BEAUTY IS NOT SIMPLICITY: AN ANALYSIS OF MATHEMATICIANS' PROOF APPRAISALS

MATTHEW INGLIS[1] AND ANDREW ABERDEIN[2]

Abstract. What do mathematicians mean when they use terms such as 'deep', 'elegant', and 'beautiful'? By applying empirical methods developed by social psychologists, we demonstrate that mathematicians' appraisals of proofs vary on four dimensions: aesthetics, intricacy, utility, and precision. We pay particular attention to mathematical beauty and show that, contrary to the classical view, beauty and simplicity are almost entirely unrelated in mathematics.

1. Introduction

Mathematical conversations are full of value judgements. Mathematicians talk of 'beautiful', 'deep', 'insightful', and 'interesting' proofs, and award each other prizes on the basis of these assessments. Validity or applicability are almost never the decisive criteria for such awards. Instead the citations for mathematical prizes are full of aesthetic judgements: nine of the eleven Abel Prize citations since its foundation have characterised the prizewinner or their work as 'deep', and the work of the remaining two was lauded for its beauty and ingenuity (Holden & Piene, 2009, 2013). Furthermore, many of the most eminent researchers have suggested that it is these value judgements which drive their research agendas. Hermann Weyl even claimed to prioritise beauty over truth: 'My work has always tried to unite the true with the beautiful; and when I had to choose one or the other, I usually chose the beautiful' (cited in Reid, 1986, p. 161).

Informal mathematical discourse suggests that a wide variety of adjectives can be meaningfully applied to mathematical objects, theorems, and proofs. The mathematician Ravi Vakil for instance, when discussing spectral sequences, remarks that

> They have a reputation for being abstruse and difficult. It has been suggested that the name 'spectral' was given because, like spectres, spectral sequences are terrifying, evil, and dangerous (Vakil, 2013, p. 57).

But what do mathematicians mean when they use such terms? And how do they reach such judgements?

Terence Tao has suggested that 'the concept of mathematical quality is a high-dimensional one' (Tao, 2007, p. 624). This raises an immediate question: how many dimensions of mathematical quality are there? Tao does not propose a definite answer. Instead, he offers an expressly non-exhaustive list of 21 features which could result in a piece of mathematics being positively assessed. These include its beauty, elegance,

[1]Mathematics Education Centre, Loughborough University, UK.
[2]School of Arts and Communication, Florida Institute of Technology, USA.
E-mail addresses: m.j.inglis@lboro.ac.uk, aberdein@fit.edu.
Date: June 20, 2014.

creativity, depth, strength, intuitiveness, and definitiveness. From this list he deduces that mathematical quality is difficult to pin down, arguing that it may appear as if 'the problem of evaluating mathematical quality, while important, is a hopelessly complicated one, especially since many good mathematical achievements may score highly on some [qualities] but not on others' (p. 626). But although the problem may appear hopelessly complicated, it cannot be truly intractable, unless the assessments of mathematical quality that seem ubiquitous in mathematical practice lack all content.

Tao's approach to resolving this conundrum begins with the observation that good mathematics (in any of his senses) tends to beget further pieces of good mathematics (in the same or other senses). Illustrating his idea with a case study of the history of Szemerédi's Theorem, a 'beautiful and celebrated result' in number theory (p. 627), Tao suggests that good pieces of mathematics represent important staging posts on a greater mathematical journey, since they promise many more pieces of good mathematics in the future. How then does a mathematician judge the quality of a novel piece of mathematics, which has not yet had time to beget any further mathematics, good or bad? Tao appeals to the mathematician's experience, suggesting that he or she would have built up over time an 'undefinable sense' that the novel mathematical work is 'a piece of a larger puzzle waiting to be explored further' (p. 633). Timothy Gowers (2002) is of like mind, arguing that good mathematical proofs provide 'intriguing hints that there is more to be discovered' (p. 51). But there are empirical findings that pose a challenge for this expertise-based account of mathematical evaluation. Notably, students, who presumably have scant experience of observing mathematical development, and therefore no obvious means of determining whether or not proofs are likely to fit into a larger picture, are apparently nevertheless capable of appreciating the aesthetics of at least some mathematical arguments (Koichu, Katz, & Berman, 2007; Papert, 1980; Sinclair, 2004). So, Tao does not seem to have resolved his conundrum.

However, if the dimensionality of mathematical qualities were substantially smaller than Tao suggests, the evaluation of mathematics need not be intractable, so the conundrum would not arise. Our goal in this paper is to investigate this question empirically. Specifically, we ask: on how many broad dimensions can perceptions of mathematical proofs be said to vary? Prior to describing our strategy, method, and findings, we briefly review existing accounts of what is sometimes considered the most valuable quality of mathematical proofs, that of beauty.

2. Mathematical Beauty

The idea that mathematical arguments could be considered 'beautiful' may appear somewhat peculiar to non-mathematicians, but this kind of language seems to be ubiquitous throughout mathematical practice. Henri Poincaré (1914) described mathematical beauty as a 'real aesthetic feeling that *all true mathematicians* recognize' (p. 59, our emphasis), and many mathematicians have claimed that it is the quest for beauty which drives their research agendas (Engler, 1990).

Although mathematical theorems and definitions are sometimes perceived as being beautiful (Wells, 1990), it seems that the mathematical objects most commonly described in this way are proofs. But the notion of beautiful proofs raises a number of serious questions. What features must a mathematical proof have in order to be 'beautiful'?

Is 'beauty' in mathematics a genuinely aesthetic quality, or just a proxy for some other property? Is 'beauty' projected onto a proof by a reader, or is it a property of the proof itself? Different authors have taken different positions on this latter question. For example, G. H. Hardy (1940) believed that beauty was an objective property of the mathematics (famously claiming that 'there is no permanent place in the world for ugly mathematics', p. 85), whereas Kant saw mathematical beauty as being highly subjective, a feeling which a reader experiences when a proof fits with their intellectual capabilities (Breitenbach, in press).[1] Before outlining our research strategy, we first briefly discuss some ways in which mathematical beauty has been accounted for in the literature.

2.1. **Beauty as Simplicity.** A classic view of mathematical beauty is to relate the notion to simplicity. James McAllister (2005) epitomises this stance:

> Mathematicians have customarily regarded a proof as beautiful if it conformed to the classical ideals of brevity and simplicity. The most important determinant of a proof's perceived beauty is thus the degree to which it lends itself to being grasped in a single act of mental apprehension (p. 22).

Similarly, Michael Atiyah claims that 'elegance is more or less synonymous with simplicity' (cited in Wells, 1990, p. 39), and similar views can be found in the works of Carl Friedrich Gauss (e.g., Tappenden, 2008), Werner Heisenberg (e.g., Engler, 1990) and Poincaré (e.g., Chandrasekhar, 2010), as well as many contemporary research mathematicians (e.g., Cherniwchan, Ghassemi, & Keating, 2013).

Simplicity has also been associated with the beauty of aspects of mathematics other than proofs. David Wells (1990) surveyed 68 mathematicians to determine which of 24 theorems they found to be the most beautiful. He found that, even though his survey stated the theorems without proof, simplicity of proof impacted upon the perceived beauty of the theorems. For example, some participants marked down the theorem 'Every prime number of the form $4n + 1$ is the sum of two integral squares in exactly one way' because it lacked a simple proof, and others expressed a desire to find a simple proof of the theorem 'A continuous mapping of the closed unit disk into itself has a fixed point'. Overall, Wells found that $e^{i\pi} = -1$, the existence of infinitely many primes, and $V + F = E + 2$ were rated as being the most beautiful theorems, and a complex polynomial identity[2] the least, again seemingly confirming the importance of simplicity in aesthetic judgements.

In sum, one attempt to provide a characterisation of which proofs mathematicians find beautiful is to suggest that the perceived beauty of a proof is identical to, or at least highly correlated with, its perceived simplicity.

2.2. **Beauty as Epistemic Satisfaction.** Some have questioned whether aesthetic judgements in science and mathematics are really related to aesthetics at all. Rom Harré (1958) argues that 'we are no more entitled to suppose that when someone calls a proof

[1] At least on Breitenbach's reconstruction of Kant's views; as she concedes, other commentators have read Kant as denying that mathematics is ever beautiful.

[2] $$\frac{5\left[(1-x^5)(1-x^{10})(1-x^{15})\ldots\right]^5}{\left[(1-x)(1-x^2)(1-x^3)(1-x^4)\ldots\right]^6} = p(4) + p(9)x + p(14)x^2 + \ldots$$
where $p(n)$ is the number of partitions of n.

"elegant" he is appraising it on aesthetic grounds [...] than we would be to suppose that he is appraising it on moral grounds when he calls it "good"' (p. 136). Cain Todd (2008) agrees, and develops Harré's argument by distinguishing two types of account in which the existence of genuine aesthetic judgement is posited: conjunctive accounts, which assert that aesthetic and epistemic judgements coincide, and disjunctive accounts, which assert that they are independent.

Many scientists seem to subscribe to the conjunctive view, believing that the perceived beauty of a scientific theory is somehow a sign of its truth (Engler, 1990; Kivy, 1991; McAllister, 1996). Paul Dirac, for instance, asserted that if a theory is 'really beautiful, then it necessarily will appear as a fine model of important physical phenomena' (cited in Stakhov, 2009, p. 615). Todd (2008) rejects this idea, on the grounds that it is very difficult to see how aesthetic factors could be systematically related to a theory's empirical adequacy. Although McAllister's (1996) 'aesthetic induction' would appear to provide a mechanism for such a relationship (essentially McAllister proposes that empirically adequate scientific theories come, in time, to be perceived as beautiful), Todd suggests that, despite their use of aesthetic terms, if scientists' judgements are actually about empirical adequacy then they are at heart epistemic not aesthetic.

In contrast, disjunctive theories suppose that aesthetic judgements are independent of epistemic judgements. This view seems to imply the existence of, or at least the possibility of, scientific theories and mathematical proofs which are true but not beautiful, and beautiful but not true. Whereas the first of these possibilities seems highly plausible, Todd questions whether the latter is tenable, arguing that 'theories cannot be beautiful that do not have properties indicative of truth' (p. 68). While this may be true for scientific theories, it is not clear that this claim is meaningful for mathematical proofs. A proof must be perceived to be 'true' (valid) before it is accepted as a proof. For proofs, 'truth' (or validity) is a qualifying trait: a mathematical proof cannot be a beautiful proof if it is not valid, not because this would entail it being ugly, but rather because it would necessarily be a non-proof.

Nevertheless, Todd rejects both conjunctive and disjunctive views. He concludes that the most parsimonious account involves rejecting the notion of scientific and mathematical aesthetic judgements entirely. Instead he contends that when mathematicians and scientists talk of beautiful proofs or theories they are actually making epistemic not aesthetic claims. Todd does, however, accept that his reductive account is open to challenge, remarking that he does not wish to argue 'that the primarily epistemic activities of science and mathematics cannot also be appreciated aesthetically, merely that if they can it needs to be shown along what axes aesthetic and epistemic appreciation and judgement operate in relation to each other' (Todd, 2008, p. 75). Therefore if the dimensionality of mathematical quality could be appropriately assessed, as we attempt to do in this paper, it would go some way to challenging Todd's argument for a reductive account of aesthetics in mathematics.

2.3. **Beauty as Enlightenment.** Todd is not alone in proposing a reductive account: Gian-Carlo Rota's (1997) mathematical aesthetics is a particularly clear instantiation. For Rota, when mathematicians talk about 'beauty', they are actually referring to 'enlightenment', an epistemic concept. To flesh out this approach, Rota diagnoses the 'light-bulb mistake'. He suggests that although understanding a piece of mathematics

is a complex and often painstaking process, mathematicians frequently remember the learning process as if the key idea 'had been perceived by an instantaneous realization, in a moment of truth, like a light-bulb suddenly being lit' (p. 179). On Rota's account, it is this perception of instantaneous enlightenment that is at the core of perceptions of mathematical beauty:

> The term 'mathematical beauty', together with the light-bulb mistake, are tricks that mathematicians have devised to avoid facing up to the messy phenomenon of enlightenment. [...] All talk of mathematical beauty is a copout from confronting the logic of enlightenment (Rota, 1997, p. 182).

So presumably mathematicians perceive simplicity to be a characteristic of beautiful mathematics simply because it is easier to gain enlightenment from a simple proof compared to a complex proof.[3]

But why would mathematicians use the term 'beautiful' in place of the, in Rota's view, more correct term 'enlightening'? Rota's suggestion is that they do so because 'beauty', unlike 'enlightenment', is a concept which does not admit degrees—something is either beautiful or not beautiful—and mathematicians 'universally dislike any concepts admitting degrees' (p. 181). This position has been criticised by Ulianov Montaño (2014), who suggests that there is no reason why beauty could not admit degrees. Indeed, 'X is more beautiful than Y, but less than Z' appears to be a perfectly well-formed sentence.

Montaño (2012) also criticises Rota for his approach to ugliness and elegance. Montaño argues that, if ugliness were the opposite of beauty, we should expect that a piece of mathematics would be labelled ugly if it failed to provide enlightenment. However, confusingly, Rota claims that ugliness (or at least, the lack of beauty) is in fact associated with a 'lack of definitiveness' (p. 178), not a lack of enlightenment. Rota also has an idiosyncratic account of mathematical elegance. Whereas many or most mathematicians and writers treat 'beauty' and 'elegance' as more or less synonymous (e.g., Sinclair, 2011; Wells, 1990), he instead claims that elegance is a pedagogical matter, that it 'has to do with the presentation of mathematics, and only tangentially does it relate to content' (p. 178). In the sections below we empirically investigate the relationship between perceived beauty, enlightenment, and definitiveness.

2.4. **Summary.** There is general agreement that the notion of mathematical beauty is widespread in mathematical practice. Mathematicians characterise each others' work as beautiful (or not), and strive to produce beautiful proofs. But there is disagreement about what beauty is in a mathematical context. Two broad classes of theory have been proposed.

Non-reductive accounts of beauty in mathematics suppose that when mathematicians talk of beautiful proofs they are genuinely making an aesthetic judgement (e.g., McAllister, 2005). A classic view is that simplicity is central to these assessments. In contrast, reductive accounts, such as those proposed by Todd (2008) and Rota (1997), suggest that

[3]Perhaps the most famous light-bulb analogy in modern mathematics, Andrew Wiles's account of proving Fermat's Last Theorem as finding the light switches in each of the rooms of a 'dark unexplored mansion', implicitly challenges Rota's supposition that mathematicians forget the process that preceded the light coming on: 'each of these breakthroughs, while sometimes they're momentary, sometimes over a period of a day or two, they are the culmination of, and couldn't exist without, the many months of stumbling around in the dark that precede them' (quoted in Singh, 1997, p. 258).

when mathematicians talk of beauty they are using the term as a proxy for an epistemic assessment. In particular, Rota argues that beautiful proofs are those which provide enlightenment. And Todd argues that, until the dimensions along which aesthetic and epistemic appreciation operate can be fully determined, reductive accounts offer the most parsimonious approach to mathematical beauty.

Our goal in this paper is to directly investigate the dimensions along which mathematical proofs are typically evaluated. In particular, is Tao (2007) correct to argue that the dimensionality of mathematical quality is high? On how many dimensions do assessments of mathematical proofs vary? And does the answer to this question offer insights into the nature of mathematical beauty?

3. Research Strategy

Our approach to addressing the dimensionality of mathematical assessment relies upon a statistical procedure known as a *factor analysis*. Factor analyses attempt to model the covariation among a set of observed variables in terms of functions of a small number of latent constructs, or factors, which are themselves unobservable. The goal is to explain as much of the original variance as possible using a small set of factors, and to express each of the original variables as a function of these new factors. The technique works by looking for patterns in a matrix of the correlations between the original variables: if a set of variables are all strongly inter-correlated they can, in some sense, be said to represent the same underlying construct. If, for example, one found a large group of men and measured their height, arm length, leg length, thigh length, shoulder width, and weight, we would expect these measurements to be strongly (albeit not perfectly) correlated, and could reasonably interpret each of these measurements as being in some sense a measure of the latent construct 'general body size'. Indeed, exactly this study was performed by Burt and Banks (1946), who looked at nine different measurements taken from 2400 adult males, and successfully identified a general size factor. Factor analyses are widely used in a wide variety of research domains, including the study of intelligence (e.g. Deary, 2001), marketing (e.g. Stewart, 1981) and, perhaps most analogously to the current project, human personalities.

Human personalities are clearly more complex than human body measurements, but an approach methodologically analogous to Burt and Banks's (1946) study can, and has been, adopted for their investigation. There are a great many natural language terms which can be used to describe someone's personality (in fact, Allport and Odbert (1936) identified almost 18,000 English words which can be used to describe a person's traits). The standard procedure used to investigate the dimensionality of personalities is to ask participants to think of a person (maybe a peer, maybe themselves), and then rate how well a series of adjectives describe that person, typically on a five-point scale from 'very inaccurately' to 'very accurately'. These ratings can then be subjected to a factor analysis to determine how many factors, or dimensions, emerge. Many such studies in dozens of countries have been conducted since Tupes and Christal's (1961) early work, with the consistent finding that adjectives which describe human traits cluster around five broad factors, which have become known as the 'Big Five' (Schmitt et al. (2007) replicated this

result across 56 countries and 28 languages; for a review, see John, Naumann, & Soto, 2008).[4]

Thus an individual's personality can, as a rough approximation, be seen as a point within five-dimensional space. There are now various methods of estimating where a given person lies within this space (e.g., Donnellan, Oswald, Baird, & Lucas, 2006), and these estimates are systematically related to a great many real-world behaviours. For example Big Five profiles predict, among other things, life satisfaction (Boyce, Wood, & Powdthavee, 2013), perceived well-being (Hayes & Joseph, 2003), job performance (Paunonen & Ashton, 2001), drug consumption (Paunonen & Ashton, 2001), frequency of group conversations in day-to-day life (Mehl, Gosling, & Pennebaker, 2006), and even philosophical intuitions (Feltz & Cokely, 2013).

Our conjecture is that classifying mathematicians' perceptions of the qualities of mathematical proofs is an analogous problem to the challenge of characterising human personalities. In both cases a great many adjectives are used to describe the traits of the object (be it a person or a proof), but inspection of these adjectives gives credence to the idea that many or most of them are describing broadly similar properties: 'profound' and 'deep' appear to be rather similar, for example. Thus we followed a research strategy analogous to that used by social psychologists interested in personalities. First, we produced a list of adjectives which have often been used to describe mathematical proofs. Second, we asked a large number of mathematicians to think of a proof that they had recently read, and to state how accurately each adjective described that proof. Finally, we subjected these ratings to an Exploratory Factor Analysis (EFA), to determine on how many broad dimensions mathematicians' perceptions of proofs vary. We were particularly interested in using these findings to interrogate the various accounts of mathematical beauty described above.

Our research strategy does not, of course, allow us to draw any conclusions about objective qualities of proofs (or indeed, to say whether or not proofs *have* objective qualities), and neither does it allow us to understand which proofs have which qualities, or whether there are between-mathematician differences in the assessment of mathematical quality. It does, however, allow us to investigate the structure of the language with which mathematicians characterise the qualities of mathematical proofs.

This project has clear affinities with recent work in 'experimental philosophy'. Like the experimental philosophers, we seek to examine 'intuitions that have long been at the center of philosophical study [...] using the methods associated with contemporary cognitive science—systematic experimentation and statistical analysis' (Knobe, 2007, p. 81). However, our work is perhaps closer in spirit to the 'empirical semantics' of the mid-twentieth century Oslo Group, and especially the work of its founder, Arne Naess (Naess, 1938, 1981; Crockett, 1959; Barnard & Ulatowski, 2013). What is distinctive about Naess's approach is his emphasis on empirically investigating the use of philosophically significant terms as a guide to their meaning, an approach we share. We differ from Naess in the more quantitative character of our methodology and our focus on the practice of a specific community, research mathematicians, rather than the population as a whole. Of course, the strategy of looking for meaning in use has a much broader philosophical

[4]The Big Five factors are: conscientiousness, extraversion, agreeableness, neuroticism, and openness to experience.

TABLE 1. The eighty adjectives used in the study.

abstract	deep	informative	profound
accurate	definitive	ingenious	rigorous
ambitious	delicate	innovative	robust
appealing	dense	insightful	shallow
applicable	difficult	inspired	sharp
awful	dull	intricate	simple
awkward	effective	intuitive	sketchy
beautiful	efficient	loose	speculative
bold	elaborate	lucid	striking
careful	elegant	meticulous	strong
careless	enlightening	minimal	sublime
charming	explanatory	natural	subtle
clear	exploratory	non-trivial	tedious
clever	expository	obscure	trivial
clumsy	flimsy	obvious	ugly
conceptual	fruitful	plausible	unambiguous
confusing	general	pleasing	unpleasant
creative	illustrative	polished	useful
crude	incomplete	practical	weak
cute	inefficient	precise	worthless

pedigree. It was one of the distinctive features of the 'linguistic turn', perhaps most closely identified with the work of Wittgenstein and Quine. It may be traced back even further to the American pragmatists of the nineteenth century, and especially to Peirce. In Wittgensteinian terms, the notion of mathematical beauty can only be understood by investigating how mathematicians use the word 'beautiful' in mathematical 'language games'. Our approach could be seen as an attempt to systematically study the language game which takes place when mathematicians evaluate proofs.

4. METHOD

4.1. **Materials and Procedure.** Our first task was to select a long list of adjectives which have been used to describe mathematical proofs. Using Tao's (2007) list of mathematical qualities as a starting point, we formed a list of 80 adjectives that we conjectured may be used by mathematicians to describe the traits of mathematical proofs. These are shown in Table 1.

Like earlier researchers interested in empirically studying research mathematicians' behaviour (Heinze, 2010; Inglis, Mejía-Ramos, Weber, & Alcock, 2013; Löwe, Müller, & Müller-Hill, 2010), we adopted a web-based approach. All mathematics departments with graduate programmes ranked by *U.S. News & World Report* were invited by email to participate in the study. If the department agreed, they forwarded an email invitation to participate to all research-active mathematicians in their departments. As with all research which requires participants to give informed consent prior to participation, our participants were self-selected and so cannot be said to be a truly random sample. The

email gave a brief outline of the purpose of the research, and provided a web link to the location of the study. Participants who decided to take part first saw an introductory page which again explained the purpose and nature of the research. On the second page participants were asked to select their research area (applied mathematics, pure mathematics, or statistics), and state their level of experience (PhD student, postdoc, or faculty). On the third page participants were given the following instructions:

> Please think of a **particular** proof in a paper or book which you have recently refereed or read. Keeping this specific proof in mind, please use the rating scale below to describe how accurately each word in the table below describes the proof. Describe the proof as it was written, not how it could be written if improved or adapted. So that you can describe the proof in an honest manner, you will not be asked to identify it or its author, and your responses will be kept in absolute confidence. Please read each word carefully, and then select the option that corresponds to how well you think it describes the proof. (Emphasis in the original)

Participants were then shown the list of eighty adjectives, presented in a random order, and asked to select how well each described their chosen proof using a five-point Likert scale (very inaccurate, inaccurate, neither inaccurate nor accurate, accurate, very accurate). Finally participants were thanked for their time, and invited to contact the research team if they wanted further information.

4.2. **Participants.** A total of 255 mathematicians participated in the study, consisting of 146 PhD students, 23 postdocs, and 86 faculty. 192 participants described themselves as pure mathematicians, 50 described themselves as applied mathematicians, 12 were statisticians, and 1 declined to answer. Sixteen participants did not respond to one or more adjectives, resulting in a total of 20 missing values (0.1% of the dataset), which were imputed using item means.

5. Results

5.1. **Analysis Strategy.** Prior to conducting the main analysis, the suitability of participants' ratings for factor analysis was investigated.[5] Typically two tests are used to investigate this issue: in our case the Kaiser-Meyer-Oklin value, at .881, exceeded the recommended .6 (Kaiser, 1974), and Bartlett's Test of Sphericity, $\chi^2(3160) = 12623, p < .001$, confirmed that the correlation matrix contained non-zero terms. Thus both tests supported the use of an EFA.

Participants' ratings for the 80 adjectives were entered into an EFA, using the maximum likelihood method.[6] Both a Scree Test (Cattell, 1966) and Parallel Analysis (Horn, 1965) supported extracting five factors, which together accounted for 44.7% of the original variance. Both the Scree Test and Parallel Analysis are approaches which attempt to find a balance between extracting sufficient factors to explain a high proportion of the original variance, and extracting so many that closely related latent constructs are represented.

[5]Main analyses were conducted with IBM SPSS Statistics 20, Horn's Parallel Analysis was conducted using Monte Carlo PCA.

[6]In separate analyses we also used the unweighted least squares and principal axis methods; both gave essentially identical results.

Initially we performed an oblimin rotation, but because the correlations between factors were low (all R^2s $< .1$), we followed Tabachnick and Fidell's (2007) advice and used a varimax rotation to ensure that all factors were orthogonal, thus aiding interpretation of the factor structure.

Loadings for the five extracted factors are given in Tables 2 and 3. These figures describe how well each adjective describes each factor: so if an adjective has a high positive loading then it is very representative of that factor, if it has a zero loading then it is independent of that factor, and if it has a high negative loading then it is very unrepresentative of that factor.

Before describing the factors in detail we first note that Factor 2 appeared to be somewhat different from the other factors. Adjectives which loaded strongly onto Factor 2 included 'crude', 'careless', 'shallow', 'flimsy', and 'inefficient', all of which had very low ratings. In fact the mean rating for adjectives which loaded strongly ($> .4$) onto Factor 2 was 1.84 (on a 1–5 scale), which was significantly lower than for other adjectives, $M = 3.14, t(78) = 12.0, p < .001$. This seemed to suggest that Factor 2 was a proxy for 'non-use', in other words that adjectives which mathematicians systematically felt were not accurate descriptions of the proof they had chosen loaded strongly onto Factor 2. To investigate this further, we correlated all 80 adjectives' mean ratings with their loadings on Factor 2, finding an extremely strong relationship, $r = -.941, p < .001$, shown in Figure 1. We therefore concluded that Factor 2 was not a true dimension upon which proofs vary (or at least, proofs which mathematicians have recently read or refereed and chose to think about do not vary substantially on this dimension). Consequently we do not discuss Factor 2 in the remainder of the paper. Repeating the EFA on the 60 adjectives which had mean ratings significantly greater than 2 (on the 1 to 5 Likert scale) resulted in four factors which were essentially identical to Factors 1, 3, 4 and 5 discussed below. The table of factor loadings for this analysis is given in the Appendix (Tables 5 and 6).

5.2. **The Four Factors.** We now discuss each of the remaining factors in turn.

The first factor contained adjectives which seemed to refer to the aesthetic qualities of proofs: 'striking', 'ingenious', 'inspired', 'profound', 'creative', 'deep', 'sublime', 'innovative', 'beautiful', 'elegant', and 'charming' all loaded strongly onto this factor. We refer to this factor as the *aesthetics* dimension.[7]

The third factor contained adjectives which seemed to refer to the intricacy of the proof: 'dense', 'difficult', 'intricate', 'unpleasant', 'confusing', and 'tedious' loaded strongly onto this factor, and 'simple' had a strong negative loading. We refer to this factor as the *intricacy* dimension.

The fourth factor contained adjectives which seemed to refer to the usefulness of the proof: 'practical', 'informative', 'efficient', 'applicable', and 'useful' loaded strongly onto this factor. We refer to this factor as the *utility* dimension.

The fifth factor contained adjectives which seemed to refer to the precision of the proof: 'careful', 'precise', 'meticulous', and 'rigorous' loaded strongly onto this factor. We refer to this factor as the *precision* dimension.

[7]The names of the four factors used in this paper are the result of researcher judgement, and are not dictated by the statistical analysis.

TABLE 2. Adjectives which loaded strongly onto Factors 1 (aesthetics) and 2 (non-proofs). Loadings greater than ±.4 are shown in bold.

Adjective	F1	F2	F3	F4	F5
striking	**0.790**	-0.089	-0.038	0.138	0.095
ingenious	**0.781**	-0.187	0.116	-0.027	0.019
inspired	**0.753**	-0.127	-0.011	0.119	0.100
profound	**0.735**	0.000	0.195	0.118	0.009
creative	**0.720**	-0.203	0.040	0.114	0.086
deep	**0.709**	-0.058	0.250	0.120	0.045
sublime	**0.700**	0.036	-0.015	0.094	0.200
innovative	**0.693**	-0.091	0.107	0.032	0.116
beautiful	**0.688**	-0.275	-0.243	0.148	0.147
elegant	**0.659**	-0.328	-0.275	0.177	0.145
charming	**0.643**	0.065	-0.311	0.173	0.102
clever	**0.593**	-0.129	0.073	0.114	0.052
bold	**0.571**	0.192	0.149	0.090	0.111
appealing	**0.555**	-0.225	-0.371	0.292	0.148
pleasing	**0.553**	-0.228	-0.314	0.277	0.242
enlightening	**0.527**	-0.131	-0.100	0.366	0.201
ambitious	**0.517**	0.125	0.189	0.178	0.071
insightful	**0.517**	-0.141	-0.028	0.304	0.125
strong	**0.511**	-0.062	0.218	0.358	0.277
delicate	**0.504**	0.011	0.219	-0.117	0.106
subtle	**0.462**	-0.087	0.187	-0.166	-0.012
sharp	**0.428**	-0.040	0.000	0.292	0.309
cute	**0.425**	0.134	-0.353	0.138	0.096
exploratory	0.304	0.194	-0.016	0.176	0.099
crude	-0.069	**0.676**	0.141	-0.007	-0.008
careless	0.021	**0.666**	0.030	0.074	-0.305
shallow	-0.153	**0.654**	-0.146	0.028	0.107
flimsy	0.150	**0.637**	0.029	-0.020	-0.096
inefficient	-0.044	**0.631**	0.213	-0.126	-0.055
ugly	-0.313	**0.613**	**0.414**	-0.038	-0.001
sketchy	0.085	**0.611**	0.061	0.057	-0.396
incomplete	-0.018	**0.600**	0.107	-0.025	-0.343
weak	-0.160	**0.595**	-0.063	-0.043	-0.032
loose	-0.050	**0.592**	0.039	0.122	-0.110
worthless	-0.018	**0.590**	0.044	-0.128	-0.081
clumsy	-0.060	**0.545**	0.324	-0.055	0.002
awkward	-0.187	**0.538**	**0.418**	0.020	-0.073
awful	-0.088	**0.535**	0.367	-0.157	-0.062
trivial	-0.233	**0.459**	-0.176	-0.024	0.148
obvious	-0.206	**0.454**	**-0.409**	0.141	0.305
speculative	0.326	**0.422**	-0.073	-0.025	-0.069
dull	**-0.405**	**0.419**	0.339	0.138	0.078

TABLE 3. Adjectives which loaded strongly onto Factors 3 (intricacy), 4 (utility), and 5 (precision). Loadings greater than ±.4 are shown in bold.

Adjective	F1	F2	F3	F4	F5
dense	0.121	0.209	**0.640**	0.043	0.059
difficult	0.364	0.112	**0.639**	-0.057	-0.002
intricate	0.323	-0.005	**0.602**	0.035	0.276
unpleasant	-0.203	**0.472**	**0.594**	-0.012	-0.001
simple	-0.119	0.057	**-0.581**	0.217	-0.034
confusing	-0.051	**0.429**	**0.566**	-0.043	-0.213
tedious	-0.186	0.365	**0.535**	0.015	0.199
elaborate	0.372	0.052	**0.486**	-0.056	0.261
non-trivial	**0.426**	-0.224	**0.463**	0.060	0.075
obscure	0.117	0.372	**0.460**	0.003	-0.161
abstract	0.276	0.129	0.393	0.177	-0.027
practical	-0.086	0.016	-0.085	**0.734**	0.061
informative	0.194	-0.119	0.013	**0.632**	0.261
efficient	0.233	-0.190	-0.186	**0.628**	0.053
applicable	0.087	0.015	0.043	**0.616**	0.206
useful	0.178	-0.138	0.036	**0.563**	0.217
effective	0.215	-0.162	0.041	**0.460**	0.255
intuitive	0.119	0.059	-0.384	**0.455**	0.297
plausible	-0.100	0.036	-0.078	**0.439**	0.173
fruitful	**0.411**	-0.060	0.125	**0.437**	0.206
conceptual	0.309	0.032	0.107	**0.422**	-0.019
illustrative	0.236	0.095	-0.270	**0.411**	0.276
natural	-0.003	0.123	-0.308	0.376	0.372
minimal	0.110	0.129	-0.167	0.282	-0.117
general	0.244	0.069	0.080	0.280	-0.020
precise	0.187	-0.277	0.103	0.253	**0.633**
careful	0.221	-0.192	0.216	0.088	**0.633**
meticulous	0.116	0.015	0.383	0.079	**0.576**
rigorous	0.102	-0.315	0.141	0.210	**0.527**
accurate	0.074	-0.265	0.085	0.373	**0.466**
clear	0.114	-0.359	**-0.423**	0.201	**0.455**
lucid	0.320	-0.043	-0.166	0.195	**0.450**
definitive	0.249	-0.013	0.051	0.258	**0.403**
unambiguous	0.169	-0.168	-0.120	0.184	**0.403**
polished	0.223	-0.258	-0.154	0.323	0.389
explanatory	0.101	0.002	-0.308	0.313	0.367
robust	0.258	-0.020	0.099	0.326	0.332
expository	0.112	0.116	-0.117	0.221	0.292

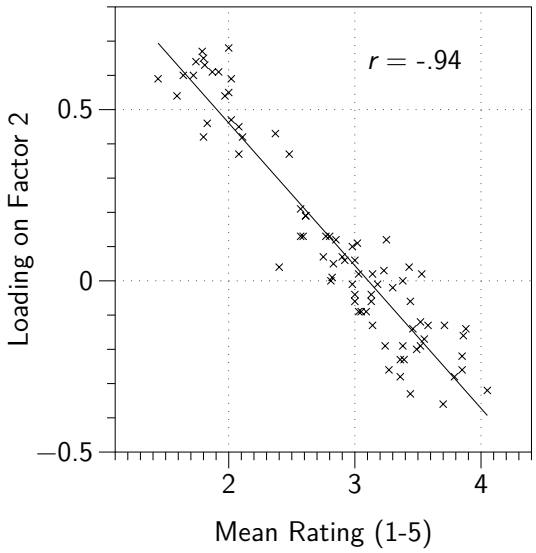

FIGURE 1. The relationship between each adjective's mean rating, and its loading onto Factor 2.

5.3. **Correlates of Beauty.** Next we investigated those adjectives which best correlated with 'beautiful'. The Spearman rank correlation between each adjective and 'beautiful' is given in Table 4. Notably, as predicted by Rota's (1997) account, 'enlightening' was reasonably strongly correlated with 'beautiful' ($r_s = .566, p < .001$) but contrary to another suggestion from Rota, 'definitive' ($r_s = .264, p < .001$) was a weak correlate. Surprisingly, given the classical view of the relationship between beauty and simplicity, 'beautiful' and 'simple' had a correlation coefficient not significantly different from zero ($r_s = .078, p = .217$). Although Atiyah (cited in Wells, 1990) holds that elegance and simplicity are 'more or less synonymous', we found that 'elegant' and 'simple' correlated very weakly ($r_s = .148, p = .018$).

6. Discussion

We structure our discussion of these findings in five sections. First we summarise our main results, then we discuss the implications of these findings for the three accounts of mathematical beauty discussed in the introduction. Finally, we discuss the role of empirical data in discussions of mathematical practice, and argue that reflective reports from mathematicians must be treated with appropriate caution.

6.1. **Summary.** Our analysis indicated that mathematicians' appreciation of the qualities of mathematical proofs can be reasonably well understood using four-dimensional space. In particular, the 'personality' of a proof can be established by considering its aesthetics, intricacy, utility, and precision, each of which forms an independent orthogonal dimension. 'Beautiful' loaded strongly onto the aesthetics factor, whereas 'simple' loaded (negatively) onto the intricacy factor. Looking directly at the correlates of beauty revealed that there appears to be no relationship between a proof's perceived beauty and its

TABLE 4. Spearman rank correlations between each of the eighty adjectives and 'beautiful'. Coefficients greater than ±.21 are significantly different from zero (after Bonferroni correction).

Adjective	r_s	Adjective	r_s	Adjective	r_s	Adjective	r_s
beautiful	1.00	sharp	.35	accurate	.21	dense	-.08
elegant	.75	strong	.34	explanatory	.20	sketchy	-.13
pleasing	.68	efficient	.33	robust	.19	obscure	-.15
appealing	.62	useful	.32	exploratory	.17	careless	-.17
ingenious	.60	bold	.30	applicable	.15	trivial	-.18
striking	.59	illustrative	.29	elaborate	.15	weak	-.23
inspired	.58	subtle	.29	abstract	.13	incomplete	-.23
enlightening	.57	precise	.28	intricate	.12	tedious	-.24
creative	.55	clear	.27	minimal	.12	shallow	-.24
charming	.54	intuitive	.27	natural	.12	worthless	-.25
deep	.53	definitive	.26	general	.12	clumsy	-.26
insightful	.53	careful	.25	speculative	.08	confusing	-.27
sublime	.52	delicate	.25	simple	.08	inefficient	-.27
clever	.49	ambitious	.25	difficult	.08	awful	-.27
profound	.46	effective	.24	expository	.08	loose	-.27
innovative	.42	informative	.24	meticulous	.06	crude	-.28
cute	.41	non-trivial	.24	practical	.06	awkward	-.39
polished	.40	conceptual	.22	plausible	-.02	unpleasant	-.39
lucid	.36	rigorous	.22	flimsy	-.07	dull	-.47
fruitful	.35	unambiguous	.22	obvious	-.07	ugly	-.48

perceived simplicity, contrary to the classical view espoused by many mathematicians and philosophers.

6.2. **Beauty as Simplicity.** As noted in the introduction, the classical idea that mathematical proofs tend to be regarded as beautiful if they are simple has been supported by many notable mathematicians and philosophers (Chandrasekhar, 2010; Cherniwchan et al., 2013; Engler, 1990; McAllister, 2005; Wells, 1990; Tappenden, 2008). We found no evidence for this view. Neither 'beautiful' nor 'elegant' were strongly correlated with 'simple', and whereas the former adjectives loaded strongly onto the aesthetics dimension, the latter loaded (negatively) onto the intricacy factor. It seems that either claims made by mathematicians and philosophers about the relationship between beauty and simplicity are simply incorrect, or that when the words 'simple' and 'beautiful' are used by philosophers to discuss mathematical practice, they are being used in a substantially different way to how practising mathematicians use the terms.

How plausible is this latter suggestion? Are there several different senses of the word 'simple'? One can imagine referring to a major mathematical breakthrough as being a 'beautiful and simple' proof, and clearly this would be quite different to a 'dull and simple' proof produced by a mediocre mathematician in a low-ranked journal. Could it be that the word 'simple' has two (or more) distinct meanings, some positive and some

negative, and that the mathematicians in our sample were using it in one sense, whereas the philosophers and mathematicians cited in the introduction were using it in the other?

We do not believe that this is the case. The fact that 'simple' can be appropriately paired with both 'beautiful' and 'dull' merely indicates that simplicity is independent of the aesthetics dimension, not that 'simple' has two meanings. Whereas 'beautiful' loaded positively onto the aesthetics dimension, 'dull' loaded negatively onto the same dimension. Because our EFA indicated that aesthetics and intricacy are orthogonal factors, it should be (at least in principle) possible to write simple proofs which have positive, zero, or negative values on the aesthetics dimension. The intuitive plausibility of both 'beautiful and simple' and 'dull and simple' proofs, provides some support for this suggestion.

If simplicity and beauty are independent, why would so many mathematicians and philosophers link the two notions? One possible answer would be to suggest that, when asked to think of an example of a beautiful mathematical proof, mathematicians are biased towards thinking of simple proofs. The psychological literature provides some support for this proposal. In general a sentence's simplicity (as rated by independent judges) is reasonably well correlated with its memorability (Gruneberg, Monks, Sykes, & Oborne, 1974; Mehler, 1963). It seems plausible to suppose that the same is true of mathematical proofs. So it may be that, faced with the difficult task of recalling a beautiful proof from memory, there is a psychological bias towards recalling a simple example. Over time one can imagine that this tendency could have become reified in reflective accounts of mathematical practice as a direct relationship between the two notions. If this proposal is correct, then it has substantial implications for the reliability of mathematicians' reflective accounts in the philosophical study of mathematical practice, and we return to this issue later in the paper.

6.3. **Beauty as Epistemic Satisfaction.** Recall that Todd (2008) argues that the most parsimonious account of aesthetics in mathematics is that mathematicians' apparant aesthetic judgements are in fact epistemic judgements. He does, however, make clear that this view could be challenged by showing 'along what axes aesthetic and epistemic appreciation and judgement operate in relation to each other' (p. 75). Does our EFA provide such evidence? Among our eighty adjectives were several that one might view as being related to epistemic appreciation. For example, one might expect proofs that are 'enlightening', 'explanatory', 'expository', 'insightful', 'illustrative', 'informative', and 'plausible' to be particularly epistemically adequate. While some of these adjectives (notably 'enlightening' and 'insightful') loaded strongly onto the aesthetics dimension, most did not. Directly correlating these words with 'beautiful' resulted in a range of coefficients, shown in Table 4, from $+.57$ (enlightening) to $-.02$ (plausible), suggesting that there is no strong relationship between aesthetic and epistemic judgements. Moreover, all these adjectives, other than 'expository', loaded reasonably strongly ($\geq .3$) onto the utility dimension, perhaps indicating that epistemic judgements are related to the proof's utility for future mathematical work.

One interpretation of these data would be to suggest that epistemic and aesthetic judgements of mathematical proofs are indeed different. Epistemic judgements concentrate largely on the utility dimension, whereas aesthetic judgements concentrate on the aesthetics dimension. Nevertheless, there do appear to be adjectives which reside at the conjunction of these two dimensions. Proofs which are 'enlightening', 'fruitful', and

'insightful', for instance, will tend to score highly on the aesthetics *and* utility dimensions (these adjectives have reasonably high loadings on both dimensions).

6.4. **Beauty as Enlightenment.** Rota's (1997) account of mathematical beauty argues that mathematicians use the word as a 'copout' to avoid having to think about enlightenment. Further, he suggested that (i) a lack of beauty is characterised not by a lack of enlightenment, but rather by a lack of definitiveness; and that (ii) beauty and elegance are unrelated concepts: whereas beauty is related to enlightenment, elegance is a pedagogical term related to how the proof is presented, not to the content of the proof itself.

Before addressing Rota's (1997) primary claim, that a beautiful proof must necessarily be enlightening, we first note that his other suggestions do not fare well when tested against our empirical data. First, 'beautiful' and 'definitive' were poorly correlated ($r_s = .264, p < .001$), suggesting that it is unreasonable to characterise a lack of beauty by a lack of definitiveness. Second, beauty and elegance appear to be very strongly related concepts. Indeed, of all pairwise combinations of the eighty adjectives, 'beautiful' and 'elegant' had the highest correlation ($r_s = .752, p < .001$), higher even than 'pleasing' and 'appealing' ($r_s = .697, p < .001$).

As noted above, 'beautiful' and 'enlightening' were strongly correlated in our data ($r_s = .566, p < .001$). Does this support Rota's (1997) claim that beautiful proofs are those which provide enlightenment? There are several reasons to doubt Rota's account based on our data. First, as shown in Table 4, there were many considerably stronger correlates of 'beautiful' than 'enlightening', and these were significantly stronger in the case of 'elegant' ($t(252) = 4.339, p < .001$) and 'pleasing' ($t(252) = 2.130, p = .034$, Williams-Steiger tests). Given this, it seems difficult to accept that 'beautiful' and 'enlightening' have an identical meaning in the context of mathematical proofs.

Furthermore, the rationale for Rota's account, that unlike enlightenment, beauty does not admit degrees—a proof is either beautiful or not beautiful—is not supported by our data. In fact, only 9% of our participants claimed that 'beautiful' was a very inaccurate description of their chosen proof, and only 18% claimed that it was a very accurate description. The remaining 73% of participants chose one of the three intervening options when rating their proof on its beauty. So, contrary to Rota's suggestion, it seems that mathematicians are perfectly happy to say that a proof is moderately beautiful.

6.5. **Empirical research in the philosophy of mathematics.** In recent years, several philosophers have suggested that the philosophy of mathematics could be productively informed by empirical research on mathematicians' behaviour. Donald Gillies (2014), for instance, describes how numbers were constructed in different societies through history, and suggests that this causes problems for a Brouwerian constructivist account of number. Using this case as an illustrative example, he argues that sociological evidence could make useful contributions to philosophical debates (although Gillies focuses on sociology, his argument applies with equal force to other empirical traditions, such as social or cognitive psychology). Similar arguments have been made elsewhere (e.g., Giaquinto, 2007; Löwe et al., 2010; Van Kerkhove & Van Bendegem, 2006).

As will be clear from the content of this paper, we are extremely sympathetic to this point of view, but we do not wish to defend it here at length. Instead, we point out that

our data suggest that there are serious weaknesses with the types of empirical evidence that have traditionally dominated our understanding of mathematical practice. Typically, descriptions of higher-level mathematical research consist of introspective accounts from well-regarded mathematicians (e.g., Hadamard, 1945/1954; Poincaré, 1913; Pólya, 1954; Thurston, 1994). Poincaré (1913), for instance, wrote at length about how he developed the ideas contained in his first memoir on Fuchsian functions (even describing precise locations where ideas first entered his consciousness).

Unfortunately, a great deal of empirical research has conclusively shown that humans are extremely bad at accurately reporting their thought processes (Dutton & Aron, 1974; Hoffman, 1992; Horsey, 2002; Nisbett & Wilson, 1977), including in mathematics (Inglis & Alcock, 2012; Weber, 2008). A striking recent demonstration of this phenomenon was reported by Johansson, Hall, Silkstöm, and Olsson (2005), who showed participants photographs of two faces and asked them to choose the most attractive. Participants were then given a closer look at their chosen photograph and asked to verbally explain their choice. However, on some occasions participants were surreptitiously given the wrong photograph, containing the face they had not chosen. Around three-quarters of the time participants failed to notice and, moreover, often confabulated detailed explanations of why they had chosen this card ('I like earrings!'). Of course, these explanations could not possibly reflect the participant's actual decision-making processes, as they had in reality chosen the other card. In short, when we are prompted for introspective accounts of our behaviour, there is strong evidence that we often 'tell more than we can know' (Nisbett & Wilson, 1977).

Many mathematicians and philosophers have associated simplicity with beauty (Engler, 1990; Chandrasekhar, 2010; Cherniwchan et al., 2013; McAllister, 2005; Wells, 1990; Tappenden, 2008). However, in this paper we have shown that, when mathematicians are asked to rate a specific proof on these two characteristics, there appears to be no relationship between them. One reasonable account of this conflict is that mathematicians, when pressed to give a coherent account of their judgements, confabulate in much the same way as the participants in Johansson et al.'s (2005) study. In fact, associating beauty with simplicity is a particularly sensible confabulation because, as we have noted above, simplicity aids memorability, and so it is reasonable to suppose that those beautiful proofs which come to mind when prompted are the simpler ones.

But given that we should be wary of mathematicians' introspective accounts of their choices, why should we be any more convinced of the choices themselves? Why should we believe that asking a mathematician to say whether or not they believe a specific proof is 'beautiful' should give us any insight into the beauty of the proof? In our view, answering this question satisfactorily requires determining how consistent the language involved in mathematical evaluations is between individuals.

In this paper we have provided evidence that the language by which mathematicians evaluate proofs has a coherent internal structure which, roughly speaking, can be described using four-dimensional space. We have not demonstrated that there is any between-mathematician consistency in their characterisations of proofs, only that there is a degree of within-mathematician consistency. We do not know whether, if asked to rate the same proof, there would be widespread agreement about where the proof falls on each of the four dimensions, but we do know that, for a given mathematician, if a proof rates highly

on one adjective associated with a given dimension, it is likely to rate highly on other adjectives associated with that dimension. Further investigations would be required to determine the between-participant consistency of mathematicians' judgements about proofs (cf. Geist, Löwe, & Van Kerkhove, 2010).

In broader terms, the four-dimensional framework developed in this paper holds great promise for future research. In the first place, there are additional philosophically contentious aspects of mathematical proof, such as its explanatory character, upon which such data may also shed light. Secondly, there is considerable scope for new studies applying the framework to the analysis of other aspects of mathematical practice related to proof, some of which are suggested above. Lastly, we believe the methodology developed in this paper significantly extends the arsenal of techniques available to experimental philosophy. An analogous approach may be of value in the empirical investigation of a wide range of questions of philosophical significance.

7. Funding

This work was supported by a Royal Society Worshipful Company of Actuaries Research Fellowship to MI. AA is grateful to Florida Institute of Technology for granting sabbatical leave.

8. Acknowledgements

We are extremely grateful to Lara Alcock and Dirk Schlimm for providing insightful comments on earlier versions of this work. Early drafts of this paper were presented at the Loughborough Proof Reading Workshop (2013), the Mathematical Cultures Research Network (London, 2013), the Second International Meeting of the Association for the Philosophy of Mathematical Practice (Urbana-Champaign, 2013), and the Rutgers Proof Comprehension Workshop (2014), and we thank the audiences for their valuable remarks.

References

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs*, *47*, 211.

Barnard, R., & Ulatowski, J. (2013). Tarski's material adequacy condition and truth as "commonsense". *Manuscript submitted for publication*.

Boyce, C. J., Wood, A. M., & Powdthavee, N. (2013). Is personality fixed? personality changes as much as 'variable' economic factors and more strongly predicts changes to life satisfaction. *Social Indicators Research*, *111*, 287-305.

Breitenbach, A. (in press). Beauty in proofs: Kant on aesthetics in mathematics. *European Journal of Philosophy*.

Burt, C., & Banks, C. (1946). A factor analysis of body measurements for British adult males. *Annals of Human Genetics*, *13*, 238-256.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245-276.

Chandrasekhar, S. (2010). Beauty and the quest for beauty in science. *Physics Today*, *63*, 57-62.

Cherniwchan, C., Ghassemi, A., & Keating, J. (2013). *Mathematical ethnographies: Is a mathematical proof beautiful?* Retrieved 13th October 2013, from `http://www.maths.bris.ac.uk/research/videos/beautiful/`

Crockett, C. (1959). An attack upon revelation in semantics. *Journal of Philosophy*, *56*(3), 103-111.

Deary, I. J. (2001). *Intelligence: A very short introduction.* Oxford: Oxford University Press.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment*, *18*, 192-203.

Dutton, D. G., & Aron, A. P. (1974). Some evidence for heightened sexual attraction under conditions of high anxiety. *Journal of Personality and Social Psychology*, *30*, 510-517.

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

Feltz, A., & Cokely, E. (2013). Predicting philosophical disagreement. *Philosophy Compass*, *8*, 978-989.

Geist, C., Löwe, B., & Van Kerkhove, B. (2010). Peer review and testimony in mathematics. In B. Löwe & T. Müller (Eds.), *PhiMSAMP. Philosophy of mathematics: Sociological aspects and mathematical practice* (pp. 155–178). London: College Publications.

Giaquinto, M. (2007). *Visual thinking in mathematics.* Oxford: Oxford University Press.

Gillies, D. (2014). Should philosophers of mathematics make use of sociology? *Philosophia Mathematica*, *22*, 12-34.

Gowers, T. (2002). *Mathematics: A very short introduction.* Oxford: Oxford University Press.

Gruneberg, M. M., Monks, J., Sykes, R. N., & Oborne, D. J. (1974). Some correlates of rated memorability of sentences. *British Journal of Psychology*, *65*, 519-527.

Hadamard, J. (1945/1954). *The psychology of invention in the mathematical field.* New York: Dover Publications.

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

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

Hayes, N., & Joseph, S. (2003). Big 5 correlates of three measures of subjective well-being. *Personality and Individual Differences*, *34*, 723-727.

Heinze, A. (2010). Mathematicians' individual criteria for accepting theorems and proofs: An empirical approach. In *Explanation and proof in mathematics: Philosophical and educational perspectives* (pp. 101–111). Berlin: Springer.

Hoffman, R. R. (Ed.). (1992). *The psychology of expertise.* New York: Springer.

Holden, H., & Piene, R. (2009). *The Abel prize 2003–2007: The first five years.* Heidelberg: Springer.

Holden, H., & Piene, R. (2013). *The Abel prize 2008–2012.* Heidelberg: Springer.

Horn, J. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179-185.

Horsey, R. (2002). The art of chicken sexing. *UCL Working Papers in Linguistics*, *14*, 107-117.

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

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

Johansson, P., Hall, L., Silkstöm, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, *310*, 116-119.

John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxomony: History, measurement, and conceptual issues. In *Handbook of personality: Theory and research* (pp. 114–158). New York: Guilford.

Kaiser, H. (1974). An index of factorial simplicity. *Psychometrika*, *39*, 31-36.

Kivy, P. (1991). Science and aesthetic appreciation. *Midwest Studies in Philosophy*, *16*, 180-195.

Knobe, J. (2007). Experimental philosophy. *Philosophy Compass*, *2*, 81-92.

Koichu, B., Katz, E., & Berman, A. (2007). What is a beautiful problem? An undergraduate students' perspective. In J. H. Woo, H. C. Lew, K. S. Park, & D. Y. Seo (Eds.), *Proceedings of the 31st Conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, p. 113-120). Seoul: PME.

Löwe, B., Müller, T., & Müller-Hill, E. (2010). Mathematical knowledge: A case study in empirical philosophy of mathematics. In B. Van Kerkhove, J. De Vuyst, & J. P. Van Bendegem (Eds.), *Philosophical perspectives on mathematical practice* (pp. 185–203). London: College Publications.

McAllister, J. W. (1996). *Beauty and revolution in science*. Ithaca, NY: Cornell University Press.

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

Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, *90*, 862-877.

Mehler, J. (1963). Some effects of grammatical transformations on the recall of English sentences. *Journal of Verbal Learning and Verbal Behavior*, *2*, 346-351.

Montaño, U. (2012). Ugly mathematics: Why do mathematicians dislike computer-assisted proofs? *The Mathematical Intelligencer*, *34*, 21-28.

Montaño, U. (2014). *Explaining beauty in mathematics: An aesthetic theory of mathematics*. Dordrecht: Springer.

Naess, A. (1938). Common sense and truth. *Theoria*, *4*, 39-58.

Naess, A. (1981). The empirical semantics of key terms, phrases, and sentences: Empirical semantics applied to nonprofessional language. In S. Kanger & S. Öhman (Eds.), *Philosophy and grammar: Papers on the occasion of the quincentennial of Uppsala University* (pp. 135–154). Dordrecht: Reidel. (Reprinted in *The Selected Works of Arne Naess*, [Vol. 8, pp. 59-78]. Dordrecht: Springer, 2005.)

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231-295.

Papert, S. (1980). *Mindstorms: Children, computers and powerful ideas.* New York: Basic Books.

Paunonen, S. V., & Ashton, M. C. (2001). Big five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology*, *81*, 524-539.

Poincaré, H. (1913). *The foundations of science.* New York and Garrison: The Science Press.

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

Pólya, G. (1954). *Mathematics and plausible reasoning: Induction and analogy in mathematics.* Princeton, NJ: Princeton University Press.

Reid, C. (1986). *Hilbert Courant.* New York: Springer.

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

Schmitt, D. P., Allik, J., Mccrae, R. R., Benet-Martínez, V., Alcalay, L., Ault, L., et al. (2007). The geographic distribution of big five personality traits: Patterns and profiles of human self-description across 56 nations. *Journal of Cross-Cultural Psychology*, *38*, 173–212.

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

Sinclair, N. (2011). Aesthetic considerations in mathematics. *Journal of Humanistic Mathematics*, *1*, 2-32.

Singh, S. (1997). *Fermat's last theorem.* London: Fourth Estate.

Stakhov, A. P. (2009). *Mathematics of harmony: From Euclid to contemporary mathematics and computer science.* Singapore: World Scientific.

Stewart, D. W. (1981). The application and misapplication of factor analysis in marketing research. *Journal of Marketing Research*, *18*, 51-62.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics.* Upper Saddle River, NJ: Pearson Allyn & Bacon.

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

Tappenden, J. (2008). Mathematical concepts and definitions. In P. Mancosu (Ed.), *The philosophy of mathematical practice* (pp. 256–275). Oxford: Oxford University Press.

Thurston, W. P. (1994). On proof and progress in mathematics. *Bulletin of the American Mathematical Society*, *30*, 161-177.

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

Tupes, E. C., & Christal, R. C. (1961). *Recurrent personality factors based on trait ratings.* Lackland Air Force Base, TX: US Air Force.

Vakil, R. (2013). *Math 216: Foundations of algebraic geometry.* Stanford, CA: math216.wordpress.com.

Van Kerkhove, B., & Van Bendegem, J. P. (Eds.). (2006). *Perspectives on mathematical practices: Bringing together philosophy of mathematics, sociology of mathematics, and mathematics education.* Dordrecht: Springer.

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

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

Appendix A. Supplementary Tables

Repeating the primary EFA on the 60 adjectives which had mean ratings significantly greater than 2 (on the 1 to 5 Likert scale) resulted in four factors which were essentially identical to the original Factors 1, 3, 4 and 5. The table of factor loadings for this analysis is given in Tables 5 and 6.

Table 5. Adjectives which loaded strongly onto Factor 1 (aesthetics), from an EFA on the 60 adjectives with mean ratings significantly greater than 2. Loadings greater than ±.4 are shown in bold.

Adjective	F1	F2	F3	F4
ingenious	**0.797**	-0.055	0.146	0.110
striking	**0.785**	0.183	0.045	0.107
inspired	**0.766**	0.121	0.020	0.157
creative	**0.741**	0.073	0.049	0.194
beautiful	**0.720**	0.162	-0.206	0.244
profound	**0.715**	0.126	0.272	0.020
elegant	**0.712**	0.133	-0.290	0.316
deep	**0.683**	0.118	0.335	0.074
innovative	**0.674**	0.056	0.178	0.144
sublime	**0.663**	0.180	0.105	0.145
charming	**0.643**	0.256	-0.204	0.054
clever	**0.599**	0.064	0.083	0.154
appealing	**0.585**	0.320	-0.331	0.227
pleasing	**0.579**	0.301	-0.278	0.317
enlightening	**0.528**	0.381	-0.044	0.238
insightful	**0.501**	0.316	0.032	0.181
bold	**0.493**	0.219	0.321	-0.034
strong	**0.474**	0.351	0.270	0.297
subtle	**0.470**	-0.196	0.173	0.046
delicate	**0.468**	-0.094	0.284	0.105
ambitious	**0.463**	0.273	0.320	-0.042
cute	**0.423**	0.245	-0.245	-0.007
sharp	**0.418**	0.295	0.041	0.315

TABLE 6. Adjectives which loaded strongly onto Factors 2 (utility), 4 (intricacy), and 5 (precision), from an EFA on the 60 adjectives with mean ratings significantly greater than 2. Loadings greater than ±.4 are shown in bold.

Adjective	F1	F2	F3	F4
practical	-0.087	**0.674**	-0.063	0.084
informative	0.169	**0.621**	0.061	0.293
applicable	0.065	**0.589**	0.098	0.200
intuitive	0.093	**0.575**	-0.253	0.193
natural	-0.046	**0.548**	-0.169	0.186
illustrative	0.200	**0.545**	-0.125	0.139
efficient	0.260	**0.525**	-0.199	0.204
useful	0.161	**0.518**	0.068	0.283
explanatory	0.072	**0.447**	-0.191	0.257
conceptual	0.265	**0.417**	0.191	0.004
plausible	-0.130	**0.415**	-0.053	0.202
fruitful	0.390	**0.405**	0.163	0.243
robust	0.216	0.369	0.167	0.290
expository	0.065	0.368	0.001	0.125
exploratory	0.241	0.350	0.166	-0.108
general	0.205	0.311	0.162	-0.053
minimal	0.115	0.228	-0.158	-0.067
difficult	0.287	-0.088	**0.698**	-0.012
dense	0.040	-0.002	**0.677**	0.022
intricate	0.258	-0.032	**0.624**	0.319
confusing	-0.130	-0.088	**0.599**	-0.295
tedious	-0.264	0.049	**0.596**	0.001
elaborate	0.289	-0.022	**0.573**	0.214
simple	-0.076	0.263	**-0.558**	-0.068
abstract	0.214	0.143	**0.438**	-0.018
non-trivial	0.404	-0.068	**0.412**	0.263
precise	0.166	0.206	0.088	**0.747**
careful	0.186	0.078	0.208	**0.679**
rigorous	0.114	0.128	0.076	**0.649**
accurate	0.071	0.285	0.042	**0.602**
clear	0.149	0.230	**-0.432**	**0.540**
meticulous	0.039	0.128	**0.443**	**0.494**
polished	0.242	0.309	-0.158	**0.464**
unambiguous	0.156	0.223	-0.088	**0.421**
lucid	0.298	0.277	-0.093	0.400
effective	0.214	0.356	0.012	0.397
definitive	0.200	0.319	0.132	0.355