Support for proof as a cluster concept:
An empirical investigation into mathematicians' practice

Keith Weber
Rutgers University

Abstract. In a previous RUME paper, I argued that proof in mathematical practice can profitably be viewed as a cluster concept in mathematical practice. I also outlined several predictions that we would expect to hold if proof were a cluster concept. In this paper, I empirically investigate the viability of some of these predictions. The results of the studies confirmed these predictions. In particular, prototypical proofs satisfy all criteria of the cluster concept and their validity is agreed upon by most mathematicians. Arguments that satisfy only some of the criteria of the cluster concept generate disagreement amongst mathematicians with many believing their validity depends upon context. Finally, mathematicians do not agree on what the essence of proof is.

Key words: Cluster concept; Mathematical practice; Proof

Mathematics educators agree that an important goal of mathematics education is to improve students' abilities to write proofs. Unfortunately, there is also a consensus that mathematics educators do not agree on what a proof is (Balacheff, 2002; Reid & Knipping, 2010; Weber, 2009). In a previous RUME theoretical report (Weber, 2014), I suggested that proof in mathematicians' practice might profitably be viewed as a *cluster concept* in the sense of Lakoff (1987). Essentially, this means that there may not be a precise definition that distinguishes a proof from a non-proof; rather, proof is actually a cluster of characteristics where a proof was expected to satisfy most or all of the characteristics but an argument might still be a proof if any one or two of the characteristics were not. I claimed that this had the following testable hypotheses:

(i) Mathematicians would believe that an argument that satisfied all characteristics of the proof cluster would be regarded as a proof by all mathematicians and would not be viewed as controversial.
(ii) Arguments that satisfied some, but not all, of the characteristics of the proof cluster would be viewed as controversial by mathematicians. There would not be a consensus on whether these arguments were proofs and such evaluations would be context-dependent.
(iii) Mathematicians would not agree on what the true essence of proof was.

In the studies reported in this contributed report, I specifically test whether these hypotheses were true.

Theoretical perspective

The goal of this paper is to test the viability of the theoretical perspective that proof is a cluster concept. I begin by briefly summarizing the arguments from Weber (2014). I start with the presumption made by many mathematics educators: we want our definition of proof to be descriptive and align with mathematical practice[1]. That is, the arguments that we define to be proofs must include the proofs that mathematicians actually read and write. As

[1] A mathematics educator need not adopt this presumption (c.f., Staples, Thanheiser, & Bartlo, 2012). This article offers consequences for those who accept the presumption that our delineation of proof should be accountable to mathematicians' practice, which I believe constitutes the majority of mathematics educators who are studying proof.

educators, we are not satisfied with define proofs as types of formal derivations that would exclude nearly all of the proofs in the published literature (CadwalladerOlsker, 2011). As Lakoff (1987) observed, when we try to define categories such as proof, we naturally try to list a set of properties that all proofs satisfy. However, Lakoff also argued that most real world concepts and many scientific ones cannot be defined in this way. In Weber (2014), I present arguments for why proof is an example of a concept that cannot be defined by properties that all proofs satisfy. One alternative that Lakoff (1987) suggested is that some concepts are cluster concepts which occurs when as “a number of cognitive models combine to form a complex cluster that is psychologically more basic than the models taken individually” (p. 74).

I suggested that proof might be a cluster concept with six components: (i) a proof is a convincing argument; (ii) a proof is a perspicuous argument; (iii) a proof is deductive and non-ampliative; (iv) a proof is sufficiently transparent so that a knowledgeable mathematician can fill in any gaps; (v) a proof is written in a representation system with agreed upon methods of inference; and (vi) a proof is an argument that is sanctioned by the mathematical community. In Weber (2014), I give a more detailed account and rationale for these criteria. I also claimed that no single criterion above is sufficient to define proof. For each criterion, we can find arguments accepted as proofs by (most) mathematicians that fail to satisfy that criterion. (e.g., computer-assisted proofs are not perspicuous and we do not expect a knowledgeable mathematician to be able to complete each of the gaps contained within that proof).

If proof can productively be conceptualized as a cluster concept, then this makes three predictions. First, an argument that satisfies all elements of the cluster concept should be viewed as prototypical and non-controversial. Mathematicians should all agree that such an argument is a proof independent of context and expect their colleagues to agree with them. Second, if an argument satisfies some but not all elements of the cluster concept, it should be viewed as an atypical proof whose validity is questionable. There should be variance in mathematicians’ responses and they should be aware of the controversial nature of these arguments. Third, proof does not have an “essence”. That is, mathematicians should not agree on which criterion in the proof concept is most important.

Citing philosophers of mathematical practice, in Weber (2014), I argued that each of the above hypotheses is plausible. In the current paper, I complement these theoretical arguments with an empirical study. As I will discuss in the contributed report, it is somewhat problematic to rely on writings about mathematical practice as the claims in the literature are often contradictory. Take, for example, the claim that computer-assisted proofs are controversial proofs that are ultimately accepted by mathematicians, a claim that has been made by philosophers (e.g., Aberdein, 2009) and mathematics educators (e.g., Dreyfus, 2004). In this contributed report, I provide empirical support for this claim, which might lead a skeptic to say that I am merely verifying the obvious. Hence, it is important to note that there are philosophers and mathematicians who write about computer-assisted proofs as being *uncontroversial*, arguing that they are clearly epistemologically on par with more conventional proofs (Fallis, 1996; Montano, 2012) and even that debates about their validity are “anachronistic” as this issue has been decided years ago (Fallis, 1996). On the other side, there are those who claim that computer-assisted proofs- in spite of being undeniably correct- are *not* recognized as proofs by the mathematical community (Rota, 1997) and others who believe that computer-assisted proofs are fundamentally unreliable for obtaining conviction (Jean-Pierre Serre; as cited in Raussen & Skau, 2004). With the exception of Serre, each of the authors cited in this paragraph used their assumptions about computer-assisted proofs as starting points to deduce strong conclusions about mathematical practice; they cannot all be right. The main point here is that citing philosophers and mathematicians to justify empirical claims about mathematical practice is problematic as there is internal disagreement between the groups; by carefully choosing who is cited, a researcher can find grounds to justify both a strong claim, its negation, and qualified version of that claim[2].

Methods

Participants. These studies are comprised of two internet-based survey in which mathematicians were asked to evaluate the validity of five purported proofs. The rationale, validity, and methodolgoy of using the internet to obtain a large sample of mathematicians has been discussed elsewhere (Inglis & Mejia-Ramos, 2009; Lai, Weber, & Mejia-Ramos, 2012) and is not discussed in detail here for the sake of brevity. For the first study, e-mails were sent to the secretaries at the mathematics departments at 25 large state universities in the Great Britain. In these e-mails, the secretaries were asked to forward a request to participate in the study with a link to the study's website to the faculty members of their department. Through this process, 95 mathematicians agreed to participate in the study and completed the survey. For the second study, the same process was completed with 25 large state universities in the United States, yielding a total of 110 mathematician participants

Procedure. In the first study, which I will call the **proof evaluation** survey, participants were told that they would be asked to make validity judgments on five mathematical arguments from number theory. The participants were told that the focus of the study was on the type of reasoning within the argument and that no attempt was being made to deceive them. They were then told that each proof was published, each sentence in the argument was true, and each calculation was carried out correctly. These provisions were put in place because my previous research has shown that validating proofs in number theory can be a time-consuming process for those who did not specialize in that area (Weber, 2008), which would limit the number of mathematicians who would invest the time to complete this survey. Further, I wanted to avoid generating disagreement amongst mathematicians due to performance errors (c.f., Inglis et al, 2013)-- that is, I did not want mathematicians to disagree on whether a proof was valid because some mistakenly thought a true statement was false. I was interested in the *types* of reasoning mathematicians considered valid in a proof rather than their evaluations of particular arguments.

The participants were then shown five arguments in a randomized order and told the publication source from where the argument came. The five arguments were:

- Prototypical Proof 1 (PP1): A conventional proof that "The n^{th} prime p_n satisfies $p_n \leq 2$^(2^{n-1}) for all $n \geq 1$" taken from Jones and Jones (1998) *Elementary Number Theory* textbook that was published by Springer.
- Prototypical Proof 1 (PP2): A conventional proof that, "if n is a number of the form $6k$-1, then n is not perfect" by Holdener (2002) that appeared in the *American Mathematical Monthly*.
- Empirical Proof (EP): An empirical argument to support "if n is an odd integer, then n^2 is an odd integer" based on verifying the claim for n = 1, 3, and 5. The participants were told this appeared in Weber (2003)[3].

[2] To be clear, quantitative studies are certainly not without their limitations as well. The point is that it is better to have a good theoretical argument *and* quantitative evidence to support it. This is especially true in the case of investigating mathematical practice, where the leading theoretical experts do not agree about factual claims about what arguments mathematicians accept.

[3] The argument did appear, but as a common type of invalid student proof. However, based on recent studies (Iannone & Inglis, 2010; Weber, 2010), I no longer think these proofs are that common amongst mathematics majors in proof-based courses.

- Visual Proof (VP): A visual proof of the claim that "if n is an odd integer, then n^2 is congruent to 1 (mod 8)" by Nelsen (2008) that appeared in *Math Horizons*.
- Computer Assisted Proof (CAP): A modification of a computer-generated proof that $\pi = \sum_{k=0}^{\infty} \frac{1}{16^k}\left(\frac{4}{8k+1} - \frac{2}{8k+4} - \frac{1}{8k+5} - \frac{1}{8k+6}\right)$ given by Adamchik and Wagon (1996) in the *American Mathematical Monthly*.

After each proof was presented, participants were asked to make four judgments:

- On a scale of 1-10, how typical was the reasoning used in this proof of the proofs that they read and write?
- In their estimation, was this argument a valid proof? (yes/no)
- What percentage of mathematicians did they think would agree with their judgment? (>90%, 71-90%, 51-70%, <50%)
- For a more nuanced judgment on validity, did they think that: (i) The proof was valid in nearly all mathematical contexts, (ii) I think the proof is valid but there are some mathematical contexts in which it would be invalid, (iii) I think the proof is invalid, but there are some mathematical contexts in which it would be valid, and (iv) The proof would be invalid in nearly all mathematical contexts.

The prediction is that the two conventional proofs, which satisfied all the criteria in the cluster concept, would not be controversial. They would be regarded as prototypical proofs (scoring high on the first judgment), widely recognized as valid (most participants would answer "yes" to the second judgment) independent of context (most participants would answer (i) for the fourth judgment), and most would believe that the mathematical community would agree with them (most participants would answer >90% on the third judgment). Likewise, the empirical argument that satisfies none of the criteria of cluster concept would also not be controversial. Most participants would say this was not a proof, independent of context, and would expect their colleagues to agree.

The visual proof and computer-generated proof satisfy some, but not all, criteria of the cluster concept. Visual proofs are not written in a conventional representation system and computer-generated proofs are not perspicuous and contain gaps that could not be necessarily filled in by a knowledgeable mathematician. Hence the prediction is that these proofs would be controversial. Mathematicians would find these to be atypical of the proofs that they read (scoring low on the first judgment), would disagree on their validity (there would be a significant percentage of participants who answered yes to the second judgment but also a significant percentage who answered no), would be aware that there was disagreement (most participants would not answer >90% on the third judgment), and would think the validity of the proof was contextual (most participants would answer (ii) or (iii) for the third question).

In the second survey, which I call the **proof essence** survey, participants were asked what they believed the essence of a proof was and were given nine options to choose from:

1. A proof provides a mathematician with certainty that a theorem is true
2. A proof provides a mathematician with a high degree of confidence that a theorem is true
3. A proof is a deductive argument with each step being a logical consequence from previous steps
4. A proof is a blueprint from which a mathematician could write a complete formal proof if he or she desired
5. A proof, in principle, can be translated into a formal argument in an axiomatized theory
6. A proof explains why a theorem is true
7. A proof convinces a particular mathematical community that a result is true

8. None of the above captures the essence of proof
9. There is no single essence of proof

If proof is a cluster concept, then we would predict that there is no single criterion that captures the essence of proof. Hence, we would not expect the majority of participants to choose any one of these responses.

Results

Proof	Mean Typicality Rating	Validity	Judgment	Anticipated	Level of	Agreement	
		Valid	Invalid	91-100%	71-90%	51-70%	0-50%
PP1	7.4	99%	1%	90%	9%	1%	0%
PP2	6.8	98%	2%	78%	20%	2%	0%
VP	2.6	62%	38%	14%	46%	33%	7%
CAP	2.7	39%	61%	10%	41%	37%	12%
EP	1.6	0%	100%	92%	0%	1%	6%

Table 1. Participants' judgments on the validity of the five proofs that they read

Proof	Valid proof in nearly all contexts	Valid proof but invalid in some contexts	Invalid proof but valid in some contexts	Invalid proof in nearly all contexts
PP1	94%	5%	1%	0%
PP2	79%	20%	0%	1%
VP	21%	33%	40%	6%
CAP	10%	33%	42%	15%
EP	1%	1%	3%	95%

Table 2. Participants' judgment on the more fine-grained question on utility

The results of the proof evaluation survey are presented in Tables 1 and 2. The results of the study confirmed the predictions. For PP1 and PP2, the large majority of participants claimed the arguments were valid, valid in nearly all mathematical contexts, and thought most of their peers (>90%) would agree with them. The median score for how representative these proofs were of what they actually read and wrote was about seven. For VP and CAP, there was substantial disagreement amongst the participants, the participants were mostly aware that at least 10% of their colleagues would disagree with them, and the majority thought the validity of the proof depended on context.

For the Essence phase of the study, no participant chose "none of the above" and 11% chose 9, that there was no single essence of proof. No choice gathered the majority of the participants; the fourth choice (that proof was a blueprint where a knowledgeable mathematician could fill in every gap) was the most popular, chosen by 25% of the participants, and the first choice (that proof provided certainty) being chosen by 22% of the participants. Every option aside from 8 (none of the above) was chosen by at least three participants.

Discussion and significance

The proof evaluation phase of the study

There are philosophers and mathematics educators who claim that there is a very high rate of agreement amongst mathematicians as to whether a particular argument is a proof or not (e.g., Azzouni, 2004; Selden & Selden, 2003). However, there are also philosophers and mathematics educators who challenge this claim (e.g., Aberdein, 2009; Auslander, 2008; Dreyfus, 2004; Inglis et al., 2013; Rav, 2007; Weber, Inglis, & Mejia-Ramos, 2014). The data presented here offer a potential approach to resolve this discrepancy. For *typical proofs*, mathematicians may indeed usually agree on their validity. Disagreements may arise due to performance errors (e.g., a reviewer overlooks a flaw in the proof), but this could presumably be resolved in a conversation between mathematicians, as Selden and Selden (2003) suggested. The disagreements do not concern the legitimacy of the *type* of reasoning being used. Hence, those who highlight mathematicians' "unusual degree of agreement about the correctness of arguments" (Selden & Selden, 2003, p. 7) seem to be correct in the following sense: for the proofs that mathematicians *typically* encounter in their working lives, it may well be the case that there is usually a high level of agreement amongst mathematicians on the validity of these proofs.

However, for *atypical proofs*, arguments that satisfy some but not all criteria of the cluster concept, disagreement on validity is common and mathematicians are aware of it. Importantly, the majority may think validity judgments about these proofs are *contextual*. (A good follow-up study would be to interview individual mathematicians to get a better sense of what these contexts are). Hence, those who challenge the claim that mathematicians share the same standard of proof are right to note that there are classes of proofs where this is not so.

The finding about the validity of atypical proofs being contextual has a useful consequence for methodological design. In a sense, we can say that asking someone whether a visual argument is a proof is not a well-formed question. The majority of the participants in this study felt the answer depended on mathematical context. In general, asking individuals to judge whether an imperfect argument without a fatal flaw is a proof to make a binary judgment on the argument's validity might be asking an artificial and unreasonable question. It might be better to ask *in what sense* is the argument a proof (and in what sense is it not) and *in what contexts* the argument would be acceptable (and in what contexts would it not be acceptable).

The essence phase of the study

The data on the proof essence phase of the study offer a strong challenge to a researcher who wants to describe what proof essentially is to mathematicians. For instance, take the claim that proof is, at its essence, a convincing argument- an assertion made by numerous mathematics educators (e.g., Balacheff, 1987; Harel & Sowder, 1998; Mason, Burton, & Stacey, 1982) and some philosophers (e.g., Davis & Hersh, 1981). If this were so, we might expect that for the essence question, most participants in this study would have chosen option 1 (proofs provide a mathematician with certainty), 2 (proofs provide a mathematician with a high degree of confidence), or 7 (a proof convinces a mathematical community). Perhaps some participants might have chosen 9 (there is no single essence of proof) on the grounds that a proof needed to be convincing *and* something else. Yet if we add the number of participants who chose 1, 2, 7, or 9, we only reach 41%. It seems difficult to claim that mathematicians essentially view proof as a convincing argument if the majority of mathematicians chose another facet of proof that proof is essentially about (in particular, choices 3, 4, 5, and 6). To avoid misinterpretation, no single study can be offered as a definitive rebuttal to the claim that many mathematicians view proof as something other than a convincing argument. What I do contend is that those who want to claim that proof is essentially about conviction (or explanation or anything else) should at least be held to *account* for these empirical findings.

This offers a practical suggestion for teachers or researchers who desires that proof in their classrooms to be epistemologically consistent with mathematicians' practice. They should not take conviction, explanation, social acceptance, or deduction as the primary criteria for what constitutes a proof. Different mathematicians place different weight on the importance of each of these. My contention is that good proofs satisfy *all* of these roles and I would encourage classroom research to reflect that.

References

Aberdein, A. (2009). Mathematics and argumentation. *Foundations of Science*, 14, 1-8.

Adamchik, V. & Wagon, S. (1997). A simple formula for π. *American Mathematical Monthly*, 104, 852-855.

Auslander, J. (2008). On the roles of proof in mathematics. In B. Gold & R. A. Simons (Eds.) *Proofs and other dilemmas: Mathematics and philosophy* (pp. 61-77). Washington, DC: Mathematical Association of America.

Azzouni, J. (2013). The relationship of derivations in artificial languages to ordinary rigorous mathematical proof. *Philosophia Mathematica*, 21, 247-254.

Balacheff, N. (1987). Processus de preuves et situations de validation. *Educational Studies in Mathematics*, 18(2), 147-176.

Balacheff, N. (2002). The researcher epistemology: A deadlock from educational research on proof. In F.L. Lin (ed.) *2002 International Conference on Mathematics Education-Understanding proving and proving to understand*. NCS and NSUT: Taipei.

CadwalladerOlsker, T. (2011). What do we mean by mathematical proof? *Journal of Humanistic Mathematics*, 1, 33-60.

Davis, P. and Hersh, R. (1981). *The mathematical experience*. New York: Viking Penguin Inc.

Dreyfus, T. (2004) What counts as proof in the mathematics classroom? In Kourkoulos, M., Troulis, G. & Tzanakis, C. (Eds.), *Proceedings of the 3rd Colloquium on the Didactics of Mathematics* (pp. 114-132). Rethymnon, Crete: Department of Education, University of Crete.

Fallis, D. (1996). Mathematical proof and the reliability of DNA evidence.*American Mathematical Monthly*, 103, 491-497.

Harel, G. & Sowder, L. (1998). Students proof schemes. *Research in Collegiate Mathematics Education, 3*, 234-282.

Holdener, J. (2002). A theorem of Touchard on the form of odd perfect numbers. *American Mathematical Monthly*, 109, 661-663.

Iannone, P. & Inglis, M. (2010). Self-efficacy and mathematical proof: Are undergraduates good at assessing their own proof production ability? In *Proceedings of the 13th Conference for Research in Undergraduate Mathematics Education*. Raleigh, North Carolina.

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

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

Jones, G. & Jones, J. (1998). *Elementary number theory*. London: Springer.

Lai, Y., Weber, K., & Mejía-Ramos, J. P. (2012). Mathematicians' perspectives on features of a good pedagogical proof. *Cognition and Instruction*, *30*(2), 146-169.

Lakoff, G. (1987). *Women, fire, and dangerous things.* Cambridge University Press: Cambridge.

Mason, J., Burton, L., & Stacey, K. (1982). *Thinking mathematically*. London: Addison Wesley.

Montaño, U. (2012). Ugly mathematics: Why do mathematicians dislike computer-assisted proofs?. *The Mathematical Intelligencer*, *34*(4), 21-28.

Nelsen, R. (2008). Visual gems in number theory. *Math Horizons*, 31, 7-9.

Raussen, M., & Skau, C. (2004). Interview with Jean-Pierre Serre. *Notices of American Mathematical Society*, *51*(2), 210-214.
Rav, Y. (2007). A critique of a formal-mechanist version of the justification of arguments in mathematicians' proof practices. *Philosophia Mathematica*, 15, 291-320.
Reid, D. & Knipping, C. (2010). *Proof in mathematics education*. Sense Publishers: Rotterdam.
Rota, G. C. (1997). The phenomenology of mathematical beauty. *Synthese*, *111*(2), 171-182.
Selden, A. & Selden, J. (2003). Validations of proofs written as texts: Can undergraduates tell whether an argument proves a theorem? *Journal for Research in Mathematics Education*, *34*, 4-36.
Staples, M., Bartlo, J., & Thanheiser, E. (2012). Justification as a teaching and learning practice: Its multifacted (potential) role in middle school classrooms. *Journal of Mathematical Behavior*, 31, 447-462.
Weber, K. (2003). Students' difficulties with proof. In A. Selden and J. Selden (eds.) *Research Sampler,* 8. Available from the following MAA website on the teaching of learning of mathematics: http://www.maa.org/t_and_l/sampler/research_sampler.html. Last downloaded April 18, 2013.
Weber, K. (2008). How mathematicians determine if an argument is a valid proof. *Journal for Research in Mathematics Education*, 39, 431-459.
Weber, K. (2009). Review of Boero's Theorems in schools. *Mathematical Thinking and Learning*, 11, 289-294.
Weber, K. (2010). Mathematics majors' perceptions of conviction, validity, and proof. *Mathematical Thinking and Learning*, *12*(4), 306-336.
Weber, K. (2014). Proof as a cluster concept. In T. Fukawa-Connelly, G. Karokak, K. Keene, & M. Zandieh (Eds.) *Proceedings of the 17th Annual Conference on Research in Undergraduate Mathematics Education*. Denver, CO.
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.