Synthese (2021) 198 (Suppl 26):S6369–S6392
https://doi.org/10.1007/s11229-019-02234-5

S.I. : ENABLING MATHEMATICAL CULTURES

Functional explanation in mathematics

Matthew Inglis[1] · **Juan Pablo Mejía-Ramos**[2]

Received: 31 May 2018 / Accepted: 25 April 2019 / Published online: 22 May 2019

Abstract
Mathematical explanations are poorly understood. Although mathematicians seem to regularly suggest that some proofs are explanatory whereas others are not, none of the philosophical accounts of what such claims mean has become widely accepted. In this paper we explore Wilkenfeld's (Synthese 191:3367–3391, 2014) suggestion that explanations are those sorts of things that (in the right circumstances, and in the right manner) generate understanding. By considering a basic model of human cognitive architecture, we suggest that existing accounts of mathematical explanation are all derivable consequences of Wilkenfeld's 'functional explanation' proposal. We therefore argue that the explanatory criteria offered by earlier accounts can all be thought of as features that make it more likely that a mathematical proof will generate understanding. On the functional account, features such as characterising properties, unification, and salience correlate with explanatoriness, but they do not define explanatoriness.

Keywords Explanation · Mathematics · Mathematical practice · Understanding

What are mathematical explanations? This question has generated substantial interest among philosophers. A number of competing accounts of mathematical explanation have been proposed (e.g., Kitcher 1981; Lange 2014; Steiner 1978), but all have well-established limitations. Our primary goal in this paper is to explore the consequences for mathematics of Wilkenfeld's (2014) notion of *functional explanation*. Roughly speaking, Wilkenfeld suggested that explanations are simply those things that, in an appropriate manner and at an appropriate time, generate understanding. We will argue that various philosophical accounts of mathematical explanation—including those offered by Steiner (1978), Kitcher (1981), and Lange (2014)—are all derivable consequences of a combination of Wilkenfeld's functional account and a modern understanding of human cognitive architecture. Consequently, we argue that Wilken-

✉ Matthew Inglis
m.j.inglis@lboro.ac.uk

1 Mathematics Education Centre, Loughborough University, Loughborough LE11 3TU, UK

2 Rutgers University, New Brunswick, USA

feld's account has a great deal of promise for philosophers interested in mathematical explanation.

It is common for mathematicians to assert that some proofs are explanatory whereas others are not. While all proofs establish their theorems, apparently some go further and explain *why* the theorem holds true. But what does it mean for a proof to be explanatory? This is a troubling question, as traditional accounts of scientific explanation seem not to work well in mathematical contexts. Two examples suffice to indicate the general problem. Salmon's (1971) statistical-relevance account of scientific explanation suggests that event X explains event Y if X and Y are probabilistically related to each other in a particular way. But this is problematic in mathematics: mathematical statements do not admit probabilities beyond zero or one, and so one cannot say that two events are statistically related (for instance, if $P(Y) \in \{0, 1\}$ then there does not exist an X for which $P(Y|X) > P(Y)$). Later in his career Salmon (1984) proposed an alternative account, the causal-mechanical model of scientific explanation. Under this proposal event A explains event B if there is a special kind of causal connection between A and B. But this too seems problematic in mathematical contexts, as mathematical statements are not causally connected to each other. It does not seem to be meaningful to say that the fact that a metric space is compact *causes* it to be separable: its compactness and separability are not temporal events to which the notion of causation can be easily applied.

These difficulties have even led some to propose that, contrary to the majority view, there is no coherent notion of explanation in mathematics, and that therefore mathematical statements can never be explained (e.g., Resnik and Kushner 1987; Zelcer 2013). In part this disagreement turns on the extent to which mathematicians describe themselves as explaining: whereas Steiner (1978) claimed that "mathematicians routinely distinguish proofs that merely demonstrate from proofs which explain" (p. 135), Resnik and Kushner (1987) contradicted him, asserting that mathematicians "rarely describe themselves as explaining" (p. 151). Both Zelcer (2013) and Weber and Frans (2017) suggested that an empirical investigation of research-level mathematical language was required to settle this question. A recent such analysis concluded that mathematicians do describe themselves (or their mathematical work) as explaining mathematics in their research papers, albeit around half as often as do physicists in their research papers, or does the general population in day-to-day English (Mejía-Ramos et al. 2019). However, these frequency data tell us nothing about what mathematical explanations actually are.

Before briefly reviewing three of the major accounts of mathematical explanation, we first clarify the type of mathematical explanations upon which we are focusing. Lyon and Colyvan (2008) distinguished between *intra-mathematical* explanations and *extra-mathematical* explanations. Both have mathematical statements as their explanans (what is doing the explaining), but they differ in the nature of their explanandums (what is being explained). Extra-mathematical explanations have non-mathematical explanandums, for instance Baker (2005) famously discussed the case of whether the length of the life cycle of a type of cicada (explanandum) can be explained by the properties of prime numbers (explanans). In contrast, in an intra-mathematical explanation both explanans and explanandum are mathematical: a mathematical proof might explain a particular mathematical theorem. Our focus here is, like Steiner's

(1978), Kitcher's (1981) and Lange's (2014), on intra-mathematical explanation, although we believe the functional account we discuss could straightforwardly be extended to extra-mathematical explanations. In the philosophical literature the focus has tended to be on *mathematical proofs* that explain, rather than other explanatory mathematical objects (e.g. an informal argument, or the justification of the aptness of a definition, D'Allesandro, in press). For this reason we too focus our discussion on proofs, although we believe both that non-proofs can offer intra-mathematical explanations, and that the functional account we outline can apply in these cases.

The remainder of the paper is organized as follows. In the next section we briefly review three accounts of mathematical explanation, those advanced by Steiner (1978), Kitcher (1981) and Lange (2014). We then motivate Wilkenfeld's (2014) notion of functional explanation by reflecting on the methods that have typically been used to warrant (and critique) these accounts, and by discussing the relationship between explanation and understanding. Having motivated it, we introduce the functional explanation account itself. By introducing a simple 'modal model' of human cognitive architecture, we then argue that Steiner, Kitcher, and Lange's accounts of explanation can all be derived from the functional account. Finally, we draw out some implications for future research on mathematical explanation, and suggest how the functional account can be empirically tested in mathematical contexts.

1 Accounts of mathematical explanation

1.1 Steiner's characterising properties

An early account of mathematical explanation was offered by Steiner (1978). He suggests that proofs are explanatory if they make critical use of some 'characterising property', defined to be those properties unique to "an entity or structure mentioned in the theorem, such that from the proof it is evident that the result depends on the property" (p. 143). For instance, on Steiner's account, the famous proof given by the schoolboy Gauss (that the sum of the first n integers is $n(n+1)/2$) is explanatory because it turns on the (characterising) symmetry properties of the sum $1+2+\cdots+n$. In contrast, a standard proof of the same result by induction is not explanatory, as it does not involve a characterising property of anything mentioned in the theorem (Steiner concedes that it may involve a characterising property of the set of natural numbers, but observed that this set is not mentioned in the theorem). Further, Steiner suggests that explanatory proofs can be generalised by varying the characterising property. By doing so, and by deforming the original proof in an analogous way, one can create an array of associated theorems and proofs. For instance, one can easily adapt Gauss's proof to derive a formula for the sum of the first n even numbers. One advantage of Steiner's approach is that it accounts for his observation that visual proofs are often explanatory: Steiner noted that characterising properties are very often visualisable (p. 146).

However, Steiner's account of mathematical explanation has been criticised on the grounds that it is sometimes difficult to find the characterising properties associated with proofs that mathematicians judge to be explanatory. For instance, Resnik and

Kushner (1987) 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). This led them to doubt Steiner's account, and also to call into question the more general claim that there is any objective distinction between explanatory and non-explanatory proofs. Hafner and Mancosu (2005) disagreed with Resnik and Kushner's scepticism about the general issue, but agreed that Steiner's criterion did not account for all mathematical explanations. Hafner and Mancosu offered a proof of Kummer's test of convergence, noted that it does not obviously use a characterising property, and reported that the proof's author explicitly described it as being explanatory. Similarly, Lange (2014) gave examples of apparently explanatory proofs which, while using characterising properties, could not obviously be deformed to create new proofs as Steiner's criterion required (for instance, he observed that a proof about isosceles trapezoids would collapse if the trapezoids were deformed to become non-isosceles). In sum, if Resnik and Kushner, Hafner and Mancosu, and Lange are right, then Steiner's focus on characterising properties cannot be the whole story behind explanation in mathematics.

1.2 Kitcher's unification

In contrast to Steiner (1978), Kitcher (1981) intended his account to apply to wider contexts than just mathematics. He suggests that the purpose of an explanation is to unify a group of different facts or concepts into a single theory. For instance, Darwin's theory of evolution unified an enormous variety of observed phenomena—from fossil records to observations of Galápagos finches—under a single account. It is therefore, on Kitcher's view, a highly explanatory theory. Kitcher suggests that two purported explanations can be compared by assessing their 'patterns of argument' (distinctive forms of argument): how 'stringent' these argument patterns are (roughly, how many restrictions are imposed on instantiations of the argument), and how many conclusions they have. Better explanations will use fewer, more stringent, argument patterns to draw more conclusions.

As with Steiner, Kitcher (1981) has been criticised with concrete examples of apparently explanatory proofs that do not seem to fit his account. For example, Hafner and Mancosu (2008) gave three different proofs of the same theorem, taken from the same textbook, and demonstrated that simply comparing the number of argumentation patterns produces an explanatory ranking inconsistent with how the textbook authors themselves judged the relative explanatoriness of the different proofs.

1.3 Lange's salience

Lange's (2014) proposal is in some ways similar to Steiner's (1978). He suggests that explanatory and non-explanatory proofs differ in the way that they use features in their theorems' premises. An explanatory proof, for Lange, is one that exploits a feature of the theorem's premises that is somehow salient in the theorem's result. For instance, Lange gave the example of a proof establishing that various outcomes in a question concerning coin tossing are equally probable. He argued: "The proof explains this

symmetry in the chances (namely, that each possible outcome has the same chance) by showing how it arises not from an algebraic miracle, but rather from a symmetry in the setup … [the] proof revealed the setup's hidden symmetry and thereby explained the result." (p. 496). But symmetry is not the only salient manner in which a proof can extract a theorem from premises. In a second example, Lange showed how 'feature unity' can accomplish the same task. He introduced a finite class of numbers, referred to as 'calculator numbers', and the theorem that all calculator numbers are divisible by 37. Two proofs were offered. The first was a brute force analysis where all calculator numbers are simply examined in turn. The other revealed a hidden property that all calculator numbers share, and from which the result immediately follows. Regarding this second proof, Lange wrote: "this proof traces the fact that every calculator number is divisible by 37 to a property that they have in common by virtue of being calculator numbers. In short, an explanation of this result consists of a proof that treats every calculator number in the same way" (p. 509). According to Lange, it is this 'feature unity' that is salient, and which therefore makes the second proof more explanatory than the first.

Like Steiner (1978) and Kitcher (1981), Lange's (2014) account has also been criticised. D'Alessandro (in press) points out that while Lange's approach allows for a theorem to explain another result via the theorem's explanatory proof (i.e. theorem T explains result R via explanatory proof P of T), mathematicians often make claims about what a statement would explain should it be true, without actually knowing whether the statement is true or false. For instance, D'Alessandro argues that $P \neq NP$ would explain a great many results in computability theory if it turned out to be true, but Lange's account—with its focus on salience between theorem and proof—has difficulty in accounting for this. There is, at present, no proof of $P \neq NP$, so clearly there is no saliency connection between the proof of $P \neq NP$, and results in computability theory.[1]

In sum, none of the accounts of mathematical explanation we have discussed here have gained widespread acceptance. In the remainder of the paper we suggest that this is because while each of the criteria offered by Steiner (1978), Kitcher (1981) and Lange (2014) are prototypical properties that one would associate with mathematical explanations, they are not defining characteristics.

2 Motivating a functional account

Our goal in this section is to motivate the functional account of explanation that we describe in the next section. To this end, we make two observations about the three accounts of mathematical explanation discussed above. First, we reflect on the evidence offered both in support of, and to critique, each of the accounts. Second, we note that none of the accounts directly address whether or how explanations can generate understanding in their audiences.

[1] Although, as pointed out by a reviewer, this does not preclude $P \neq NP$ featuring in a proof that is explanatory on Lange's account.

All three of the accounts outlined above share the property that they were justified by their proponents with reference to explicit examples of allegedly explanatory proofs. Steiner (1978), for instance, started his paper by rejecting an earlier characterisation of explanatoriness—Feferman's (1969) suggestion that explanatory proofs are more general than non-explanatory proofs—by offering a proof of Pythagoras's theorem that he judged to be more general but less explanatory. Steiner then offered his own characterisation of explanatoriness, and justified it with reference to another proof that he judged to be explanatory and which met his criterion. As discussed above, Resnik and Kushner (1987) critiqued Steiner's criterion by presenting proofs that either met his criterion without being explanatory, or which were explanatory but did not meet his criterion. Hafner and Mancosu (2005) adopted a similar approach, but this time appealed to the intuitive judgement of the author of the proof they presented. All these contributions involved presenting exemplars of explanatory proofs (or non-explanatory proofs) and analysing their properties (Inglis and Aberdein 2016). For these arguments to work, their proponents must make two assumptions. First, that their readers are easily able to intuitively decide whether or not a given proof is explanatory; and second, that readers' judgements of explanatoriness are likely to coincide. This second assumption is questionable. Resnik and Kushner (1987) themselves argued that "our intuitions concerning candidate explanatory proofs are weak, sparse and ill-defined" (p. 153). More concretely, Inglis and Aberdein (2016) presented empirical evidence that mathematicians regularly disagree about the properties of mathematical proofs. But our focus here is on the first assumption.

If individual mathematicians (and philosophers of mathematics) are intuitively able to decide whether or not a proof is explanatory, and if accounts of mathematical explanation can be evaluated by assessing their consistency with these intuitive judgements, it would be valuable for a notion of mathematical explanation to account for how these intuitive judgements are made. Consider the debate between Steiner and Feferman described by Resnik and Kushner: "Feferman proposed an intuitively explanatory proof to Steiner with the challenge to find the characterizing property used. To his credit, Steiner rose to the challenge, deformed the proof and arrived at another theorem." (pp. 146–147). But if Feferman, an eminent mathematician, could not immediately see the characterising property, clearly his intuitive judgement about the explanatoriness of this particular proof was not based on whether or not there was a characterising property. So what was it based upon? It would be desirable for an account of mathematical explanation to be able to answer this question. One major advantage of the functional account described in the next section is that it provides a natural account of how intuitive judgements of explanatoriness are reached.

A second observation that can be made about Steiner's (1978) and Lange's (2014) accounts, but perhaps not Kitcher's (1981), is that they fail to answer a challenge first posed by Michael Friedman in the context of scientific explanation. Friedman (1974) argued that any coherent account of scientific (or, presumably, mathematical) explanation needed to show how an explanation generates understanding: "I don't see how the philosopher of science can afford to ignore such concepts as 'understanding' and 'intelligibility' when giving a theory of the explanation relation." (p. 8).

In mathematical contexts Friedman's (1974) demand appears to be controversial. Delarivière, Frans and Van Kerkhove (2017) drew a distinction between what they

called *ontic* and *epistemic* accounts of mathematical explanation. Ontic accounts are those which do not concern themselves with Friedman's challenge: such accounts would be content with labelling a proof as explanatory even if it never generated any understanding for any agent. How could this be? Ontic accounts are willing to contemplate explanations that are not understandable by others: they are interested in saying that X explains Y when X is the objective reason why Y is the case, regardless of whether anyone has or could understand the relationship. Delarivière et al. emphasized that ontic approaches do not necessarily deny the relation between mathematical explanations and understanding, they simply do not pose understanding as a defining characteristic of mathematical explanations. In contrast, epistemic accounts have the advantage of directly addressing Friedman's challenge. Such accounts describe a proof as having explanatory value if and only if it grants some agent an understanding of why the associated theorem is true. Delarivière et al. pointed out that adopting an epistemic account leads to a contextual understanding of explanation: what is an explanation for one mathematician may not be for another. Delarivière et al.'s contextual epistemic account of mathematical explanation has several important similarities with our use of Wilkenfeld's (2014) functional account, and we discuss the similarities and differences of the two approaches later in the paper.

3 Wilkenfeld's functional explanation

Whereas Friedman (1974) demanded that philosophers show how explanations, suitably defined, generate understanding, Wilkenfeld's (2014) functional account does the reverse, by defining explanations to be those things that generate understanding. He suggested that such an account has only recently become tenable, as it is only in recent years that philosophical accounts of understanding have become sufficiently sophisticated. Our primary goal in this section is to outline the main aspects of Wilkenfeld's account; although in our view his proposal is very promising, we do not have the space here to defend it in a manner that would convince a sceptic (such readers should consult Wilkenfeld 2014). Instead, our aim is to explore the consequences of a functional account in the context of mathematical explanation.

Wilkenfeld (2014) offered this definition of functional explaining (FE):

> Explainer B *explains* (engages in an explaining-act with respect to) proposition p to audience A if and only if B produces a representation R that, when properly internalized, causes improved understanding of p in A (where the content of that understanding is similar to R), and B intends to explain or cause understanding via this production (p. 3371).

Several points are worth making about FE. First, the definition is neutral with respect to which account of understanding one adopts. There are several plausible options, one offered by Wilkenfeld himself (understanding as 'representation manipulability'). In the next section we adopt Kelp's (2016) approach and link it to work from educational psychology, but many of our arguments would work equally well under any cognitive theory of understanding [but perhaps not under ability-based accounts of the sort advanced by Avigad (2008)].

Second, FE sees explaining as an intentional act that takes place between an explainer and an audience. But it is perfectly possible that these might be the same person. For example, educational psychologists have established that instructing students to explain to themselves as they read a text (to generate 'self-explanations') leads to higher levels of understanding than if they are left to read the text without advice (e.g., Hodds et al. 2014; Rittle-Johnson and Loehr 2017).[2]

Third, FE requires that the representations produced by the explainer are "properly internalised" by the audience, a constraint designed to insist that the increased understanding is reached in the normal manner, involving some cognitive effort. Wilkenfeld pointed out that this restriction is necessary to prevent strange non-explanatory acts being counted as explanations (for instance, if a conjuror's spell created the necessary understanding in their audience through some supernatural route, one would not want to call this an explanation).

Fourth, FE does not directly involve any notion of veridicality. However, this criterion can be devolved to the account of understanding. If one favours a factive account of understanding—one that insists that understanding must involve knowledge—then explanations too must be veridical to at least some extent.

Fifth, FE provides a natural account for how intuitive judgements about the explanatoriness of mathematical proofs can be made. Individuals have a degree of insight into the extent to which they understand a given phenomenon (e.g., Schraw and Moshman 1995). While this kind of metacognitive ability is far from perfect (understanding when metacognitive failure occurs is an active area of educational research, e.g., Goos 2002; Schoenfeld 1985), it is likely that if reading a proof has a substantial effect on an individual's understanding, there is a good chance that they will be aware of it. Further, even when suffering from metacognitive failure (i.e., where a reader believes they have understood to a greater degree than they actually have), we would expect a reader to have a perception of their level of understanding, and therefore be able to report on the explanatoriness or otherwise of the proof they have read, even if this report was not completely accurate. To take a concrete example, FE would say that when Steiner (1978) found Gauss's proof concerning the sum of the first n numbers to be explanatory, he metacognitively compared his level of understanding before and after having read the proof. He did not search for a characterising property. On this account, and assuming Steiner was not suffering from metacognitive failure, his increased understanding was not just a useful heuristic for diagnosing that Gauss's proof was explanatory, it was constitutive of its explanatoriness.

Finally, 'understanding' in FE is not intended to refer to all types of understanding. Explanations must generate *objectual* understanding and not merely *propositional* understanding. This distinction can be illustrated by considering the difference between the assertions "John understands that the train goes through Derby" (propositional) and "Sally understands marine biology" (objectual). Propositional and objectual understanding are distinct because the latter admits degrees whereas the former does not. While it is meaningful to say that one person understands marine biology

[2] The self-explanation effect provides a further reason to favour a simple and intuitive account of mathematical explanation. It is hard to see how a simple instruction to generate explanations can improve students' understanding if none of us, including the students themselves, really know what mathematical explanations are.

to a greater extent than another, it is hard to see how someone could understand that a train goes through Derby more than anyone else (Baumberger 2014). Some epistemologists further distinguish between objectual understanding and understanding-why (Kvanvig 2003; Pritchard 2010), where the latter refers to a more localised type of understanding (compare "Cally understands furniture design" to "Beth understands why the chair leg snapped"). This global/local distinction is reflected in Pritchard's (2010) use of the terms *holistic* and *atomistic* understanding for objectual understanding and understanding-why respectively.[3] We concur with Grimm's (2016) assessment that the distinction between holistic and atomistic understanding (or objectual understanding and understanding-why/explanatory understanding) is a matter of degree rather than of kind,[4] and here we use 'objectual understanding' to refer to both.[5]

Clearly, for FE to avoid circularity we need to adopt an account of objectual understanding that does not itself appeal to explanation. Various accounts have been offered. Some, such as Avigad (2008), favour characterising an individual's understanding in terms of their abilities. Avigad argues that if an individual understands a proof this means that they have the ability to supply missing inferences, draw appropriate analogies, prove related theorems and so on. Here we focus on a different approach, which characterises understanding in cognitive terms, either as a special type of relationship between knowledge (e.g., Grimm 2006) or as a special type of relationship between beliefs (Kvanvig 2003).[6]

The idea that objectual understanding concerns the relationships between information (regardless of whether that information is knowledge or belief) is critical to cognitive accounts: Kvanvig (2003) wrote that "the grasping of relations between items of information is central to the nature of understanding" (p. 197). There are several plausible ways of fleshing out this observation into a full account of objectual understanding. Here we very briefly introduce Kelp's (2016) knowledge-based approach and highlight its similarity with some approaches adopted by educational psychologists. However, we emphasise our belief that the arguments we advance in the remainder of the paper will work equally well under any reasonable cognitive account.

Kelp (2016) defines *maximal understanding*[7] of some phenomenon P to be "fully comprehensive and maximally well-connected knowledge" about P (p. 252); this allows him to define a person's degree of understanding by their distance from maximal understanding. The closer one is to having maximal understanding, the better one understands. Binary judgements about whether or not someone understands can be made by assessing a person's distance from maximal understanding against some con-

[3] The emphasis on 'why' questions has led some to refer to understanding-why as *explanatory* understanding (e.g., Baumberger 2014).

[4] Grimm's (2016) point is that both understanding marine biology and understanding why the chair broke involve dependency relations, it's just that the former involves a great many more.

[5] Wilkenfeld (2014) himself preferred the term 'a cognitive sense of understanding' over what we are referring to as 'objectual understanding'. He "roughly" characterised this as "understanding-why or understanding-how, in opposition to merely understanding-that" (p. 3373).

[6] One issue of dispute between those who characterise understanding as knowledge and those who characterise it as belief concerns whether understanding is vulnerable to Gettier cases. This distinction is not so important for our purposes in this paper, and we consider both approaches to be cognitive accounts.

[7] For the remainder of the paper, when we refer to 'understanding' we mean objectual understanding.

textual threshold. For example, we might make different judgements about whether a person understands gravity depending on whether they are a schoolchild or a research-active physicist; in Kelp's terms we would tolerate a larger distance from maximal understanding in the child's case than in the physicist's case.

Kelp's (2016) account fits naturally with the ways in which educational psychologists like to think about understanding. Consider schema theory (e.g., Chi et al. 1982; Piaget 1926). A schema is a cognitive structure that permits us to treat multiple elements of information as if it were a single element. For example, we might have a schema for objects such as flowers that allow us to recognise a flower immediately despite the complex variety of components that make up the flower (its petals, stamen, stalk, colour, etc.) and despite the wide variety of different flowers. It is well established that experts have better organised and more comprehensive schemas, and that this increases the fluency and accuracy of their performance. For instance, chess grandmasters have schemas that allow them to readily identify a very large number of board configurations. It is the comprehensive and well-organised nature of the grandmasters' schemas that distinguish them from novice players (e.g., Chase and Simon 1973), and analogous observations can be made of experts in other domains, such as problem solving (e.g., Larkin et al. 1980). To interpret Kelp's account of understanding in terms of schema theory, what distinguishes expert-like understanding from novice-like understanding is the distance between their cognitive schemas and the cognitive schemas that would be required for maximal understanding. It is schemas that cognitively instantiate the relationships an individual has between their knowledge (or, if you prefer, their beliefs), and therefore it is schemas that are constitutive of understanding (cf. Kvanvig 2003).

Two clarifications are in order. First, because Kelp's (2016) account focuses on knowledge it is factive, at least in a moderate sense. In Kelp's sense, person P can only maximally understand X if X is true. The account can be said to be moderately factive as in some contexts it is possible that certain falsehoods believed by P about X would not be sufficient to create enough distance from maximal understanding to assert that P did not understand X. In other words, while a few peripheral falsehoods would challenge P's claim to understand, they might not undermine it completely.

Second, Kelp's (2016) account goes well beyond a "sense of understanding" (Trout 2002) or an "aha feeling" (Delarivière et al. 2017). While these phenomenological experiences may be correlated with possessing understanding in Kelp's sense, they do not constitute understanding itself. Indeed, as Trout (2002) pointed out, there are many reasons to suppose that our "sense of understanding" is not faultless: metacognitive failure is a well-documented phenomenon in mathematics education (e.g., Goos 2002; Schoenfeld 1985). In Kelp's terms, it is possible (albeit probably quite rare) for P to understand X without realising it, and possible (but again, probably quite rare) for P to think that they understand X without actually understanding it (or at least not to the same extent).

As we have seen, the involvement of philosophical theories of understanding are required to successfully account for mathematical explanations if we assume FE. However, the educational psychology literature may well also be of assistance. Whereas epistemologists such as Kelp (2016) have focused on what understanding actually is, educational psychologists are more concerned with the mechanisms by which peo-

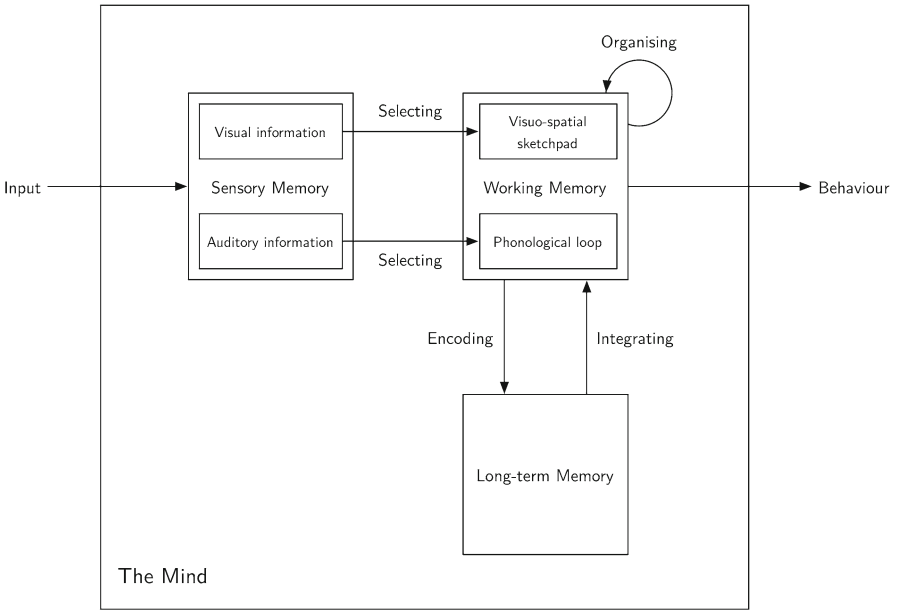

Fig. 1 A modal model of the mind

ple come to understand, i.e. the mechanisms by which they create and enrich their schemas. If explanations are those things that cause understanding, then psychological theories ought to offer some insight into what sorts of properties good explanations are likely to have. In the next section we very briefly outline a 'modal model' of the mind, adapted from work in educational psychology. We then use this model to derive what properties we would expect successful mathematical explanations to have.

4 Human cognitive architecture: a modal model

Willingham (2017) recently argued that schoolteachers need to understand the basic mechanisms by which humans learn, but that it would be counterproductive to expose them to all contemporary theoretical debates taking place within the psychological research literature. Instead, he proposed that teachers should be offered a *modal model* of human cognitive architecture, a limited model which essentially all researchers would agree with (here 'modal' is intended to invoke the mathematical notion of 'most commonly occurring'). Modal models necessarily omit a great deal of complexity and involve considerable simplification, but as a result are relatively easy to understand and are uncontroversial among relevant researchers. Here we present a modal model of human cognitive architecture, based on Atkinson and Shiffin's (1968) three-component model (Fiorella and Mayer 2015; Mayer and Moreno 2003). The model is shown in Fig. 1.

According to most accounts of human cognitive architecture, humans have three memory systems: sensory memory, working memory and long-term memory. Sensory

memory is the least relevant for our purposes. It allows sensory information to be stored long enough for relevant components to be selected and transferred to working memory. To allow this, impressions of sensory information can be retained in sensory memory for short periods after the stimulus itself has ceased.

Working memory is the site of all conscious thought: whenever we are aware of a piece of information, it is being processed in working memory. Working memory has two important limitations. First, it can only hold information for short periods. Peterson and Peterson (1959) found that information stored in working memory is lost within 30 s unless rehearsed. Second, working memory has extreme capacity limits. Miller (1956) famously claimed that only around seven ("plus or minus two") elements can be processed in working memory at a time, although more recent work has suggested this figure might be even lower depending on the nature of the processing (e.g., Cowan 2001).

Information can enter working memory from two different sources. First, one might consciously process a stimulus selected from sensory memory. To do this requires attention: in particular one must select which part(s) of the representation of the complex environment stored in sensory memory is worth devoting one's limited working memory capacity to. Second, one may integrate schemas stored in the third memory system, long-term memory.

Long-term memory is a critical component of human cognition: it is where schemas formed during previous working memory processing are stored. As noted above, the reason expert chess players outperform novices is because of the superior quality of their schemas. These are stored in their long-term memory and then retrieved and integrated into working memory when required (Chase and Simon 1973). In contrast to working memory, long-term memory appears to have no practical capacity limits, and can be used to bypass the temporal limits of working memory. If one has stored a schema in long-term memory, it can be repeatedly re-integrated into working memory, rendering the 30 s limit irrelevant. Long-term memory also provides a mechanism for circumventing working memory's capacity limits. If knowledge is structured into a coherent schema and stored in long-term memory, then the whole schema can be integrated into working memory and use less capacity than if the same knowledge had not been coherently organised, say by it having been organised into disjoint schemas. For example, imagine someone informed you their telephone number was 01123 581321. It would be a challenge to hold this number in working memory while conducting some secondary task (perhaps searching for a public phone box). However, if they told you that their phone number was the first eleven digits of the Fibonacci sequence, the same task would be much easier. In the latter case the same knowledge would have been organised into a coherent schema that would reduce the level of working memory resources required.

A final important component of our modal model of human cognitive architecture is known as the dual channel assumption (e.g., Mayer and Moreno 2003). This states that there are two separate channels in sensory and working memory, one for auditory/verbal stimuli and one for visual stimuli. This assumption is central to both Clark and Paivio's (1991) dual-coding theory, and Baddeley and Hitch's (1974) model of working memory. This latter theory proposes that working memory has two 'slave systems': verbal and auditory processing takes place in the so-called phonological

loop, whereas visual processing takes place within the visuospatial sketchpad (e.g., Baddeley 1992). Both these systems have limited capacity, meaning that if processing can be split between them total working memory capacity is, in effect, increased. In other words, information processing is more effective if the to-be-processed information is split between visual and verbal modalities (e.g., Mayer and Moreno 2003; Sweller et al. 2011).

We are now in a position to ask how understanding, in the sense of Kelp (2016) and schema theory, can be generated. From a cognitive perspective, one can be said to have understood something when a sufficiently well-organised schema (i.e., one that is sufficiently similar to the schema corresponding to Kelp's maximal understanding) has been encoded into long-term memory. Accomplishing this would typically involve selecting relevant aspects of the environment from a representation held in sensory memory, processing it in working memory, integrating relevant existing schemas from long-term memory, and re-organising this knowledge into a new schema that can subsequently be encoded into long-term memory.

5 The properties of good mathematical explanations

Using this account of the cognitive mechanisms involved in generating understanding, and since FE defines those things that generate understanding (in the right sort of way) to be explanations, we can consider what properties successful mathematical explanations are likely to have. Our account suggests that the archetypal explanatory proof would have at least three properties. First, it would have features that make it easy, or at least as easy as possible, to select the information from sensory memory into working memory that is necessary for a successful processing stage. In other words, such a proof would effectively direct the reader's attention to its conceptually important sections. Second, it would have features that make it easier to coordinate the new knowledge contained in the proof with existing schemas retrieved from long-term memory, and therefore to reorganise the new and existing information into coherent new schemas. Finally, it would be likely to split the working memory load it gives to its readers between their visual and verbal/auditory channels so that the chances of their working memory capacity being exceeded during the schema-organisation process is minimised.

The characteristics that Steiner (1978), Kitcher (1981) and Lange (2014) offered as definitions of explanatoriness would all seem to aid readers with this cognitive process of generating understanding. We explore each in turn.

Recall that Steiner argued that explanatory proofs make reference to some characterising property of an entity mentioned in the theorem. In our terms, making reference to a characterising property is a mechanism that will assist readers to integrate relevant schemas stored in long-term memory into working memory, where they can be organised into a new schema that incorporates the additional information from the proof. Retrieving information from long-term memory, and its integration in working memory, is facilitated in the presence of a cue that references that information (e.g., Tulving and Pearlstone 1966; Unsworth et al. 2013). The fact that the theorem and proof both refer to the same (characterising) mathematical property, means that the characterising

property can function both as a retrieval cue that will assist the reader to select the relevant schemas from long-term memory, and also as a signal for how they should be integrated into working memory (Mautone and Mayer 2001). This therefore improves the chances of a successful reorganisation of knowledge. In other words, integrating new information into an existing schema is facilitated by the new information directly referring to a component of the existing schema (Mayer and Moreno 2003). These considerations point to a clarification of Steiner's suggestion: if the reader has no existing schema that involves the characterising property, then it is more difficult for a proof to effectively generate understanding. In other words, if a proof's characterising property concerns some piece of mathematics that the reader has yet to encounter (or which they have encountered but have failed to successfully understand, in the sense of encoding a well-organised schema about it into their long-term memory), then it is harder for the proof to be explanatory for that reader.

In sum, we can think about a characterising property as being both a mathematical property in Steiner's sense, but also as a cognitive property that helps to link representations in sensory memory, working memory and long-term memory. Since creating new understanding from external stimuli requires these three memory systems to work together, if a proof has such a property the generation of understanding is facilitated and consequently, in view of FE, the proof is more likely to be explanatory. We emphasise that if a proof does not have a characterising property in Steiner's sense, this does not mean that it *cannot* generate understanding and be explanatory—there are other mechanisms which can facilitate understanding, some discussed below—but only that, all things being equal, it is harder for it to do so.

A similar analysis applies in the case of Lange's (2014) account of mathematical explanation. Recall that he suggested that explanatory proofs have some sort of salience, often symmetry, that links the problem set-up, the statement of the result, and the proof itself. In precisely the same way that a characterising property can be seen as a cognitive feature as well as a mathematical feature, so can Lange's notion of saliency. In particular, a proof that exhibits this kind of saliency will facilitate selecting the information from sensory memory (the salient information) necessary for a successful knowledge reorganisation (cf. the 'signalling effect', Mautone and Mayer 2001), and will facilitate integrating schemas from long-term memory relevant to this knowledge into working memory (Mautone and Mayer 2001; Mayer and Moreno 2003). In other words, things that are mathematically salient will also, for the right reader at least, be cognitively salient, and therefore help generate understanding for that reader.

What of Kitcher's (1981) unification account? Recall that he suggested that explanations are those things which unify phenomena by showing that they are all derivable by stringent arguments of the same form. In our terms, the process of (cognitively) unifying two different phenomena involves integrating two or more schemas from long-term memory into working memory, reorganising them into a single schema, and encoding it back into long-term memory. In general, the more unified a schema the closer it is to the kind of maximally well-organised and dense schema that would be associated with maximal understanding in the sense of Kelp (2016). To take the earlier example, prior to understanding Darwin's theory of evolution, anyone who had knowledge of Galápagos finches and dinosaur fossils will likely have encoded this knowledge using disjoint schemas in their long-term memory. Darwin's theory allows

for these schemas to be integrated into a more coherent single schema, and therefore reduces the distance to maximal understanding. Although Kitcher conceptualised the unification as taking place at the mathematical level, it also takes place at the cognitive level and, according to the functional account, it is this which gives a unifying proof its explanatory status.

Finally, we remark upon the common observation that visual proofs are often seen as being explanatory (Hanna 2000; Steiner 1978). Such a finding is precisely what our cognitive functional account would predict. Visual proofs usually (but not always, see below) split the processing load required to understand them between the visual and verbal/auditory channels in sensory and working memory. Because of working memory capacity limits, the successful re-organisation of schemas is facilitated if information is processed using both channels (e.g., Clark and Paivio 1991; Mayer and Sims 1994). In effect, this way of presenting proofs increases readers' working memory capacities (e.g., Mousavi et al. 1995). Of course, this argument would not apply to *purely* visual proofs—those that contain no text of any kind. Such proofs do exist (e.g., Nelsen 1993) but, in contrast to Steiner's, our account suggests they would be less explanatory than an equivalent proof that also included information conveyed by text.[8]

In sum, all three earlier accounts of mathematical explanation discussed here—those proposed by Steiner (1978), Kitcher (1981), and Lange (2014)—seem to follow from a functional explanation account coupled with a modern understanding of human cognitive architecture. In other words we have suggested that, while Steiner, Kitcher and Lange were all wrong to suggest that the criteria they proposed successfully *defined* the class of mathematical explanations, they were all correct in the sense that their criteria are all *prototypical* of mathematical explanations. Good explanatory proofs will often involve characterising properties, they will often unify different phenomena, and they will often involve some kind of salience. However this is not because these features are necessary and sufficient criteria for explanatoriness, it is because these features facilitate the generation of understanding.

6 Example applications

How does our account fare when applied to specific examples of more explanatory and less explanatory proofs? Consider the following two proofs, both by contradiction. The first, according to Colyvan (2012), is not explanatory whereas the second is.

Theorem 1. There are infinitely many primes.
Proof 1. Assume that there is a largest prime, p. Consider the number one greater than the product of all the primes: $n = 2 \times 3 \times 5 \times \ldots \times p + 1$. Either n is a product of primes or it is a prime larger than p. The latter would contradict our premise, so n must be a product of primes. But if n is a product of primes and has no

[8] Some tangentially relevant empirical evidence pertains to this question. Inglis and Mejía-Ramos (2009) found that if a purely visual argument was accompanied by a passage of descriptive text (that described the image, but which had no further content), then both research-active mathematicians and undergraduate mathematics students perceived it to be more persuasive than if the descriptive text was omitted.

prime factors greater than p, then one of its factors, q, must be in the sequence 2, 3, 5, …, p, and therefore divides the product $2 \times 3 \times 5 \times \ldots \times p$. However, since it is a factor of n it also divides n. But a number which divides two numbers also divides their difference, so q must also divide $n - (2 \times 3 \times 5 \times \ldots \times p) = (2 \times 3 \times 5 \times \ldots \times p + 1) - (2 \times 3 \times 5 \times \ldots \times p) = 1$. However, no prime divides 1 so q is not in the sequence 2, 3, 5, …, p. It follows that if n is composite, it has at least one factor greater than p. This is a contradiction. Therefore there is no largest prime number; there are infinitely many primes. □

The functional account of explanation provides a clear reason why most proofs by contradiction are unlikely to generate much understanding, and therefore are unlikely to be explanatory. The bulk of a proof by contradiction involves reasoning about logically inconsistent mathematical worlds. Since we very rarely encounter such worlds—we rarely spend time reasoning about mathematical concepts that are logically inconsistent—we do not have schemas about these concepts stored in our long-term memory. Neither can we easily combine our existing schemas to create coherent new schemas about such concepts: by design, such objects are inconsistent, and so our understanding of them will necessarily involve incoherent schemas. For instance, few people who read Proof 1 will have a well-organised schema about the largest prime number. Because of this, when reading contradiction proofs there are relatively few existing schemas upon which we can draw to assist us, and into which we can integrate our new knowledge. Furthermore, it is clearly impossible to create a coherent schema about the largest prime number. In other words, a proof by contradiction creates a situation for us where we cannot easily link input from sensory memory to schemas encoded in long-term memory or construct new coherent schemas. Moreover, in most such contexts we probably do not want to do this (we typically do not wish to encode schemas about a mathematical world where there are finitely many primes). Given these factors, we should not be surprised that many proofs by contradiction are not explanatory. As a general rule, they are unlikely to generate understanding.

However, Colyvan (2012) argued that this is not universally true, and gave the following example of an explanatory proof by contradiction.

Theorem 2. 2 is the only even prime.
Proof 2. Assume that there exists another even prime, $p > 2$. Since p is even it can be divided by 2 so it can be written $p = 2q$ for some integer $q > 1$. But this means that p is composite, contradicting the original assumption that p is prime. □

Our account provides a justification for Colyvan's suggestion. This proof by contradiction seems different to the first. It simply draws on two existing schemas—those that encode the definition of prime and the definition of evenness—and links them in a straightforward way. A reader who has understood this proof needs to have reorganised and re-encoded their existing schemas for even numbers and prime numbers so that they are connected. If they have successfully done this, it seems reasonable to suppose that they will have reduced their distance to a maximal understanding of the prime numbers.

As another example, consider these two proofs of the same theorem, offered by Dreyfus and Eisenberg (1986, p. 3), and discussed by Lange (2016, p. 267). The first, according to Lange, is non-explanatory.

> *Theorem 3*. In a list of integers from 1 to 99,999 the digit 7 appears 50,000 times.
> *Proof 3.1*. The digit 7 appears once between 1 and 10, once between 11 and 20, in fact once in every 'regular' 10-plet of numbers; here 'regular' means there are no digits 7 in the tens or higher places. Between 61 and 70 there are two digits 7 s; between 71 and 80 there are 10; collecting all of these, one concludes that the digit 7 appears 20 times between 1 and 100, and thus 20 times in every 'regular' 100-plet. In the 100-plet from 601 to 700 there is an additional 7, i.e. the digit 7 occurs 21 times, and in the following 100-plet there are 99 additional ones, yielding altogether 300 digits between 1 and 1000. Proceeding in a similar way, one finds that the digit 7 appears 4000 times between 1 and 10,000, and it appears 50,000 times between 1 and 100,000. As it does not appear in the number 100,000 the digit appears 50,000 times when listing all the integers from 1 through 99,999. □

Consider this alternative proof, described by Lange as explanatory.

> *Proof 3.2*. Include 0 among the numbers under consideration—this will not change the number of times the digit 7 appears. Suppose all numbers from 0 to 99,999 are written down with five digits each, e.g. 306 is written 00,306. All possible five digit combinations are now written down, once each. Because in this set of all possible combinations every digit will take every position equally often, every digit must, overall, occur the same number of times. Since there are 100,000 numbers with five digits each, that is 500,000 digits, each of the 10 digits appears 50,000 times. In particular, this is true of 7. □

Can we account for why Proof 3.1 is seen as non-explanatory whereas Proof 3.2 is seen as explanatory? The answer to the first part of this question seems clear. Proof 3.1 is a simple counting argument. For the vast majority of readers it will activate only existing schemas that are exceptionally well understood (those concerned with natural numbers represented in base 10). Since most readers who encounter this proof will already have extremely well-developed schemas concerning these matters, there is little room for the proof to develop new understanding. Instead it is a simple matter of applying existing understanding to a particular context.

Proof 3.2 is more interesting. It involves a shift in the way natural numbers are represented. Although in the number 306 the digit 3 represents 3 hundreds, the proof highlights that in some contexts this information can be productively ignored. Instead numbers such as 306 can be considered as a string of characters—00306—devoid of their numerical magnitude meaning (a directly analogous proof could be written about five-character long strings of letters). Having established this unusual way of thinking about numbers, the proof draws on our existing knowledge of combinatoric processes and establishes the result. Understanding is generated in at least two ways by this proof. First, the reader's schemas of natural numbers are likely to be enriched with the new knowledge that it is possible and sometimes productive to ignore numerical

magnitude information: base 10 numerical representations (or representations in other bases for that matter) can be productively thought of as strings of characters rather than numbers with magnitude meanings. A maximal understanding of natural numbers would likely include this knowledge. Second, by viewing numbers in this way, methods more commonly associated with combinatorics can be applied and, depending on the reader's existing knowledge, their schemas associated with natural numbers will likely become connected to an array of problem-solving techniques previously stored in disjoint schemas. Again, a maximal understanding of prime numbers would likely include this kind of array of problem solving techniques.

One objection which could be levelled against our account of the explanatoriness of these two proofs is as follows.[9] On Kelp's (2016) account, understanding is a type of knowledge. So any proof which increases a reader's knowledge to any extent will also increase their understanding, and therefore qualify as explanatory under FE. Since all proofs do this, even if only to the extent that their readers now know that the theorem is true, then all proofs must be explanatory. Such a conclusion would clearly be inconsistent with mathematical practice.

There are two ways to respond to this challenge. First, FE is concerned with objectual understanding, not propositional understanding. The knowledge that a particular theorem holds constitutes propositional not objectual understanding. This can be seen by the observation that this type of understanding is not gradated: a given reader either understands that the theorem is true or does not.

Second, even if one asserted that all proofs generate objectual understanding to some degree, it would not follow that all proofs must be classified as explanatory under FE. Because objectual understanding is continuous, so is explanatoriness: some proofs are more explanatory than others. Nevertheless, the continuous nature of a proof's explanatoriness can be used to form discrete judgements of the form "this proof is explanatory, that proof is not". A natural way to do this is suggested by Kelp's (2016) observation that one can use contextual thresholds to coherently assess whether someone understands gravity, even though this understanding comes in degrees. Much as Kelp (2016) suggests that contextual thresholds can be used to determine the level of understanding required to assert that a schoolchild or a physicist understands gravity, contextual thresholds concerning the extent to which a proof *increases* objectual understanding can be used to assert that a given proof is explanatory or not. So if Proof 3.1 did develop a reader's objectual understanding, but only by some trivial amount, then it is unlikely that it would exceed the increased-understanding threshold for explanatoriness. In such a scenario, we could coherently describe the proof as being non-explanatory.

[9] We are grateful to an anonymous reviewer for raising this important objection.

7 Discussion

7.1 Are explanations contextual?

What are the consequences of taking Wilkenfeld's (2014) functional explanation approach seriously in the context of mathematics? Whereas earlier accounts of explanatory proofs were all concerned with the properties of the proofs themselves—they were *ontic*, in the sense of Delarivière et al. (2017)—our proposal locates explanatoriness at the conjunction of the reader and the proof. A proof is explanatory for a given reader if it brings about understanding for that reader. Does this mean our account is context dependent? Are proofs explanatory for some and non-explanatory for others?

Our functional account provides a nuanced answer to this question. While there are some aspects of human cognitive architecture outlined earlier that do vary between individuals, there are other aspects that do not. For instance, it is clearly the case that there are individual differences in the content of long-term memory: we have all had different experiences and have consequently encoded different memories. Similarly, there are individual differences in working memory capacity: some people are capable of holding and processing more information than others (e.g., Baddeley 1992). However, the overall structure of human cognitive architecture is universal (perhaps with the exception of some individuals with serious neurological conditions and the like). For example, we all have two channels in sensory and working memory, and we all have (low) working memory capacity limits.

These observations suggest that we should expect both individual differences in what proofs mathematicians find explanatory, but also a degree of consensus. Imagine that Proof P has a Steiner-like characterising property X. Supposing Reader A has never previously encountered X, then P is highly unlikely to be explanatory for A. In our terms, in A's context, X is a mathematical characterising property but not a cognitive characterising property. For a different reader, B, who has a well-organised schema about X stored in their long-term memory, then P may very well generate understanding in the manner described above (for B, X is both mathematically and cognitively characterising). For a third reader, C, who already has extremely well-organised and dense schemas concerning all the mathematical concepts involved in P, then it is hard to see how P could ever be personally explanatory: C already has all the understanding that P has to offer.

However, it would be a mistake to deduce from this that our expert C would never describe a basic proof as being explanatory. Consider Colyvan's (2012) proof by contradiction, discussed above, of the claim that 2 is the only even prime. This proof was described by Colyvan as being explanatory, but we can with some confidence assume that Colyvan has a very detailed understanding of even numbers and prime numbers. The proof presumably therefore did not generate new understanding for Colyvan himself, but he was nevertheless happy to call it explanatory. How can we explain this? The literature on instructional explanations (e.g., Leinhardt 2001; Stein and Kucan 2010; Treagust and Harrison 1999) provides us with an answer. In many of the contexts where we make judgements about the (non-)explanatoriness of proofs,

and particularly in educational contexts, we are imagining a hypothetical audience and judging the extent to which the proof will generate understanding for them. Indeed, making such judgements constitutes a large component of what instructors do when preparing materials for teaching, and one characteristic of a successful teacher is that they choose explanations that will facilitate learning rather than interfere with learning (e.g., Eisenhart et al. 1993). When choosing which of several elementary proofs of a basic theorem to present to an introductory class, we choose the proof which we judge the class will find most explanatory, that which is most likely to help develop their understanding (Schoenfeld 2010).[10] So while Colyvan himself probably did not gain additional understanding from his proof, by labelling it explanatory he was, we suggest, implying that some appropriate audience would gain significant additional understanding from it.

What of visual images? Consider what would happen if P could be deformed into a new proof, P', by including some relevant visual images. Our account would suggest that it would become easier to process for all three of A, B and C. Although this would not make much difference to C (who already understands everything about the proof), P' will be *more* explanatory than P for both A and B, as they will effectively have more working memory capacity to use when reading it (of course this assumes that P is sufficiently complex to tax the working memory capacity of all our readers). Further, if C has a novice audience in mind when reading P and P', he or she will probably also be willing to say that P' is more explanatory than P.

These kinds of analyses suggest numerous methods by which our proposal could be tested empirically. For instance, we could experimentally manipulate mathematicians' working memory capacities as they read proofs. This is easily achieved using a secondary task paradigm. One approach would be to ask mathematicians to generate random numbers as they read one proof (e.g., Baddeley et al. 1998) while asking them to read another without this secondary task. If the proofs were counterbalanced between participants so as to remove effects associated with their respective difficulties, then our clear prediction would be that mathematicians would rate the proof they read while generating random numbers as less explanatory. This load task taxes working memory, making it less likely that the participants would be able to effectively organise their knowledge into new schemas.

7.2 Ontic versus epistemic approaches to mathematical explanation

Recall that Delarivière et al. (2017) distinguished between two broad categories of accounts of mathematical explanation: ontic and epistemic. Clearly the functional account proposed by Wilkenfeld (2014) and adopted here is an epistemic account, as it takes the generation of understanding to be the defining characteristic of an explanation. Different challenges emerge depending on whether one adopts an ontic or an epistemic account.

[10] That said, we recognise Schoenfeld's (2010) point that a teacher may have instructional goals beyond simply generating understanding of a piece of mathematics (for instance, they may prioritise generating an understanding of mathematical habits of mind rather than of particular mathematical concepts), and that this will influence their choice of instructional explanation.

Those who favour an ontic approach need to have a ready response to Friedman's (1974) challenge: why and how do explanations generate understanding? One response is to reject the challenge and simply assert that explanations do not generate any more understanding than non-explanations. But this approach seems difficult to reconcile with mathematical (or educational) practice. Alternatively, an ontic theorist could attempt to construct an inverted version of the arguments we have presented here. For instance, one could plausibly say that a Steiner-style characterising property generates understanding for the cognitive reasons we have described, but that it is the characterising property that defines the proof's explanatoriness, not the fact that it generated understanding. For this approach to succeed the ontic theorist would need to justify why understanding generated via a characterising property is somehow qualitatively different—and special in some way—to the same understanding generated through some alternative route (Lange-style saliency for instance). In our view no ontic account of mathematical explanation has, to date, successfully given such an account.

The challenge for the epistemic theorist is the direct inverse of that for the ontic theorist. If understanding defines explanation, then how can we characterise the mathematical properties of explanatory proofs? Delarivière et al.'s (2017) response to this challenge was to argue for pluralism in mathematical explanation. They suggested that the various existing accounts of mathematical explanation are all valuable, but distinct, and that therefore "there is no single account of explanation" (p. 16).[11] If this statement is read on a mathematical level we agree: the functional account suggests that there is no single mathematical characteristic possessed by an explanatory proof that will allow it to be distinguished from a non-explanatory proof. However, read more broadly we disagree. The functional account has a single model: explanatory proofs generate understanding. In this paper we have tried to demonstrate why this single model leads to the variety of criteria proposed by Steiner (1978), Kitcher (1981) and Lange (2014). The answer is that, by drawing on modern models of cognitive architecture, we can see that these criteria are precisely those properties that a proof which generates understanding is likely to have. In that sense, the functional account unifies the various existing ontic accounts.

Acknowledgements We are grateful to Brendan Larvor, Fenner Tanswell, and three anonymous reviewers for helpful comments on earlier drafts of this paper. This work was first presented at the Enabling Mathematical Cultures workshop in Oxford (2017), and we thank the organisers and attendees.

References

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *Psychology of Learning and Motivation, 2,* 89–95.

[11] While not endorsing an epistemic account, Hafner and Mancosu (2005) also emphasized the heterogeneity of mathematical explanation arguing that "the variety of mathematical explanations cannot be easily reduced to a single model" (p. 222).

Avigad, J. (2008). Understanding proofs. In P. Mancosu (Ed.), *The philosophy of mathematical practice* (pp. 317–353). Oxford: Oxford University Press.
Baddeley, A. (1992). Working memory. *Science, 255,* 556–559.
Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47–89). New York: Academic Press.
Baddeley, A., Emslie, H., Kolodny, J., & Duncan, J. (1998). Random generation and the executive control of working memory. *Quarterly Journal of Experimental Psychology A, 51,* 819–852.
Baker, A. (2005). Are there genuine mathematical explanations of physical phenomena? *Mind, 114,* 223–238.
Baumberger, C. (2014). Types of understanding: Their nature and their relation to knowledge. *Conceptus, 40,* 67–88.
Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4,* 55–81.
Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. Sternberg (Ed.), *Advances in the psychology of human intelligence* (pp. 7–75). Hillsdale, NJ: Erlbaum.
Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review, 3,* 149–210.
Colyvan, M. (2012). *An introduction to the philosophy of mathematics*. Cambridge: Cambridge University Press.
Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24,* 87–114.
D'Alessandro, W. (in press) Mathematical explanation beyond explanatory proof. *British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axy009.
Delarivière, S., Frans, J., & Van Kerkhove, B. (2017). Mathematical explanation: A contextual approach. *Journal of Indian Council of Philosophical Research, 34,* 309–329.
Dreyfus, T., & Eisenberg, T. (1986). On the aesthetics of mathematical thought. *For the Learning of Mathematics, 6*(1), 2–10.
Eisenhart, M., Borko, H., Underhill, R., Brown, D., Jones, D., & Agard, P. (1993). Conceptual knowledge falls through the cracks: Complexities of learning to teach mathematics for understanding. *Journal for Research in Mathematics Education, 24,* 8–40.
Feferman, S. (1969). Systems of predicative analysis. In J. Hintikka (Ed.), *The philosophy of mathematics*. Oxford: Oxford University Press.
Fiorella, L., & Mayer, R. E. (2015). *Learning as a generative activity*. Cambridge: Cambridge University Press.
Friedman, M. (1974). Explanation and scientific understanding. *Journal of Philosophy, 71,* 5–19.
Goos, M. (2002). Understanding metacognitive failure. *Journal of Mathematical Behavior, 21,* 283–302.
Grimm, S. R. (2006). Is understanding a species of knowledge? *British Journal for the Philosophy of Science, 57,* 515–535.
Grimm, S. R. (2016). Understanding and transparency. In S. R. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining understanding: New perspectives from epistemology and philosophy of science* (pp. 251–271). London: Routledge.
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). Dordrecht: Springer.
Hafner, J., & Mancosu, P. (2008). Beyond unification. In P. Mancosu (Ed.), *The philosophy of mathematical practice* (pp. 151–178). Oxford: Oxford University Press.
Hanna, G. (2000). Proof, explanation and exploration: An overview. *Educational Studies in Mathematics, 44,* 5–23.
Hodds, M., Alcock, L., & Inglis, M. (2014). Self-explanation training improves proof comprehension. *Journal for Research in Mathematics Education, 45,* 62–101.
Inglis, M., & Aberdein, A. (2016). Diversity in proof appraisal. In B. Larvor (Ed.), *Mathematical cultures: The London meetings 2012–2014* (pp. 163–179). Basel: Birkhäuser Science.
Inglis, M., & Mejia-Ramos, J. P. (2009). On the persuasiveness of visual arguments in mathematics. *Foundations of Science*, *14*, 97–110.
Kelp, C. (2016). Towards a knowledge-based account of understanding. In S. R. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining understanding: New perspectives from epistemology and philosophy of science* (pp. 251–271). London: Routledge.
Kitcher, P. (1981). Explanatory unification. *Philosophy of Science, 48,* 507–531.

Kvanvig, J. L. (2003). *The value of knowledge and the pursuit of understanding*. Cambridge: Cambridge University Press.
Lange, M. (2014). Aspects of mathematical explanation: Symmetry, unity, and salience. *Philosophical Review, 123,* 485–531.
Lange, M. (2016). *Because without cause: Non-causal explanations in science and mathematics*. Oxford: Oxford University Press.
Larkin, J. H., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Models of competence in solving physics problems. *Cognitive Science, 4,* 317–348.
Leinhardt, G. (2001). Instructional explanations: A commonplace for teaching and location for contrast. In V. Richardson (Ed.), *Handbook of research on teaching* (pp. 333–357). Washington, DC: American Educational Research Association.
Lyon, A., & Colyvan, M. (2008). The explanatory power of phase spaces. *Philosophia Mathematica (III), 16,* 227–243.
Mautone, P. D., & Mayer, R. E. (2001). Signaling as a cognitive guide in multimedia learning. *Journal of Educational Psychology, 93,* 377–389.
Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38,* 43–52.
Mayer, R. E., & Sims, V. K. (1994). For whom is a picture worth a thousand words? Extensions of a dual-coding theory of multimedia learning. *Journal of Educational Psychology, 86,* 389–401.
Mejía-Ramos, J. P., Alcock, L., Lew, K., Rago, P., Sangwin, C., & Inglis, M. (2019). Using corpus linguistics to investigate mathematical explanation. In E. Fischer & M. Curtis (Eds.), *Methodological advances in experimental philosophy* (pp. 239–264). London: Bloomsbury.
Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97.
Mousavi, S. Y., Low, R., & Sweller, J. (1995). Reducing cognitive load by mixing auditory and visual presentation modes. *Journal of Educational Psychology, 87,* 319–334.
Nelsen, R. B. (1993). *Proofs without words: Exercises in visual thinking (No. 1)*. Washington, DC: MAA.
Peterson, L., & Peterson, M. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology, 58,* 193–198.
Piaget, J. (1926). *The language and thought of the child*. London: Kegan, Paul, Trench, Trubner, and Company.
Pritchard, D. (2010). Knowledge and understanding. In D. Pritchard, A. Miller, & A. Haddock (Eds.), *The nature and value of knowledge: Three investigations* (pp. 3–90). New York: Oxford University Press.
Resnik, M. D., & Kushner, D. (1987). Explanation, independence and realism in mathematics. *British Journal for the Philosophy of Science, 38,* 141–158.
Rittle-Johnson, B., & Loehr, A. M. (2017). Eliciting explanations: Constraints on when self-explanation aids learning. *Psychonomic Bulletin & Review, 24,* 1501–1510.
Salmon, W. C. (1971). *Statistical explanation and statistical relevance*. Pittsburgh, PA: University of Pittsburgh Press.
Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.
Schoenfeld, A. H. (1985). *Mathematical problem solving*. Orlando: Academic Press.
Schoenfeld, A. H. (2010). How and why do teachers explain things the way they do? In M. K. Stein & L. Kucan (Eds.), *Instructional explanations in the disciplines* (pp. 83–106). Boston, MA: Springer.
Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review, 7,* 351–371.
Stein, M. K., & Kucan, L. (Eds.). (2010). *Instructional explanations in the disciplines*. Boston, MA: Springer.
Steiner, M. (1978). Mathematical explanation. *Philosophical Studies, 34,* 135–151.
Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. New York: Springer.
Treagust, D., & Harrison, A. (1999). The genesis of effective scientific explanations for the classroom. In J. Loughran (Ed.), *Researching teaching: Methodologies and practices for understanding pedagogy* (pp. 28–43). London: Falmer.
Trout, J. D. (2002). Scientific explanation and the sense of understanding. *Philosophy of Science, 69,* 212–233.
Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior, 5,* 381–391.
Unsworth, N., Brewer, G. A., & Spillers, G. J. (2013). Working memory capacity and retrieval from long-term memory: The role of controlled search. *Memory & Cognition, 41,* 242–254.

Weber, E., & Frans, J. (2017). Is mathematics a domain for philosophers of explanation? *Journal for General Philosophy of Science, 48,* 125–142.
Wilkenfeld, D. A. (2014). Functional explaining: A new approach to the philosophy of explanation. *Synthese, 191,* 3367–3391.
Willingham, D. T. (2017). A mental model of the learner: Teaching the basic science of educational psychology to future teachers. *Mind, Brain, and Education, 11,* 166–175.
Zelcer, M. (2013). Against mathematical explanation. *Journal for General Philosophy of Science, 44,* 173–192.