Running head: MATHEMATICIANS’ DIFFERENT STANDARDS WHEN EVALUATING PROOFS

On Mathematicians' Different Standards When Evaluating Elementary Proofs

Matthew Inglis

Loughborough University

Juan Pablo Mejia-Ramos

Rutgers University

Keith Weber

Rutgers University

Lara Alcock

Loughborough University

Address correspondence to Matthew Inglis, Mathematics Education Centre, Loughborough University, Loughborough, United Kingdom. Email: m.j.inglis@lboro.ac.uk.

Abstract

In this paper, we report a study in which 109 research-active mathematicians were asked to judge the validity of a purported proof in undergraduate calculus. Significant results from our study were: (a) there was substantial disagreement among mathematicians regarding whether the argument was a valid proof, (b) applied mathematicians were more likely than pure mathematicians to judge the argument valid, (c) participants who judged the argument invalid were more confident in their judgments than those who judged it valid, and (d) participants who judged the argument valid usually did not change their judgment when presented with a reason raised by other mathematicians for why the proof should be judged invalid. These findings suggest that, contrary to some claims in the literature, there is not a single standard of validity among contemporary mathematicians.

On Mathematicians' Different Standards When Evaluating Elementary Proofs

1. Introduction

Mathematical proof is a primary means of conveying mathematical content. While it may be impossible to provide an operational definition of what constitutes a mathematical proof (e.g., Davis & Hersh, 1981), a preliminary definition describes a proof as a sequence of statements each of which is either accepted as true a priori or is a necessary logical consequence of previous statements in the proof (e.g., Auslander, 2008; Kitcher, 1984). One purported virtue of this genre of argumentation is that each statement in a proof is either a valid logical consequence of previous assertions or it is not. As Rota (1993) noted, as opposed to argumentation in most other disciplines, "mathematical proof does not admit degrees" (p. 93). As a consequence, many in the mathematical community believe that the validity of a proof is not a subjective issue but an objective fact, and cite the high level of agreement between mathematicians on what constitutes a proof as support for their position.

Azzouni's (2004) article, for instance, attempted to explain why "mathematicians are so good at agreeing with one another on whether some proof convincingly establishes a theorem" (p. 84). McKnight, Magid, Murphy and McKnight (2000) asserted that "all agree that something is either a proof or it isn't and what makes it a proof is that every assertion in it is correct." (p.1). Selden and Selden (2003) remarked on "the unusual degree of agreement about the correctness of arguments and the truth of theorems arising from the validation process" (p.7); they contended that validity is a function only of the argument and not of the reader: "Mathematicians say that an argument proves a theorem, not that it proves it for Smith and possibly not for Jones" (p. 11).

Other mathematicians and philosophers, however, have been more skeptical about whether there are universally-adopted standards of mathematical proof. Auslander (2008) suggested that "standards of proof vary over time and even among mathematicians at a given time" (p. 62). Rav (2007) argued that mathematical practice has a "pluralistic nature" and concluded that "not only is mathematical proof 'time-dependent', but because of the historical and methodological wealth of proof practices (plural), any attempt to encapsulate such multifarious practices in a unique and uniform one-block perspective is bound to be defective" (p. 299).

The general purpose of this article is to provide empirical support for the latter position. To situate our contribution more precisely, we note two different sources of disagreements that mathematicians may have when evaluating a proof: performance errors in the validation of a proof, and different standards about what constitutes a proof in an established branch of mathematics.[1]

With regards to performance errors, two mathematicians may disagree on whether a proof is valid because one mathematician has overlooked a flaw in the proof. Indeed, some argue that this occurs surprisingly frequently in mathematical practice. Davis (1972), for instance, estimates that as many as half of all published proofs contain logical errors, perhaps because journal referees do not always check every line of a proof that they review (Geist, Löwe, & Van Kerkhove, 2010; Szpiro, 2003; Weber & Mejia-Ramos, 2011). However, when disagreements are due to performance errors, it is believed that the mathematicians could resolve these

[1] We distinguish here between disagreements about standards in an established branch of mathematics and disagreements about standards in a new and developing branch (for example, debates surrounding the way computers contributed to the proof of the Four Color Theorem). The proof we discuss in this paper is from a well-understood and non-controversial area of mathematics: elementary calculus.

disagreements with discussions about the problematic areas in the proof (as Selden and Selden, 2003, suggest).

In this paper, we demonstrate that there is a second source of disagreement about what mathematicians consider to be a proof. By asking 109 mathematicians to determine whether an elementary proof from undergraduate calculus was valid, we found that mathematicians used different standards in judging the validity of this proof and that the standards held by mathematicians were related to their research areas.

2. A negative characterization of individual validity judgments

In order to frame both the design and results of our empirical study, we first discuss the ways in which individuals make judgments about validity. We argue that while mathematicians are apparently willing to make positive judgments that particular proofs are valid, these can more accurately be characterized as negative judgments that such proofs are not invalid. We suggest here that this characterization leads to testable predictions for an individual validator’s behavior.

A *positive* characterization of individual validity judgments is based upon the idea that, when reading a proof, mathematicians attempt to mentally reconstruct the flow of inferences in the proof using the written text and their knowledge of the mathematical domain (perhaps using ‘inference packages’ as suggested by Azzouni, 2005). In our *negative* characterization, on the other hand, we suggest that a validation attempt can more accurately be seen as a search for problems with a purported proof. Such problems might be of (at least) two types, the detection of either of which might not be trivial. One type of problem is a genuine logical error. Such an error might be identified through a detailed line-by-line check, which reveals a statement that does not follow as a necessary consequence of previous statements. Or it might be identified through a higher-level check that establishes that methods have been applied inappropriately or that a new method is invalid (see Rav, 1999; Weber & Mejia-Ramos, 2011, for further discussion of approaches to proof validation). Of course, errors can be subtle: contemporary mathematics contains many invalid arguments that were believed to be proofs because the errors in these arguments were difficult to detect (Devlin, 2003, described several such cases in detail). Consequently, any individual validator knows that they might fail to detect such a problem.

Another type of problem is a serious gap. Gaps are not always problematic: in general they are a natural feature of the proofs typically found in textbooks and research papers. Such proofs differ from what Rav (1999) described as derivations, the latter being formal sequences of formulae where each element in the sequence is either an axiom or follows from an axiom by an accepted rule of inference. Because the vast majority of proofs are not derivations, not every statement in a proof follows *directly* from the earlier statements; often a reader must bridge gaps by constructing subproofs. Consequently, when a validator is reading a proof they must decide whether any gaps that they find are sufficiently serious or large to warrant rejecting the proof as invalid. Again, any individual validator knows that they might fail to make an appropriate judgment about such a gap.

For the mathematical community as a whole, the potential existence of errors and gaps in proofs is not unduly disruptive. Even with a negative characterization of validity judgments in mind, very high levels of confidence in a theorem and its proof can arise in time, for two reasons. First, if many mathematicians study a proof, we gain confidence that if an error existed, it would be discovered. Second, as Dawson (2006) argued, very high levels of confidence may also be gained through the generation of different proofs of the same theorem, which in a sense may serve as independent verifications of one another.

For individual validators, however, the potential existence of errors and gaps has consequences in terms of the balance of confidence with which validity judgments can be made. If a gap or a problematic statement or method is located, the validator can be confident that the proof (as written) is not correct. If, however, no such gap, statement or method is found, the validator cannot with absolute confidence conclude that none exists: a problem might simply have eluded detection. Why then would a validator ever conclude that a proof is valid? We suggest that this happens when, after a validation attempt conducted with (what they perceive to be) due care and attention, the validator has failed to reject it as invalid.

Thus the positive and negative characterizations described here differ in the way they characterize the validator's goals. In one case the goal is to find a problem, in the other it is to reconstruct the flow of inferences. Based on the assumption that a validator will be more confident about their judgment (and will find it easier to justify) when it is based on reaching their goal than when it is based on failing to reach their goal, our negative characterization, therefore, leads to the following two predictions about validator behavior:

1. When validating a purported proof, those who regard it as invalid will be more confident in their judgment than those who regard it as valid (because they have found a problem, rather than merely having failed to find one).
2. It will be easier for validators to justify their response if they have rejected the proof as invalid rather than accepting it as valid (because those who rate it valid have nothing to say beyond that they have failed to find a problem).

Note that a positive characterization of validity would make the opposite predictions. If validators successfully reconstruct the flow of inferences from premises to conclusions, they would confidently accept the proof as being valid, so we would expect those mathematicians who rated the proof as 'valid' would be more confident in their judgments than those who rated the proof 'invalid' (as they had successfully achieved their aim of reconstructing the flow of inferences). Similarly, under a positive characterization of proof, those mathematicians who rated the proof invalid would have nothing to offer in justification beyond reporting that they failed to reconstruct the proof; whereas those who rated it valid would be able to elaborate upon their reconstruction and report that it had occurred successfully.

We therefore believe that predictions 1 and 2 offer a method of empirically distinguishing between positive and negative characterizations of proof validation, and one of the goals of our study was to test these accounts.

3. Study design in relation to prior work on validity judgments

The work reported here builds on two psychological studies (Inglis & Alcock, in press; Weber, 2008) that we conducted in response to Selden and Selden's (2003) research on the cognitive processes of undergraduate mathematics majors engaged in proof validation. In their study, Selden and Selden (both published research mathematicians) asked students to evaluate an argument they labeled "the real thing", which they judged to be a fully valid proof, and another they labeled "the gap", which they evaluated as invalid. Weber (2008) gave these and other proofs to eight mathematicians and asked them to determine whether or not they were valid. One declared "the real thing" invalid, two judged "the gap" to be valid and another said that such a judgment about "the gap" was impossible without context. In a separate study, Inglis and Alcock (in press) found similar results – 5 of 12 mathematicians judged "the real thing" invalid and 5 of 12 mathematicians judged "the gap" to be valid. Selden and Selden's proofs were short and simple so we were surprised to see this level of disagreement among other mathematicians.

However, these findings gave only limited information about actual mathematical practice since they were based on variants of student-produced proofs and were therefore somewhat awkwardly written.

This drawback was addressed to some extent by the inclusion in Inglis and Alcock's (in press) study of the following purported proof:

Theorem. $\int x^{-1} dx = \ln(x) + c$.

Proof. We know that $\int x^k dx = \dfrac{x^{k+1}}{k+1} + c$ for $k \neq -1$.

Rearranging the constant of integration gives $\int x^k dx = \dfrac{x^{k+1} - 1}{k+1} + c'$ for $k \neq -1$.

Set $y = \dfrac{x^{k+1} - 1}{k+1}$, and take the limit as $k \rightarrow -1$ as follows.

Let $m=k+1$, and rearrange $y = \dfrac{x^{k+1} - 1}{k+1}$ to give $x^m = 1 + ym$ or $x = (1+ym)^{\frac{1}{m}}$.

Set $n = \dfrac{1}{m}$. Then $x = (1+ym)^{\frac{1}{m}} = \left(1 + \dfrac{y}{n}\right)^n \rightarrow e^y$ as $n \rightarrow \infty$, by properties of *e*.

As $n \rightarrow \infty$ we have $m \rightarrow 0$, so $k \rightarrow -1$.

In other words, $x \rightarrow e^y$ as $k \rightarrow -1$, so $y \rightarrow \ln(x)$ as $k \rightarrow -1$.

So $\int x^k dx = \dfrac{x^{k+1} - 1}{k+1} + c' = y + c' \rightarrow \ln(x) + c'$ as $k \rightarrow -1$. So $\int x^{-1} dx = \ln(x) + c'$.

This proof is more substantial than those used by Selden and Selden, and is written in a standard format and style. Nonetheless, it involves relatively straightforward undergraduate mathematics, and the results were similar: 6 of the 12 mathematicians who participated in Inglis and Alcock's study judged the proof valid and 6 judged it invalid.

These studies all had small sample sizes because they were all designed to focus on the methods by which mathematicians and undergraduates validate purported proofs, not their final judgments. To investigate whether these apparent differences in validity judgments were genuine, we thus asked a large number of research-active mathematicians to judge the validity of this proof. This paper reports the results.

4. Method

Given the general difficulty of obtaining large samples of research-active mathematicians, we decided to maximize our sample size by collecting our data through the internet. Web-based research methods present some practical difficulties, such as the possibility of participants submitting multiple responses. However, by taking certain precautions (outlined by Reips, 2000), these methods have been found to produce results that are consistent with those found by traditional experimental methods (Gosling, Vazire, Srivastava, & John, 2004; Krantz & Dalal, 2000). We followed the strategies employed by Inglis and Mejia-Ramos (2009a, 2009b) to conduct internet studies in mathematics education research.

4.1 Participants

Participants were 109 research-active mathematicians (56 PhD students and 53 academic staff) associated with Australian and Canadian universities. They were recruited through an email sent via their departmental secretary. Those mathematicians who chose to take part in the study clicked a link contained in the email, which directed them to the study website.

4.2 Procedure

The study website consisted of six pages of information and questions which participants moved through at their own pace.

Page 1 contained background to the study and reminded participants that they should only continue if they were a research-active mathematician.

Page 2 asked participants to provide demographic information about themselves: their status (PhD student or academic staff), number of years of experience in teaching undergraduates, broad research area (applied mathematics, pure mathematics or statistics), and their specific research area (AMS subject classification).

Page 3 gave the following instruction, which contextualized the task: "Below is a proof of the type that might be submitted to a recreational mathematics journal such as *The Mathematical Gazette*. Please read the proof and decide whether or not you think it is valid." Participants were then presented with the proof of the theorem stating that $\int x^{-1}dx = \ln(x) + c$, as given earlier. After reading the proof, participants were asked two questions: "Do you think the proof is valid or invalid?" and "How certain are you that your answer is correct?". They responded to the first by selecting either "valid" or "invalid", and to the second via a five point Likert scale (from "1 – It was a complete guess" to "5 – I am completely certain"). Finally, they were given the opportunity of explaining their answer via a free response text box.

Page 4 asked participants to estimate what percentage of mathematicians would agree with their judgment about the validity of the purported proof (they responded by selecting 0-9%, 10-19% etc.). They were also asked to suggest reasons another mathematician might have for disagreeing with their judgment.

Page 5 presented participants with an explicit objection to the proof that had been given by a mathematician in a pilot study:

> "They do not say anything about the type of convergence they are dealing with. Therefore, in the last line they seem to assume that the limit and the integral commute, which is false in general."

The original proof was presented alongside the objection, and participants were asked to state whether this objection was "reasonable" (responding "yes", "no" or "unsure"), and whether the objection (on its own) was "enough to render the proof invalid" (again responding "yes", "no" or "unsure"). Finally they were given the opportunity to explain their response via a free text box.

Page 6 thanked participants for their time and gave information about how to receive information about the purpose of the experiment.

5. Results and discussion

5.1 Overview: Disagreement and results by disciplinary area

Of the 109 participants who completed the survey[2], 29 (27%) judged the argument valid, and 80 (73%) judged it invalid. There was no significant difference between the responses of academic staff compared and those doctoral students, p=.522 (throughout this paper, if no test statistic is reported – as here – this indicates that the significance level was calculated using a 2×2 Fisher's Exact Test), consequently we do not distinguish between these groups in the analyses that follow.

Participants' responses were related to their research area, with applied mathematicians more likely than pure mathematicians to judge the argument valid, p=.002. These data are shown in Table 1. The association between research area and validity judgment retained significance when doctoral students were removed from the analysis, p=.011.

Predictions Based on the Negative Characterization of Validity Judgments

[2] Around 65% of those participants who started the survey completed it. There were no significant differences in research area or validity judgments between those who completed the survey and those who dropped out.

As predicted by our negative characterization of mathematical validity judgments, those who judged the argument invalid had higher confidence in their responses, *M*=4.16, than those who rated it valid, *M*=3.41, *U*=658, *p*<.001. Also as predicted, participants who judged the argument invalid seemed to find it easier to justify their responses: they were more likely to leave a comment explaining their answer than were those who rated the argument valid, 64% versus 35%, *p*=.011.

Comments left by those who judged the proof invalid fell into three main categories. Some participants complained about the interchange of the limit and the integral in the last line of the proof (thus flagging the same problem as the pilot study participant whose comment we used). For example, one wrote:

> "Even if all the statements in a 'proof' are correct, it is not a correct proof unless there is a justification of the transition from each to the next. What is the justification for taking the limit under the integral sign?"

Another wrote:

> "The issue is whether or not it is valid to exchange integration with taking the limit $k \to -1$ $(n \to \infty)$. Since the limit is a priori pointwise, one must use a convergence theorem here (e.g. restrict to a compact interval away from zero and use the dominated convergence theorem, or show that the limit is uniform, etc.) It seems that this detail can be corrected, but it would be imprudent to call the proof valid without explaining this point."

A second broad category of comments expressed concern about the manipulation of the constant of integration. One participant wrote:

> "The line 'rearranging the constant of integration' is invalid, since *c'* now depends on *k*, but previously it did not. I did not read further."

The final category of comments were complaints about a lack of clarity regarding what definitions the author was using:

> "The proof doesn't make any sense without defining what 'ln' and 'e' mean. Defining ln(x) by means of the integral of $1/x$ and exp to be the inverse function of ln is one of the standard ways; if this proof were showing something, it would be that this definition is equal to some other definition – but which one?"

We note that all of these comments have a claim to be valid criticisms of the proof, meaning that the 27% of participants who judged the argument valid (among them a disproportionate number of applied mathematicians) must have either not noticed these problems or not considered them sufficient to reject the proof as invalid. We consider various hypotheses to account for these results in the next section.

5.2 Agreement estimates

Both those who judged the proof valid and those who judged it invalid believed that theirs would be the majority view. The overall mean estimate of the number of mathematicians who would agree with participants' judgments was 75.0%. The mean agreement estimate for those who believed the proof was valid was 64.7%, significantly higher than 50%, $t(28)=4.59$, $p<.001$. The equivalent figure for those who believe the proof was invalid was 78.7%, again significantly higher than 50%, $t(79)=13.28$, $p<.001$. However, these figures indicate that participants did not all believe that the majority judgment would be overwhelming. Allowing that (say) one quarter of one's colleagues might disagree seems to indicate a recognition that one might have missed a problem or that others might evaluate gaps as more or less serious. Indeed, relatively few participants thought that there would be uniform agreement: only 30% of participants believed that over 90% of mathematicians would agree with them. This is perhaps surprising given the suggestion that mathematicians exhibit near uniform agreement about validity that has been expressed by Azzouni (2004), McKnight et al. (2000), and others.[3]

5.3 Willingness to change a judgment

All the results presented so far, however, leave open the question of whether the participants who judged the proof valid did not find any problems, or whether they found problems but judged them insufficient to render the proof invalid. Thus we do not yet know why pure and applied mathematicians differed in their evaluations of the proof. The hypothesis that we wish to advance is that pure and applied mathematicians use different standards when deciding whether a problem in a proof is sufficient to render the proof invalid: that they evaluate potential problems differently. However, there are alternative hypotheses. Perhaps applied

[3] As explained in the methods section, we conducted a pilot study in order to obtain an appropriate quote to be used in the second half of the study. Because the pilot was identical to the main study until after participants had read and rated the proof, it can be seen as a replication for the results discussed above. A total of 111 research-active mathematicians took part in the pilot, all from US universities. 77% rated the proof as invalid and 23% as valid. These ratings were related to participants' research areas, p=.005, with applied mathematicians more often rating the argument as valid (43%) than pure mathematicians (17%). Those who rated the argument invalid were more confident in their judgment than those who rated it valid, U=637, p=.001, and there were more comments left by those who rated the argument invalid than by those who rated it valid, p<.001. Note that because the participants in the pilot study were all from US universities, these data render implausible an alternative account suggested by a reviewer: that the relationship between research area (pure/applied) and validity judgment found in our main study could have been confounded by a relationship between geography (Australia/Canada) and validity judgment (as perhaps the balance of pure and applied research being conducted in these two countries is different). All participants in the pilot were from the USA, and we found the same relationship.

mathematicians, who in their practice may be more concerned with computation than deduction, are less adept at seeking logical errors. Or perhaps they read the proof less carefully than the pure mathematicians and consequently were more likely to overlook the problems cited by others. These accounts attribute the disagreement among participants to performance errors on the part of those who rated the proof valid. Alternatively, perhaps applied mathematicians simply look for different sorts of problems when validating a proof, so did not spot the particular problems raised by our proof. All of these alternative hypotheses suggest that instead of noticing a problem with this proof but not judging it sufficiently serious, those mathematicians who rated the proof valid simply did not notice any problems at all.

To investigate this issue we turn to participants' responses to the latter section of the instrument, where they were asked to read a specific objection given by a participant in the pilot study. This objection, given earlier, related to how the author of the proof commuted the limit and integral in the last line.

Recall that participants were asked two questions about the presented objection: whether or not it was "reasonable", and whether or not, on its own, it was enough to render the proof invalid. A substantial majority of all participants believed the objection was reasonable (82% said it was, with a further 9% saying they weren't sure), and these reasonableness judgments were not significantly related to their original validity ratings, p=.113. Crucially, however, participants' responses to the second question *were* related to their original validity ratings, p<.001. A majority (65%) of those who had originally claimed the proof was invalid said that this objection was, on its own, enough to render the proof invalid. In contrast only one participant who had originally rated the proof valid (3%) believed that this objection was sufficient. These data are shown in Table 2.

This result allows us to address the above hypotheses. If those participants who judged the proof valid simply had not noticed that the exchange of limit and integral was problematic, then we would have expected them to change their minds when this was explicitly pointed out to them. In fact only one participant did so.[4] These data provide strong support for our claim that the reason for the disagreement about the validity of the original proof was not that a subset of participants failed to notice problems that others spotted. Rather, we suggest that some participants, disproportionately applied mathematicians, applied different standards when deciding whether a potential problem with the proof was sufficient to render it invalid.

6. General Discussion

The results of this study provide empirical support for the claim that there is not universal agreement among mathematicians regarding what constitutes a valid proof, in the context of a submission to a journal such as *The Mathematics Gazette*. Our findings suggest that pure and applied mathematicians adopt different standards in judging proof validity; in particular, that they apply different standards when judging whether a potential problem in a proof is sufficient to render it invalid. This claim is true even in the domain of relatively elementary mathematics, such as the content of undergraduate calculus. Our study thus has implications for at least three academic audiences.

[4] Note that participants took part in the study anonymously on the internet, so there was no social pressure placed upon them which could have encouraged them to remain faithful to their earlier answer. Nevertheless it is possible that participants' desires to confirm their earlier response was so strong that, despite the lack of social pressure, they were still biased towards unreasonably dismissing an objection they might otherwise have agreed with. Future research could productively investigate this issue.

First, mathematicians are usually aware that students struggle to engage in formal proof-based mathematics. They are also usually aware that they as individuals might require different standards of proof from students at different levels – a beginning student might be required to spell out every step in a way that a more advanced student is not. However, mathematicians might not be aware of the extent to which students could be grappling with contradictory messages; our study indicates that even within central content areas and with a requirement to judge in a particular way, mathematicians disagree about what constitutes validity. They may therefore be giving students inconsistent feedback on what features a proof needs to have in order to be considered valid.

Second, because of these different judgments, researchers conducting psychological studies on proof should take care in their interpretations, particularly when classifying specific proofs as "obviously" valid or invalid. Our studies cast doubt on the validity status of two proofs used by Selden and Selden (2003) (Inglis & Alcock, in press; Weber, 2008). Judgments about the validity of proofs, even at the undergraduate level, appear to be far more nuanced than is commonly thought. Our interpretations of research on student understanding of such proofs should perhaps be similarly nuanced.

Finally, our results allow us to address the widely (albeit, not universally) held philosophical assumption that there is a remarkably high degree of agreement amongst mathematicians regarding whether an argument constitutes a proof (Azzouni, 2004; McKnight et al., 2000). Our results indicate that this assumption may not be correct. Instead, they indicate heterogeneity in mathematical practice that is more in line with Rav's claim (2007) that mathematical practice is "pluralistic". We thus argue that philosophers should be cautious in any attempts to explain the high level of agreement among mathematicians. Such agreement might simply not exist.

Acknowledgements

This work was partially funded by a Royal Society Worshipful Company of Actuaries Research Fellowship (to M.I.). We are grateful to Chris Sangwin for bringing to our attention the purported proof used in this study, and to three anonymous reviewers for insightful comments on an earlier draft of the paper.

References

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. (2004). The derivation-indicator view of mathematical practice. *Philosophia Mathematica*, *12*, 81-106.

Davis, P. (1972). Fidelity in mathematical discourse: Is one and one really two? *American Mathematical Monthly*, *79*, 252-263.

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

Dawson, J.W. (2006). Why do mathematicians re-prove theorems? *Philosophia Mathematica*, *14*, 269–286.

Devlin, K. (2003). The shame of it. From *Devlin's Angle*, an on-line publication of the Mathematical Association of America. Downloaded from: http://www.maa.org/devlin/devlin_archives.html. Last downloaded: April 7, 2011.

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

Geist, C. Löwe, B., & Van Kerkhove, B. (2010). Peer review and testimony in mathematics. In B. Löwe & T. Müller (Eds.) *Philosophy of Mathematics: Sociological Aspects and Mathematical Practice* (pp. 155-178). London: College Publications.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, *59*, 93–104.

Hanna, G. (1995). Challenges to the importance of proof. *For the Learning of Mathematics*, *15*(3), 42-49.

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

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

Inglis, M., & Mejia-Ramos, J. P. (2009b). On the persuasiveness of visual arguments in mathematics. *Foundations of Science*, *14*, 97-110.

Kitcher, P. (1984). *The nature of mathematical knowledge*. Oxford: Oxford University Press.

Krantz, J. H., & Dalal, R. (2000). Validity of web-based psychological research. In M. H. Birnbaum (Ed.), *Psychological experiments on the internet* (pp. 35–60). San Diego: Academic Press.

McKnight, C., Magid, M., Murphy, T.J., & McKnight, M. (2000). *Mathematics education research: A guide for the research mathematician*. American Mathematical Society: Washington, D.C.

Rav, Y. (1999). Why do we prove theorems? *Philosophia Mathematica*, *7*, 5-41.

Rav, Y. (2007). A critique of formalist-mechanist version of the justification of arguments in mathematicians' proof practices. *Philosophia Mathematica*, *12*, 291-320.

Reips, U.-D. (2000). The web experiment method: Advantages, disadvantages, and solutions. In M. H. Birnbaum (Ed.), *Psychological experiments on the internet* (pp. 89–117). San Diego: Academic Press.

Rota, G-C. (1993). The concept of mathematical truth. In A. White (Ed.) *Essays in Humanistic Mathematics* (pp. 91-96). Washington, D.C.: Mathematical Association of America.

Selden A. & Selden J. (2003) Validations of proofs considered as texts: Can undergraduates tell whether an argument proves a theorem? *Journal for Research in Mathematics Education*, *34*, 4-36.

Szpiro, G.C. (2003). *Kepler's conjecture*. New York: John Wiley, 2003.

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

Weber, K. & Mejia-Ramos, P. (2011). How and why mathematicians read proofs: An exploratory study. *Educational Studies in Mathematics*, *76*, 329-344.

Tables

Table 1. Participants' responses to the proof, by research area.

Research Area	Number of Valid Ratings	Number of Invalid Ratings
Applied Mathematics	12	12
Pure Mathematics	13	64

Table 2. Participants' responses to the question "Do you think this objection (on its own) is enough to render the proof invalid?"

Original Rating	Yes	No	Unsure
Valid	1	24	4
Invalid	52	19	9