DIVERSITY IN PROOF APPRAISAL

MATTHEW INGLIS[1] AND ANDREW ABERDEIN[2]

Abstract. We investigated whether mathematicians typically agree about the qualities of mathematical proofs. Between-mathematician consensus in proof appraisals is an implicit assumption of many arguments made by philosophers of mathematics, but to our knowledge the issue has not previously been empirically investigated. We asked a group of mathematicians to assess a specific proof on four dimensions, using the framework identified by Inglis and Aberdein (2014). We found widespread disagreement between our participants about the aesthetics, intricacy, precision and utility of the proof, suggesting that *a priori* assumptions about the consistency of mathematical proof appraisals are unreasonable.

1. Proof Appraisals

A clichéd view of research-level mathematics, or at least research-level pure mathematics, is that it is simply and solely concerned with logic: purported proofs are either valid or invalid, and the job of a mathematician is to produce as many valid ones as possible. On this account, there is little place for the appraisal of proofs in anything other than a straightforwardly descriptive fashion. Proofs might be valid or invalid, published or unpublished, short or long, but under the clichéd view it is hard to see how they could be elegant, beautiful, or deep. However, a cursory glance at mathematical practice reveals that mathematicians regularly make such appraisals. For example, in the citation for the 2003 Abel Prize, Jean-Pierre Serre's work was described as being "profound", "spectacular", and "magnificent". So it is clear that a mathematician's appraisal of a given piece of work can go well beyond its validity. In this chapter we specifically focus on the evaluation of mathematical proofs: how are such appraisals made, and what is their status?

A more systematic investigation of the ways in which mathematicians characterise mathematical proofs is given in Table 1. It shows the most common adjectives used to characterise proofs on MathOverflow, a website where research mathematicians ask and answer questions about each others' research.[1] Although many of the adjectives are straightforwardly descriptive ('original', 'short', 'direct', 'algebraic', 'new', 'combinatorial')

[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: September 25, 2014.

[1]We downloaded an archive of all posts made on MathOverflow between September 2009 and May 2013. After data cleaning (i.e. removing html links and so on) this yielded a corpus of 1.83 million words of what could reasonably be called informal mathematical discourse. We searched for all two-word clusters with 'proof' as the second word. This yielded a total of 21,208 occurrences. Of course, many of these clusters were not adjectival: nearly half were "the proof" (27.3%) or "a proof" (17.9%). Table 1 shows all those adjectival clusters with 20 or more occurrences.

these data make clear that mathematicians regularly used their aesthetic judgement (proofs can be 'nice', 'slick', 'elegant', 'conceptual', 'beautiful').

The status of proof appraisals is an issue of fundamental interest to mathematical practice researchers. Several approaches to studying such appraisals have been adopted during the course of the Mathematical Cultures project. Hanna and Mason (2014, MC2) suggested that studying what makes a proof memorable provides an avenue in which less well-specified characteristics can be investigated. They adopted Gowers's (2007) notion of the 'width' of a proof, and related this idea to Raman's (2003) characterisation of 'key ideas'. Raman (MC2) herself hypothesised that the notion of mathematical beauty could be related to 'fit'. She analysed two proofs of Pythagoras's Theorem and argued that one exhibited 'intrinsic fit', in the sense that the proof captured "the essence of why the theorem is true", and that one did not.

This chapter builds directly on two further contributions to the Mathematical Cultures project. Ernest (MC2) discussed different types of mathematical value, and noted that "it is an open controversy as to whether beauty and aesthetics are objective or subjective mathematical values". Ernest's question is fundamental to understanding the status of mathematicians' proof appraisals, and our goal in this chapter is to investigate it empirically. But first we need to qualify Ernest's distinction between objective and subjective values. This might be understood as turning on whether an appraisal of the value in question is factive, that is, on whether it reports upon matters of fact. This is not an issue we can resolve empirically. However, we can empirically determine whether a consensus of mathematicians are in agreement over a specific appraisal. Strictly speaking, these questions are conceptually distinct: a spurious consensus may arise if mathematicians are all wrong in the same way, whether or not there is even a fact of the matter; conversely, mathematicians may disagree over an issue that is factive. Yet, it seems to us to be prima facie implausible that appraisals of mathematical values such as beauty or explanatoriness should motivate this sort of distinction: how might a proof be explanatory if no mathematician finds it so, or lack beauty although most mathematicians regard it as beautiful? Thus, although we wish to remain agnostic whether mathematical values are factive, we hold it to be sufficient for present purposes to focus on the following empirical question: are these values *subjective*, in the sense that they are primarily an idiosyncratic property of the mathematician doing the judgement? Or are they *intersubjective*, broadly shared across the community of mathematicians? To answer this question we build on our work that was reported at the second Mathematical Cultures Conference (Inglis & Aberdein, 2014, MC2). Before reviewing this contribution we offer some remarks about the importance of the subjectivity/intersubjectivity distinction for the validity of typical arguments deployed by philosophers of mathematics.

2. The Exemplar Philosophers

A common methodological move made by philosophers of mathematics is to offer an example of a proof, or a mathematical object, assert that the proof has a given property, and appeal to the readers' intuitions for agreement. Here we characterise those who adopt this approach as *exemplar philosophers*. Perhaps the most clear cut example of a discussion between exemplar philosophers concerns mathematical explanation. Steiner (1978) proposed an account of explanation based on characterising properties, which he

TABLE 1. Most frequent adjectives used to describe proofs on MathOverflow, those adjectives with a frequency less than 20 are omitted. Percentages are of all 2-word clusters, including non-adjectives ('the proof' constituted 27% of 2-word clusters).

Cluster	Raw Freq	% Freq
elementary proof	269	1.27
simple proof	223	1.05
original proof	164	0.77
short proof	156	0.74
direct proof	147	0.69
standard proof	117	0.55
formal proof	107	0.50
algebraic proof	104	0.49
complete proof	95	0.45
nice proof	92	0.43
usual proof	91	0.43
rigorous proof	84	0.40
new proof	83	0.39
easy proof	82	0.39
first proof	80	0.38
constructive proof	78	0.37
combinatorial proof	77	0.36
simpler proof	61	0.29
quick proof	59	0.28
geometric proof	55	0.26
theoretic proof	54	0.25
bijective proof	47	0.22
full proof	42	0.20
general proof	42	0.20
alternative proof	41	0.19
detailed proof	41	0.19
slick proof	38	0.18
analytic proof	37	0.17
mathematical proof	37	0.17
elegant proof	36	0.17
classical proof	35	0.17
inductive proof	32	0.15
conceptual proof	31	0.15
correct proof	29	0.14
consistency proof	28	0.13
shortest proof	28	0.13
topological proof	28	0.13
beautiful proof	23	0.11
similar proof	23	0.11
probabilistic proof	21	0.10
published proof	21	0.10
valid proof	20	0.09

defined to be "a property unique to a given entity or structure within a family or domain of such entities or structures" (p. 143). He suggested that an explanatory proof was one which "makes reference to a characterising property of an entity or structure mentioned in the theorem, such as that from the proof it is evident that the result depends on the property" (p. 143).

Steiner's argument is a model of the exemplar philosophers' approach. First he rejected an earlier characterisation of explanatoriness (Feferman's (1969) suggestion that explanatory proofs were those which are more general) by offering a proof of Pythagoras's theorem about which "it would be hard to claim that" it were more explanatory than the standard proof, despite it being more general (p. 139). Next, Steiner offered his own characterisation and justified it with reference to a proof (of the lemma that there are no integers a and b such that $a^2 = 2b^2$) which he claimed was explanatory, and which satisfied his characterisation. Finally, Steiner offered an example of a supposedly non-explanatory proof of the identity $1 + 2 + 3 + \ldots + n = n(n+1)/2$, and showed that it did not satisfy his characterisation. At each point in his argument Steiner asserted that the proofs he presented were exemplars of explanatoriness or non-explanatoriness, and offered no justification beyond his own judgement and an implicit appeal to his readers' intuitions.

Steiner's characterisation of explanatoriness was criticised by Resnik and Kushner (1987), again using the exemplar approach. They offered two proofs, one "that meets Steiner's criterion but doesn't explain and one which ought to explain if any proof does but fails to meet Steiner's criterion" (p. 146). With respect to the second proof, of the intermediate value theorem, the authors suggested that it was explanatory because "We find it hard to see how someone could understand this proof and yet ask why the theorem is true" (p. 149). As with Steiner then, no substantive evidence was offered, beyond their own judgement, for the (non-)explanatoriness of Resnik and Kushner's exemplar proofs.

This reliance on personal intuitions was criticised by Hafner and Mancosu (2005), who suggested that relying upon exemplar proofs identified by philosophers might not accurately reflect those proofs considered explanatory by working mathematicians. They offered a different exemplar to challenge Steiner, a proof of Kummer's Convergence Test produced by Pringsheim. In contrast to Steiner and Resnik and Kushner, Hafner and Mancosu appealed to the intuition of the proof's author, not simply to their own intuitions: "According to Pringsheim this proof gives 'the true reason why the C_n [...] can eventually be replaced by *completely arbitrary positive* numbers B_n' " (p. 229). Based on this evidence, Hafner and Mancosu concluded that their proof was genuinely an exemplar of explanatoriness, and used it to probe the adequacy of Steiner's characterisation.

A critical assumption of all existing exemplar accounts of explanatoriness is that intuitions about whether a proof is explanatory or non-explanatory are widely shared.[2] For the arguments offered by Steiner and Resnik and Kushner to persuade, it is crucial that their own judgements about explanatoriness are representative of the mathematical community at large. Hafner and Mancosu require a weaker assumption, that the judgement made by Pringsheim, a working mathematician, is representative of the larger mathematical community. Nevertheless, both approaches require the assumption of intersubjectivity,

[2]Note that, although we have focused on explanatoriness here, there are accounts of other mathematical values which use exemplars (e.g., Montaño, 2014; Tappenden, 2008a, 2008b).

that judgements about the properties of proofs are broadly shared across the mathematical community. If this assumption did not hold, then the domain of applicability of Steiner's theory would be substantially smaller than the whole mathematical community (it would merely be the collection of mathematicians who shared Steiner's intuitions, a group of unknown size). Further, as Ernest (MC2) pointed out, whether or not the assumption of intersubjectivity is reasonable is currently an unresolved open question. If the assumption is incorrect, and if there are disagreements between mathematicians about the explanatoriness of the exemplars offered by the exemplar philosophers, then the whole exemplar approach to characterising mathematical qualities such as explanatoriness or beauty seems problematic.

Our goal in this chapter is to empirically investigate the extent to which proof appraisals are shared between mathematicians. To do this, we build on our earlier analysis of the structure of mathematical proof appraisals (Inglis & Aberdein, 2014, MC2).

3. The Structure of Proof Appraisals

In earlier work (Inglis & Aberdein, 2014) we argued that the ways in which mathematicians evaluate mathematical proofs can be considered an analogous problem to the ways in which people evaluate human personalities. In both cases there are a large number of adjectives which can be applied (proofs can be appealing, bold, dense, etc.; humans can be bashful, creative, moody, etc.). And in both cases some of these adjectives appear to capture very similar ideas (elegant proofs seem intuitively likely to also be characterised as beautiful; rude people seem likely to also be characterised as unsympathetic). Social psychologists have approached the study of human personalities by asking participants to rate how accurately a given person (perhaps themselves, perhaps an acquaintance) would be described by a long list of adjectives. These ratings can then be subjected to an exploratory factor analysis, a statistical procedure which clusters the adjectives depending on how well they are correlated. For example, if a person who is accurately described by the word 'unsympathetic' is highly likely to be accurately described by the word 'rude', then in some sense the two adjectives are measuring the same trait. One of the most robust findings in social psychology is the observation that human personality traits cluster around five broad dimensions (Donnellan, Oswald, Baird, & Lucas, 2006; John, Naumann, & Soto, 2008).

We adopted an analogous strategy by asking a large group of mathematicians to think of a specific proof that they had recently refereed or read, and to state how accurately a long list of adjectives described it. We found that there were four broad dimensions upon which mathematical proofs vary, which we labelled Aesthetics, Intricacy, Precision and Utility (Inglis & Aberdein, 2014).[3] Importantly, all other adjectives in our study could be approximated by linear combinations of these dimensions. For example, proofs were likely to be rated as explanatory if they were useful, precise and non-intricate.[4] Thus, to

[3]We also found a fifth group of adjectives which consisted of those which were uniformly poor descriptors of the participants' chosen proofs (e.g. very few of the participants' chosen proofs were characterised as careless, crude, or flimsy). We characterised these as the Non-Use adjectives.

[4]The adjective 'explanatory' had loadings of 0.101, 0.002, -0.308, 0.313 and 0.367 on the Aesthetics, Non-Use, Intricacy, Utility and Precision dimensions respectively.

investigate the subjectiveness of any of our original adjectives, including 'explanatory', it suffices to consider the subjectiveness of these four dimensions.

Although our earlier study gives an indication of the structure of the space in which mathematical appraisals operate, it does not indicate whether proof appraisals are idiosyncratic subjective judgements which vary greatly between mathematicians, or intersubjective judgements that are broadly shared across the mathematical community. In the studies reported in this paper we sought to address this question by asking a group of mathematicians to study a particular proof, and then appraise it within the four-dimensional space earlier identified. In order to achieve this aim, we first produced a short research instrument which could reliably capture mathematicians' proof appraisals. The two studies involved in this process are reported in the Appendix. In the next section, we describe the methods used in our main study.

4. Method, Procedure and Participants

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. Potential participants were asked to visit a website where the purpose of the study was explained. If they agreed to participate, they clicked through to the first page, where they were asked to state their research area (pure mathematics, applied mathematics or statistics) along with the AMS subject classification which best characterised their work, and their position (PhD student, postdoc, faculty with less than 5 years experience, or faculty with more than 5 years experience).

On the next page participants were presented with a proof of the Sylvester-Gallai theorem, shown in Figure 1, taken from *Proofs from the Book* (Aigner & Ziegler, 2000). Aigner and Ziegler attributed the proof to L. M. Kelly and described it as being "simply the best" (p. 63). We chose this proof because it seemed to be relatively accessible, but also non-trivial. In addition, because it appeared in *Proofs from the Book*, we had reason to believe that it would elicit aesthetic reactions from at least some participants.

After studying the proof, participants were asked to select how accurately each of the twenty adjectives shown in Table 2 described it. The adjectives were presented in a random order and participants were asked to respond using a five-point Likert scale (very inaccurate, moderately inaccurate, neither inaccurate nor accurate, moderately accurate, very accurate).

Finally, participants were thanked for their time and invited to contact the research team should they have any questions.

A total of 112 mathematicians completed the study, consisting of 47 PhD students, 12 postdocs, 52 faculty (of whom 11 had less than five years experience), and 1 participant who declined to answer. The majority, 83% of participants, were pure mathematicians, 15% were applied mathematicians, and only 2% were statisticians.

5. Results and Discussion

Internal consistency is a critical aspect of psychometric instrument development. Because such instruments typically consist of several different Likert-scale items designed to measure the same construct, it is important to determine that each of these items

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

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

Claim: This line ℓ_0 does it!

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

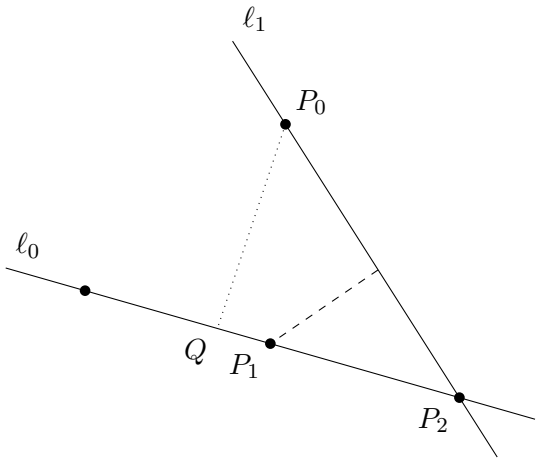

FIGURE 1. The proof used in the study, taken from Aigner and Zeigler (2000, p. 63).

results in a similar score to every other item which purportedly measures that construct. For example, the mini IPIP personality scales consist of four items for each of the five dimensions of human personality (Donnellan et al., 2006). If two items designed to assess how neurotic a person was resulted in substantially different responses, then we would say that the neuroticism dimension had poor internal consistency, and scores derived from it should be treated with caution. Typically the internal consistency of a scale is assessed using either a split-half reliability coefficient or the Cronbach's alpha statistic. To calculate the split-half reliability of a scale one calculates individuals' overall score on half the items (the odd numbered items, say) and correlates this figure with that from the other half of the items. If the scale has high internal consistency, this correlation coefficient (once adjusted for the reduced test length) should be high. The Cronbach's alpha coefficient

TABLE 2. The adjectives used in the short scale.

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

*reverse scored

results from a more involved calculation, but operates on the same principles and can be interpreted in an similar manner (e.g., Knapp & Mueller, 2010). Typically a split-half or alpha coefficient of 0.7 or greater is considered to indicate acceptably high internal consistency (e.g., Nunnally, 1978).

We first calculated the internal consistencies of the four dimensions. This yielded Cronbach's alphas of .877, .743, .839 and .797 for the Aesthetics, Intricacy, Precision and Utility dimensions respectively. As in the final pilot study (reported in the Appendix) all alphas were above the typical guideline of .7 (Nunnally, 1978).

We then calculated dimension scores for each participant by adding their Likert scale responses for each dimension (with "very inaccurate" given a score of 1, and "very accurate" a score of 5). The responses for 'simple' were reverse scored. This yielded four scores for each participant, one for each dimension, which could vary from 4 to 20 (so, for example, if a participant had a score of 20 on the Intricate dimension, they found the proof to be highly intricate). We then plotted the distributions of the scores, these are shown in Figure 2. Inspection of these histograms revealed widely spread distributions of scores for each of the four dimensions.

We further analysed participants' responses by conducting a hierarchical cluster analysis (using Ward's method with a Euclidean squared metric). This is a statistical procedure which attempts to cluster participants into groups based on the similarity of their scores

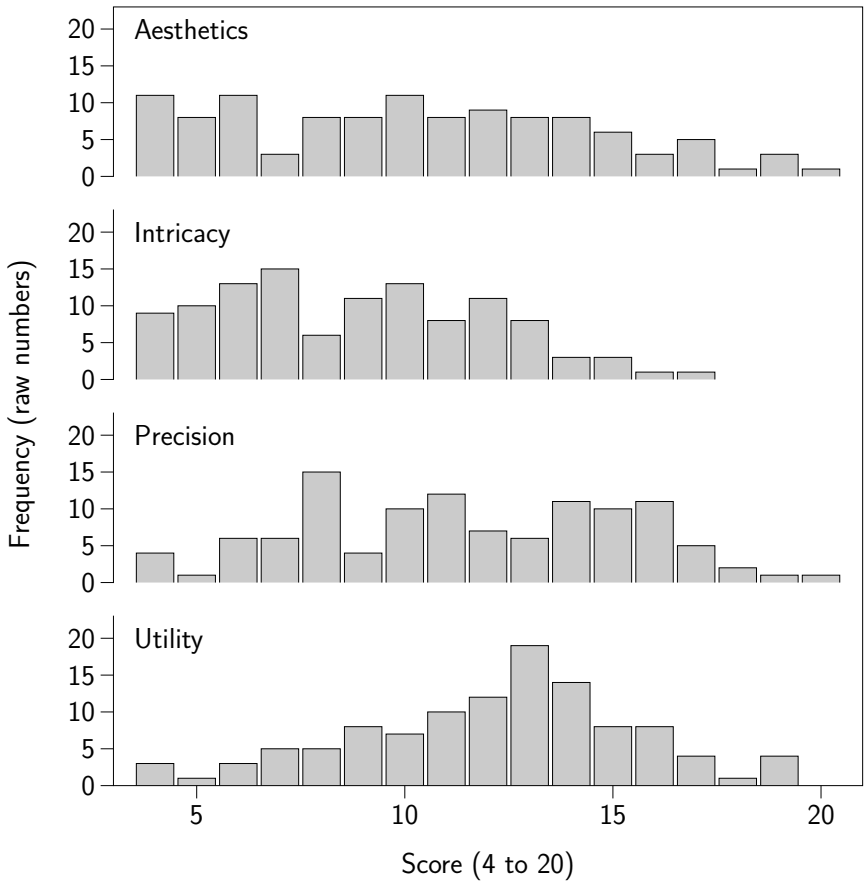

FIGURE 2. Histograms showing how participants rated the proof on each of the four dimensions.

on different dimensions. Inspection of the resulting dendrogram suggested that a three cluster solution was optimal. The mean ratings of each group for each dimension are shown in Figure 3. Participants in Cluster 1 ($N = 51$) rated the proof as being high on the Aesthetics, Precision and Utility dimensions, and low on the Intricacy dimension; participants in Cluster 2 ($N = 25$) rated it as being low on the Aesthetics, Intricacy and Precision dimensions, and high on the Utility dimension; and participants in Cluster 3 ($N = 36$) rated the proof as being low on all dimensions, and especially low on the Aesthetics dimension.

Next we investigated whether participants' responses could be predicted using their research area. We ran a multivariate analysis of variance (Manova), with research area (pure or applied mathematics) as the predictor, and scores on the four dimensions as the dependent variables. For the purposes of this analysis we classified the two statisticians in the sample as being applied mathematicians. Overall, there was no main effect of research area, $F(4, 107) = 1.733, p = .148$. Looking at the dimensions separately revealed a trend for the applied mathematicians to find the proof more intricate than the pure

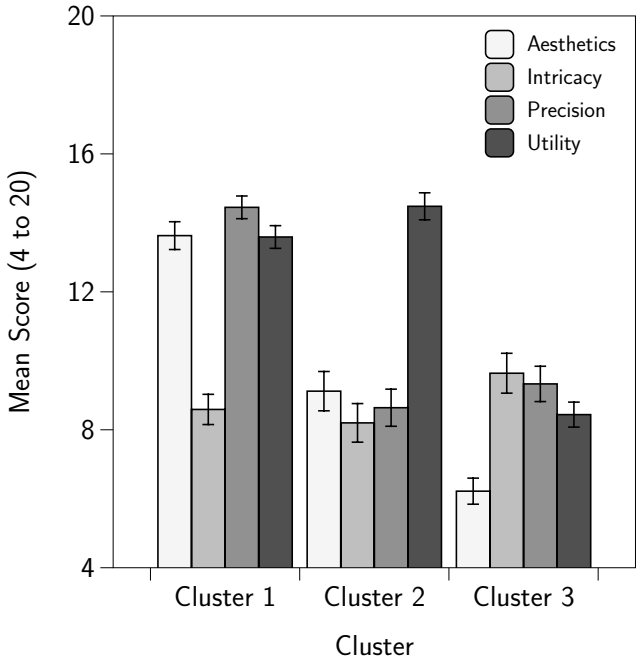

FIGURE 3. The mean ratings on each dimension of the three clusters. Error bars show ±1 SE of the mean.

mathematicians (mean ratings: 10.2, 8.6; $t(110) = 1.895, p = .061$), and a trend for the applied mathematicians to find the proof less useful than the pure mathematicians (means: 10.5, 12.4; $t(19.8) = 1.840, p = .081$), but neither of these trends approached the Bonferroni-corrected significance level of .013. Similarly, there was no significant relationship between cluster membership and research area, $\chi^2(2) = 4.038, p = .133$.

Finally, we ran a Manova predicting scores on the four dimensions with our career stage variable (because we had relatively few faculty with less than 5 years experience in our sample, we merged the experienced and inexperienced faculty categories in this analysis; this left three categories: research students, postdocs and faculty). We found no main effect of career stage, $F(8, 212) = 1.261, p = .265$, and neither did career stage predict any of the dimension scores individually. There was also no association between career stage and cluster membership, $\chi^2(4) = 4.181, p = .382$. Overall, we found no evidence that participants' appraisals of this proof were strongly predicted by either their research area or their career stage.

6. Discussion

We found a remarkable level of disagreement between our participants' ratings of the proof. For each of the four dimensions of proof appraisal there were participants who thought the proof should score high on that dimension, and there were participants who thought the proof should score low on that dimension. Furthermore, neither research area nor career stage seemed to be predictive of mathematicians' appraisals on any of the four dimensions.

Recall that the proof was taken from *Proofs from the Book*, a collection of proofs modelled on Paul Erdős's suggestion that there is a book "in which God maintains the perfect proofs for mathematical theorems, following the dictum of G. H. Hardy that there is no permanent place for ugly mathematics" (Aigner & Ziegler, 2000, p. V). Proofs from the book are said to contain "brilliant ideas, clever insights and wonderful observations" (p. V). Given this, we expected that the proof used in this study would be seen by most participants as being relatively strong on the aesthetics dimension. But this was not the case: there was widespread disagreement about how aesthetic the proof was, and in fact a majority of participants (60.4%) rated it below the midpoint (12) of the aesthetic scale, with just 31.5% rating it above the midpoint.

Recall that Ernest (MC2) noted that "it is an open controversy as to whether beauty and aesthetics are objective or subjective mathematical values". Our study here provides some support for the latter position. If beauty and aesthetics were objective, or at least intersubjective, we would have expected much greater clustering around a mean rating on the Aesthetics dimension. Indeed, our findings allow us to go further, and suggest that no qualities which can be represented as linear combinations of aesthetics, intricacy, precision and utility are intersubjective. Our findings are, in this respect, in line with other recent empirical evidence about mathematical practice. For example, Weber, Inglis, and Mejía-Ramos (2014) reviewed a series of studies which demonstrated that there is substantial heterogeneity about how persuasive mathematicians find different types of evidence for mathematical assertions.

Exemplar philosophers have typically relied upon their own intuitions about the qualities of a proof to draw philosophical conclusions. The data we have presented here suggest that these intuitions may not be widely shared, potentially causing a serious problem for this approach. Of course, our study involved only a single proof, and we cannot say that our findings would generalise to all mathematical arguments. In particular, perhaps the approach of the exemplar philosophers could be rescued by supposing that there *is* widespread agreement about the qualities of the proofs chosen as exemplars by the exemplar philosophers. We cannot refute this suggestion, but we do suggest that our data indicate that assuming *a priori* that there would be agreement is unwarranted. We have demonstrated that, for at least one proof (one which was deemed worthy of inclusion in *Proofs from the Book*), there is no consensus, so whether or not there is agreement among mathematicians about the qualities of any particular proof (including, for example, Steiner's (1978) exemplars) should be regarded as an open empirical question.

What of Hafner and Mancosu's (2005) alternative approach? They criticised Steiner (1978) and Resnik and Kushner (1987) for relying on personal intuitions, and instead appealed to the judgement of Pringsheim, the author of their exemplar proof. While this approach certainly seems preferable to Steiner's and Resnik and Kushner's, our data suggest that it still may be insufficient for Hafner and Mancosu's needs. While Pringsheim found his proof to be explanatory, our data suggest that it is entirely plausible that he was an outlier in this respect. Whether or not this is the case is uncertain, a matter which can only be resolved by sampling a sufficiently large number of mathematicians, and asking them to assess the explanatoriness of his proof.

Of course we are not the first to suggest that philosophers should be wary of assuming that their personal intuitions about semantics are widely shared. Concerns about the

validity of this assumption were central to the 'empirical semantics' approach of the Oslo Group in the early-to-mid twentieth century (e.g., Gullvåg, 1955; Naess, 1938, 1981; Tönnessen, 1955). Gullvåg, for instance, pointed out that any suggestion about a term's meaning "is merely an unsupported guess as long as no systematic testing of it has been attempted", and that "to test it systematically it is hardly sufficient that a single person registers his own reactions to this or that sentence, or makes pronouncements based on intuitions, or undertakes scattered observations of others' usage" (p. 343). Similar concerns are at the root of more recent work on experimental philosophy where, among other topics, empirical methods have been used to explore the generality of philosophers' intuitions about ethical dilemmas (e.g., Appiah, 2008; Nadelhoffer & Nahmias, 2007). The results we have presented here strongly suggest that analogous concerns are valid in the context of mathematical practice, and that empirical data which demonstrate that personal intuitions about exemplar proofs are shared (i.e. which demonstrate that exemplars are indeed exemplary) are necessary if the exemplar approach is to yield productive insights.

One unresolved question from our study concerns the origin of mathematicians' proof appraisals. If we are correct that there are large individual differences in how mathematicians evaluate proofs, and if these differences cannot be predicted by the mathematicians' experience or research area, then what lies behind these differences? This is a question for which we do not have a good answer, or even a good hypothesis. It seems ripe for future research.

Acknowledgements

We are grateful to Dirk Schlimm and an anonymous reviewer for insightful comments on an earlier draft of this chapter. This work was partially funded by a Royal Society Worshipful Company of Actuaries Research Fellowship (to MI).

Appendix A. Producing a Short Scale

The goal of the two studies reported in the Appendix was to create a short scale which could reliably capture mathematicians' proof appraisals. Specifically, we were concerned to develop an instrument which showed sufficiently high internal consistency on all four of the dimensions identified by Inglis and Aberdein's (2014) exploratory factor analysis.

We constructed our initial scale for testing by taking the four adjectives which had the highest loadings on each of the four dimensions. These are shown in Table 3. We also included four adjectives from the Non-Use dimension. Although we did not believe that this formed a genuine dimension, we felt it useful to include adjectives which were likely to elicit negative responses, in order to reduce the likelihood of participants simply selecting "very accurate" for each adjective.

A.1. **Study 1.**

A.1.1. *Method, Participants and Procedure.* Participants were 53 research-active mathematicians recruited from Australia, Canada and New Zealand. Departments in the three countries were invited by email to participate in the study. If they agreed, they forwarded an email to all research-active mathematicians in the department inviting them to participate. The email gave an outline of the purpose of the study, and provided a link

TABLE 3. The adjectives used in Studies 1 and 2. Changed adjectives are shown in italics.

Study 1		Study 2	
Adjective	Dimension	Adjective	Dimension
ingenious	Aesthetics	ingenious	Aesthetics
inspired	Aesthetics	inspired	Aesthetics
profound	Aesthetics	profound	Aesthetics
striking	Aesthetics	striking	Aesthetics
dense	Intricacy	dense	Intricacy
difficult	Intricacy	difficult	Intricacy
intricate	Intricacy	intricate	Intricacy
unpleasant	Intricacy	**simple*	*Intricacy*
careless	Non-Use	careless	Non-Use
crude	Non-Use	crude	Non-Use
flimsy	Non-Use	flimsy	Non-Use
shallow	Non-Use	shallow	Non-Use
careful	Precision	careful	Precision
meticulous	Precision	meticulous	Precision
precise	Precision	precise	Precision
rigorous	Precision	rigorous	Precision
applicable	Utility	applicable	Utility
efficient	Utility	*useful*	*Utility*
informative	Utility	informative	Utility
practical	Utility	practical	Utility

*reverse scored

to the study's website. Participants who clicked on the link first saw an introductory page which again explained the purpose of the study. 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, which were identical to those used by Inglis and Aberdein (2014):

> 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 twenty adjectives given in Table 3 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.

A.1.2. *Results and Discussion.* We calculated the internal consistency of each of the four dimensions (excluding the Non-Use dimension) using the Cronbach's alpha statistic. Recall that an alpha of 0.7 or above is typically considered acceptable (e.g., Nunnally, 1978). The Cronbach's alpha for each dimension are shown in Table 4.

TABLE 4. The Cronbach's alphas of each dimension in Studies 1 and 2, and the Main Study.

Dimension	Study 1	Study 2	Main Study
Aesthetics	.850	.864	.877
Intricacy	.595	.770	.743
Precision	.805	.788	.839
Utility	.485	.832	.797

The alphas associated with the Aesthetics and Precision dimensions were considerably above the 0.7 guideline, but those for Intricacy and Utility dimensions fell somewhat short, indicating a lack of consistency between the items on these dimensions. To address this problem we calculated, for these two dimensions, the item without which the resultant three-item scale had the highest alpha. These were 'unpleasant' and 'efficient' for the Intricacy and Utility dimensions respectively. We then replaced these items with two new adjectives, each of which had loaded strongly onto these dimensions in Inglis and Aberdein's (2014) factor analysis: 'simple' and 'useful'. Because a very simple proof would have a low score on the Intricacy dimension, we reverse scored the item (i.e. a participant choosing "very accurate" for 'simple' would be given a score of 1 rather than 5). We then conducted a second study to investigate the performance of our revised scale.

A.2. **Study 2.**

A.2.1. *Method, Participants and Procedure.* The procedure was identical to that of Study 1 except that 'efficient' was replaced with 'useful' on the Utility dimension, and 'unpleasant' with 'simple' on the Intricacy dimension (with 'simple' reverse scored). The full list of adjectives is given in Table 3. Participants were 53 research active mathematicians from universities in Ireland and Scotland. They were recruited in a similar manner to participants in Study 1.

A.2.2. *Results and Discussion.* The Cronbach's alphas for the four dimensions are given in Table 4. With the revised scale, all four alphas were above the 0.7 guidelines, suggesting that each dimension had acceptable internal consistency. We therefore used this revised scale in our main study.

References

Aigner, M., & Ziegler, G. (2000). *Proofs from the book* (Second ed.). Berlin: Springer.

Appiah, K. A. (2008). *Experiments in ethics*. Cambridge, MA: Harvard 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.

Feferman, S. (1969). Systems of predicative analysis. *Journal of Symbolic Logic*, *29*, 1–30.

Gowers, W. T. (2007). Mathematics, memory and mental arithmetic. In M. Leng, A. Paseau, & M. Potter (Eds.), *Mathematical knowledge* (pp. 33–58). Oxford: Oxford University Press.

Gullvåg, I. (1955). Criteria of meaning and analysis of usage. *Synthese*, *9*, 341–361.

Hafner, J., & Mancosu, P. (2005). The varieties of mathematical explanation. In P. Mancosu, K. F. Jørgensen, & S. A. Pedersen (Eds.), *Visualization, explanation and reasoning styles in mathematics* (pp. 215–250). Berlin: Springer.

Hanna, G., & Mason, J. (2014). Key ideas and memorability in proof. *For the Learning of Mathematics*, *34*(2), tbc.

Inglis, M., & Aberdein, A. (2014). Beauty is not simplicity: An analysis of mathematicians' proof appraisals. *Philosophia Mathematica*.

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.

Knapp, T. R., & Mueller, R. O. (2010). Reliability and validity of instruments. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 337–342). New York: Routledge.

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

Nadelhoffer, T., & Nahmias, E. (2007). The past and future of experimental philosophy. *Philosophical Explorations*, *10*, 123–149.

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.)

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw Hill.

Raman, M. (2003). Key ideas: What are they and how can they help us understand how people view proof? *Educational Studies in Mathematics*, *52*, 319–325.

Resnik, M. D., & Kushner, D. (1987). Explanation, independence and realism in mathematics. *British Journal of the Philosophy of Science*, *38*, 141–158.

Steiner, M. (1978). Mathematical explanation. *Philosophical Studies*, *34*, 135–151.

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

Tappenden, J. (2008b). Mathematical concepts: Fruitfulness and naturalness. In P. Mancosu (Ed.), *The philosophy of mathematical practice* (pp. 276–301). Oxford: Oxford University Press.

Tönnessen, H. (1951). The fight against revelation in semantical studies. *Synthese*, *8*, 225–234.

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