (0.035)	0.029 (0.051)
Elegance	0.266 (0.055)***	0.683 (0.051)***
Intricacy	0.019 (0.049)	0.060 (0.046)
Sophistication	0.007 (0.051)	0.113 (0.050)*
Mechanical Turk (with higher math)		
Seriousness	−0.081 (0.030)**	0.041 (0.028)
Universality	0.057 (0.032)°	−0.012 (0.028)
Profundity	0.180 (0.034)***	0.133 (0.030)***
Novelty	−0.008 (0.033)*	0.023 (0.028)
Clarity	0.170 (0.035)***	0.072 (0.033)*
Simplicity	−0.008 (0.028)**	0.069 (0.033)*
Elegance	0.343 (0.041)***	0.678 (0.032)***
Intricacy	0.077 (0.038)*	0.026 (0.032)
Sophistication	0.066 (0.040)*	−0.068 (0.032)*

Note. Entries are the coefficients (SEs) in a hierarchical regression predicting ratings of "beautiful" from ratings of all other attributes.
° $p_{2t} < 0.10$.
* $p_{2t} < 0.05$.
** $p_{2t} < 0.01$.
*** $p_{2t} < 0.001$.

References

Aigner, G., & Ziegler, M. (1998). *Proofs from THE BOOK.* Berlin, Germany: Springer.
Axelsson, Ö. (2007). Individual differences in preferences to photographs. *Psychology of Aesthetics, Creativity, and the Arts, 1*, 61–72.
Berlyne, D. E. (1960). *Conflict, arousal, and curiosity.* New York, NY: McGraw-Hill.
Bilalic, M., Langner, R., Ulrich, R., & Grodd, W. (2011). Many faces of expertise: Fusiform face area in chess experts and novices. *Journal of Neuroscience, 31*, 10206–10214.
Blijlevens, J., Thurgood, C., Hekkert, P., Chen, L.-L., Leder, H., & Whitfield, T. W. A. (2017). The aesthetic pleasure in design scale: The development of a scale to measure aesthetic pleasure for designed artifacts. *Psychology of Aesthetics, Creativity, and the Arts, 11*(1), 86–98.
Bolzano, B. (1817). Purely analytic proof of the theorem that between any two values which give results of opposite sign there lies at last one real root of the equation. (Trans. S. Russ.).
Bonawitz, E. B., & Lombrozo, T. (2012). Occam's rattle: Children's use of simplicity and probability to constrain inference. *Developmental Psychology, 48*, 1156–1164.
Boselie, F., & Leeuwenberg, E. (1985). Birkhoff revisited: Beauty as a function of effect and means. *American Journal of Psychology, 98*, 1–39.
Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process.* Chicago, IL: Unviersity of Chicago Press.
Brielmann, A. A., & Pelli, D. G. (2017). Beauty requires thought. *Current Biology, 27*, 1506–1513.
Caivano, J. L. (1994). Color and sound: Physical and psychophysical relations. *Color Research and Application, 19*, 126–133.
Carey, S. (2009). *The origin of concepts.* Oxford, UK: Oxford University Press.
Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*, 55–81.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121–152.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Ed.). Hillsdale, NJ: Erlbaum.

Cova, F., Garcia, A., & Liao, S. (2015). Experimental philosophy of aesthetics. *Philosophy Compass, 10*, 927–939.

Cova, F., & Pain, N. (2012). Can folk aesthetics ground aesthetic realism? *The Monist, 95*, 241–263.

Dehaene, S. (1996). *The number sense: How the mind creates mathematics.* Oxford, UK: Oxford University Press.

Descartes, R. (1954/1684). Rules for the direction of the mind. In E. Anscombe & P. T. Geach (Trans.), Descartes: Philosophical writings. London, UK: Pearson.

Desmet, P., & Hekkert, P. (2007). Framework of product experience. *International Journal of Design, 1*, 57–66.

Dirac, P. A. M. (1963). The evolution of the physicist's picture of nature. *Scientific American, 208*, 45–53.

Douven, I., & Schupbach, J. N. (2015). The role of explanatory considerations in updating. *Cognition, 142*, 299–311.

Dutton, D. (2009). *The art instinct: Beauty, pleasure, and human evolution.* New York, NY: Oxford University Press.

Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision, 10*, 1–12.

Fauvel, J., Flood, R., & Wilson, R. (2006). *Music and mathematics: From pythagoras to fractals.* New York, NY: Oxford University Press.

Fayn, K., Silvia, P. J., Erbas, Y., Tiliopoulous, N., & Kuppens, P. (2018). Nuanced aesthetic emotions: Emotion differentiation is related to knowledge of the arts and curiosity. *Cognition and Emotion, 32*, 593–599.

Frege, G. (1974). The foundations of arithmetic: A logico-mathematical enquiry into the concept of number. (J. L. Austin, Trans.). Oxford, UK: Blackwell. (Original work published 1884.).

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought.* Cambridge, MA: MIT Press.

Goodman, N. (1972). *Problems and projects.* Indianapolis, IN: Hackett.

Gopnik, A. (1998). Explanation as orgasm. *Minds and Machines, 8*, 101–118.

Halberda, J., Mazzocco, M. M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature, 455*, 665–668.

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

Hekkert, P. (2014). Aesthetic responses to design: A battle of impulses. In P. P. L. Tinio & J. K. Smith (Eds.), The Cambridge handbook of the psychology of aesthetics and the arts (pp. 277–299).

Hoyer, W. D., & Stokburger-Sauer, N. E. (2012). The role of aesthetic taste in consumer behavior. *Journal of the Academy of Marketing Science, 40*, 167–180.

Hubbard, T. L. (1996). Synesthesia-like mappings of lightness, pitch, and melodic interval. *American Journal of Psychology, 109*, 219–238.

Hume, D. (1978). *A treatise on human nature.* Oxford, UK: Oxford University Press (Original work published 1738.).

Hume, D. (1985). *Of the standard of taste. Essays moral, political, and literary*. Indianapolis, IN: Liberty Fund (Original work published 1757.).

Harré, R. (1958). Quasi-aesthetic appraisals. *Philosophy, 33*, 132–137.

Hauser, M. D., & McDermott, J. (2003). The evolution of the music faculty: A comparative perspective. *Nature Neuroscience, 6*, 663–668.

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* (pp. 163–179). Basel, Switzerland: Birkhäuser Verlag.

Ishizu, T., & Zeki, S. (2011). Toward a brain-based theory of beauty. *PLoS One, 6*, e21852.

Johnson, S. G. B. (2017). *Cognition as sense-making (Unpublished doctoral dissertation).* New Haven, CT: Yale University.

Johnson, S. G. B., Johnston, A. M., Koven, M. L., & Keil, F. C. (2017). Principles used to evaluate mathematical explanations. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.). *Proceedings of the 39th annual conference of the cognitive science society* (pp. 612–617). Austin, TX: Cognitive Science Society.

Johnson, S. G. B., Johnston, A. M., Toig, A. E., & Keil, F. C. (2014). Explanatory scope informs causal strength inferences. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.). *Proceedings of the 36th annual conference of the cognitive science society* (pp. 2453–2458). Austin, TX: Cognitive Science Society.

Johnson, S. G. B., Rajeev-Kumar, G., & Keil, F. C. (2016). Sense-making under ignorance. *Cognitive Psychology, 89*, 39–70.

Johnson, S. G. B., & Steinerberger, S. (2018). The aesthetics of mathematical explanations. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.). *Proceedings of the 40th annual conference of the cognitive science society* (pp. 572–577). Austin, TX: Cognitive Science Society.

Johnson, S. G. B., & Steinerberger, S. (2019). The universal aesthetics of mathematics. *Mathematical Intelligencer*.

Johnson, S. G. B., Valenti, J. J., & Keil, F. C. (2017). Simplicity and complexity preferences in explanation: An opponent heuristic account. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.). *Proceedings of the 39th annual conference of the cognitive science society* (pp. 606–611). Austin, TX: Cognitive Science Society.

Johnson-Laird, P. N. (1999). Deductive reasoning. *Annual Review of Psychology, 50*, 109–135.

Johnston, A. M., Johnson, S. G. B., Koven, M. L., & Keil, F. C. (2017). Little Bayesians or little Einsteins? Probability and explanatory virtue in children's inferences. *Developmental Science, 20*, e12483.

Johnston, A. M., Sheskin, M., Johnson, S. G. B., & Keil, F. C. (2018). Preferences for explanation generality develop early in biology, but not physics. *Child Development*.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.). *Heuristics and biases: The psychology of intuitive thought* (pp. 49–81). New York, NY: Cambridge University Press.

Kanigal, R. (2016). *The man who knew infinity: A life of the genius Ramanujan.* New York, NY: Simon & Schuster.

Kant, I. (2002). *Critique of the power of judgment* (Translated by P. Guyer and E. Matthews). Cambrige, UK: Cambridge University Press (Original work published 1790).

Khemlani, S. S., Sussman, A. B., & Oppenheimer, D. M. (2011). *Harry Potter* and the sorcerer's scope: Latent scope biases in explanatory reasoning. *Memory & Cognition, 39*, 527–535.

Kitcher, P. (1975). Bolzano's ideal of algebraic analysis. *Studies in the History and Philosophy of Science, 6*, 229–269.

Krull, W. (1987). The aesthetic viewpoint in mathematics. *Mathematical Intelligencer, 9*, 48–52.

Leder, H., Gerger, G., Brieber, D., & Schwarz, N. (2014). What makes an art expert? Emotion and evaluation in art appreciation. *Cognition and Emotion, 28*, 1137–1147.

Leder, H., Gerger, G., Dressler, S. G., & Schabmann, A. (2012). How art is appreciated. *Psychology of Aesthetics, Creativity, and the Arts, 6*, 2–10.

Levitan, C. A., Ren, J., Woods, A. T., Boesveldt, S., Chan, J. S., McKenzie, K. J., ... van den Bosch, J. J. F. (2014). Cross-cultural color–odor associations. *PLoS One, 9*, e101651.

Lipton, P. (2004). *Inference to the best explanation* (2nd ed.). London, UK: Routledge.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology, 55*, 232–257.

Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences, 20*, 748–759.

Margulies, S. (1977). Principles of beauty. *Psychological Reports, 41*, 3–11.

Marks, L. E. (1987). On cross-modal similarity: Auditory–visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception & Performance, 13*, 384–394.

Mazzocco, M. M. M., Feigenson, L., & Halberda, J. (2011). Impaired acuity of the approximate number system underlies mathematical learning disability (dyscalculia). *Child Development, 82*, 1224–1237.

Medin, D. L., & Atran, S. (2004). The native mind: Biological categorization and reasoning in development and across cultures. *Psychological Review, 111*, 960–983.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100*, 254–278.

Medin, D. L., Lynch, E. B., & Coley, J. D. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognitive Psychology, 32*, 49–96.

Moore, R. C. (1994). Making the transition to formal proof. *Educational Studies in Mathematics, 27*, 249–266.

Nabokov, V. (1969). *Poems and problems.* New York, NY: McGraw-Hill.

Nieminen, S., Istók, E., Brattico, E., Tervaniemi, M., & Huotilainen, M. (2011). The development of aesthetic responses to music and their underlying neural and psychological mechanisms. *Cortex, 47*, 1138–1146.

Orians, G. H., & Heerwagen, J. H. (1992). Evolved responses to landscapes. In J. Barkow, L. Cosmides, & J. Tooby (Eds.). *The adapted mind* (pp. 555–579). New York, NY: Oxford University Press.

Palmer, S. E., Schloss, K. B., Xu, Z., & Prado-Léon, L. R. (2013). Music–color associations are mediated by emotion. *Proceedings of the National Academy of Sciences, 110*, 8836–8841.

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

Pinker, S. (1997). *How the mind works.* New York, NY: W. W. Norton.

Poincaré, H. (2007). *Science and method* (F. Maitland, Trans.). New York, NY: Cosimo (Original work published 1914).

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review, 8*, 364–382.

Rhodes, G., Proffitt, F., Grady, J. M., & Sumich, A. (1998). Facial symmetry and the perception of beauty. *Psychonomic Bulletin & Review, 5*, 659–669.

Rips, L. J. (1994). *The psychology of proof.* Cambridge, MA: MIT Press.

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

Senk, S. L. (1985). How well do students write geometry proofs? *The Mathematics Teacher, 78*, 448–456.

Silvia, P. J. (2013). Interested experts, confused novices: Art expertise and the knowledge emotions. *Empirical Studies of the Arts, 31*, 107–115.

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

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics, 73*, 971–995.

Starkey, P., & Cooper, R. G., Jr. (1980). Perception of numbers by human infants. *Science, 210*, 1033–1035.

Tao, T. (2007). What is good mathematics? *Bulletin of the American Mathematical Society, 44*, 623–634.

Tetlock, P. (2005). *Expert political judgment: How good is it? How can we know?* Princeton, NJ: Princeton University Press.

Thornhill, R., & Gangestad, S. W. (1993). Human facial beauty: Averageness, symmetry and parasite resistance. *Human Nature, 4*, 237–269.

Thornhill, R. (2003). Darwinian aesthetics informs traditional aesthetics. In E. Voland, & K. Grammar (Eds.). *Evolutionary aesthetics* (pp. 9–35). Berlin, Germany: Springer.

Todd, C. S. (2008). Unmasking the truth beneath the beauty: Why the supposed aesthetic judgments made in science may not be aesthetic at all. *International Studies in the Philosophy of Science, 11*, 61–79.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327–352.

Van Fraassen, B. C. (1989). *Laws and symmetry*. Oxford, UK: Clarendon Press.
Velasco, C., Woods, A. T., Deroy, O., & Spence, C. (2015). Hedonic mediation of the crossmodal correspondence between taste and shape. *Food Quality and Preference, 41*, 151–158.
Velasco, C., Woods, A. T., Hyndman, S., & Spence, C. (2015). The taste of typeface. *i-Perception, 6*, 1–10.
Verhavert, S., Wagemans, J., & Augustin, M. (2018). Beauty in the blink of an eye: The time course of aesthetic experiences. *British Journal of Psychology, 109*, 63–84.
Vessel, E. A., & Rubin, N. (2010). Beauty and the beholder: Highly individual taste for abstract, but not real-world images. *Journal of Vision, 10*.
Wells, D. (1990). Are these the most beautiful? *The Mathematical Intelligencer, 12*, 37–41.
Westphal-Fitch, G., & Fitch, W. T. (2017). Beauty for the eye of the beholder: Plane pattern perception and production. *Psychology of Aesthetics, Creativity, and the Arts, 11*, 451–456.
Wynn, K. (1992). Addition and subtraction by human infants. *Nature, 358*, 749–750.
Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition, 74*, B1–B11.
Zee, A. (1999). *Fearful symmetry: The search for beauty in modern physics*. Princeton, NJ: Princeton University Press.
Zeki, S., Romaya, J. P., Benincasa, D. M. T., & Atiyah, M. F. (2014). The experience of mathematical beauty and its neural correlates. *Frontiers in Human Neuroscience, 8*, 68.