Running head:

Reading Mathematical Proofs

Expert and Novice Approaches to Reading Mathematical Proofs

Matthew Inglis

Lara Alcock

Loughborough University, United Kingdom

This work was supported by a grant from the MSOR Network of the Higher Education Academy (L. A. & M. I.) and a Royal Society Worshipful Company of Actuaries Research Fellowship (M. I.). We are grateful to Chris Sangwin for bringing Proof 5 to our attention.

Abstract

This article presents a comparison of the proof validation behavior of beginning undergraduate students and research-active mathematicians. Participants' eye movements were recorded as they validated purported proofs. The main findings are that (a) contrary to previous suggestions, mathematicians sometimes appear to disagree about the validity of even short purported proofs; (b) compared with mathematicians, undergraduate students spend proportionately more time focusing on "surface features" of arguments, suggesting that they attend less to logical structure; and (c) compared with undergraduates, mathematicians are more inclined to shift their attention back and forth between consecutive lines of purported proofs, suggesting that they devote more effort to inferring implicit warrants. Pedagogical implications of these results are discussed, taking into account students' apparent difficulties with proof validation and the importance of this activity in both school- and university-level mathematics education.

Key words: Advanced mathematical thinking, College mathematics, Logic and Proof, Proof, Reasoning.

Proof is central to the practice of academic mathematicians, and is increasingly seen as essential to a coherent school-level mathematics curriculum (e.g., Cowen, 1991; Hanna, 2007; National Council of Teachers of Mathematics [NCTM], 2000; Selden & Selden, 2003; Weber, 2008). Consequently, in recent years there has been substantial growth in educational research that focuses on how students engage with proof and proving. There is now an extensive literature on teachers' and students' conceptions of proof (e.g., Bell, 1976; Harel & Sowder, 1998; Knuth, 2002), on how students construct mathematical proofs (e.g., Moore, 1994; Weber, 2001; Weber & Alcock, 2004), and on how pedagogy can be designed to develop students' understanding of proof (e.g. Rowland, 2002; Stylianides, 2007; Stylianides & Stylianides, 2009b). However, it is also recognized that despite this growing literature, studies that have focused on the activity of reading proofs (e.g., Hazzan & Zazkis, 2003; Mamona-Downs & Downs, 2005; Selden & Selden, 2003) are relatively unusual. This is a surprising omission, because it is by reading and evaluating the proofs of others that mathematicians are said to learn new mathematics (Rav, 1999; Selden & Selden, 2003), and because undergraduates are expected to spend substantial study time reading proofs presented in lectures or textbooks (Weber, 2004). Despite this expectation, and despite the apparent ubiquity of proof reading as an activity in higher level mathematics and mathematics education, we apparently fail to teach the skills needed to read effectively: Undergraduate students and precollege teachers typically perform at chance level when asked to determine the validity of even relatively simple purported proofs (Knuth, 2002; Selden & Selden, 2003).

It is clear that pedagogical strategies in this area could usefully be informed by a deeper understanding of the types of behavior that characterize more- and less expert readers of proof.

In this paper we report the first direct comparison of the proof reading behavior of research-active mathematicians and beginning undergraduate students.

PROOF VALIDATION

Our review of the literature is structured around four main themes. First, we review theoretical and empirical investigations into how students and mathematicians judge the validity of purported mathematical proofs. We pay particular attention to disagreements between researchers regarding the extent to which expert mathematicians agree about validity. Second, we discuss theoretical ideas that might account for students' difficulties with proof validation. Third, we review empirical and theoretical discussions about how successful proof validations may be conducted. Specifically, we concentrate on Weber and Mejia-Ramos's (2011) suggestion that there are two distinct strategies that can be adopted. Finally, we discuss methodological difficulties with studying the processes by which students and mathematicians validate proofs. To date, most researchers have presented theoretical ideas based on self-report methods of data collection, and we argue that recording the eye movements of participants as they validate is a valuable way of testing and refining those ideas.

Judging the Validity of Proofs

For anyone involved in mathematics, one important activity is reading a mathematical proof with the aim of determining whether or not it is valid. Selden and Selden (2003) characterized this activity as *proof validation* and suggested that it is a complex process involving evaluating statements, posing and answering questions, constructing subproofs, and recalling definitions and theorems. Validation is only one reason why someone may read a proof: Mejía-Ramos and Inglis (2009) argued, for instance, that a reader's behavior when validating is potentially quite different from their behavior when reading for comprehension, because in proof

comprehension the validity of the proof is assumed (or at least initially assumed) by virtue of its author or source, and the goal of the reader is to understand the proof, not to check it for correctness. Nevertheless, to date, most research on the reading of proofs has focused on proof validation, and this is also our focus here.

Proof validation is clearly important for any teacher or lecturer who may be required to grade students' work in proof-oriented courses, but it is also important for students. *Principles and Standards for School Mathematics* (NCTM, 2000) argued that students should be encouraged to "seek, formulate, and critique explanations so that classes become communities of inquiry" (p. 346); clearly, one way of critiquing explanations is by validating purported proofs. Furthermore, it has been suggested that validating proofs is fundamental to proof construction. Selden and Selden, for example, argued that "constructing or producing proofs is inextricably linked to the ability to validate them reliably, and a proof that could not be validated reliably would not provide much of a warrant" (p. 9).

Proof validation is thus extremely important but appears to be extremely challenging. Selden and Selden (2003) found that without help, the 8 undergraduate students in their study performed no better than chance when asked to determine the validity of four purported number theory proofs. Similar results were found by Alcock and Weber (2005) in the domain of real analysis. These difficulties are not restricted to students: Teachers at both primary and secondary levels often appear to accept invalid arguments as valid proofs (Knuth, 2002; Martin & Harel, 1989).

Selden and Selden (2003) drew a sharp contrast between the chance-level performance of their student participants and the behavior of research mathematicians, claiming that mathematicians exhibit "remarkably uniform agreement" (p. 11) about whether a proof is valid

(note, however, that Selden and Selden did not empirically study the behavior of mathematicians). Similar assertions about mathematicians' uniform agreement have been made by philosophers of mathematics when debating the nature of proof (e.g., Azzouni, 2004; Rav, 2007), and in popular expository mathematics texts (e.g., Singh, 1997). However, when Weber (2008) empirically studied the processes by which mathematicians validate proofs, he found that there was not uniform agreement between his participants' final judgments, even on the relatively simple proofs used by Selden and Selden. Of those four proofs, Weber's research-active participants had uniform agreement for only two; for the other two there were minorities (1 of 8 and 2 of 8) who disagreed with the verdicts of the majority. Given the small sample size in Weber's study, it may be unreasonable to draw firm conclusions from this finding, but his results certainly suggest that the assertions made by Selden and Selden, Rav, and Azzouni require empirical support.

In this article, our first aim is to contribute to these discussions by reporting further empirical evidence regarding the uniformity of agreement about the validity of purported mathematical proofs.

Accounts for Students' Difficulties With Proof Validation

Although it is unclear whether mathematicians exhibit uniform agreement about the validity of purported proofs, it is clear that validation is challenging for students. Several reasons have been proposed to account for students' difficulties when validating. One possible account is that students might have insufficient understanding of what constitutes a proof. Many researchers have found that students sometimes regard the provision of examples or visual images as valid proofs (e.g., Coe & Ruthven, 1994; Harel & Sowder, 1998), or believe that proofs must be structured in a particular format (e.g., geometry proofs must be written in two columns, Martin &

Harel, 1989). We note, however, that these conclusions have typically been drawn, for university students at least, from research based on proof construction activities rather than proof evaluation activities. For various methodological reasons, such studies may not give an accurate picture of students' conceptions of proof (Stylianides & Stylianides, 2009a; Vinner, 1997; Weber, 2010). Indeed, when Weber (2010) studied the argument evaluation behavior of successful undergraduate students, he found that few participants believed empirical arguments could be valid proofs, but that many judged invalid deductive arguments to be valid proofs. He concluded that high-achieving students' difficulties with proof validation were primarily related to their skill at validating deductive arguments, not to the misconception that nondeductive arguments might constitute valid proofs.

Another possible account for why students find proof validation difficult is related to the aspects of proofs to which they devote most attention. Selden and Selden (2003) found that, rather than attending to the underlying logical structures of purported proofs, students concentrated on what Selden and Selden called *surface features*, such as algebraic notation and computations. Similarly, Healy and Hoyles (2000), in their study of students' perceptions of proof, found that algebraic arguments—even when mathematically nonsensical—were regarded by school students as likely to receive the best mark from the teacher. This suggests that students' difficulties with proof validation may be associated with a tendency to overvalue algebraic manipulations and thus to allocate too little attention to the logical relationships between the various components of the proof. This interpretation—in which students may be distracted from logic by symbolism—would also be consistent with Österholm's (2005) finding that undergraduate students better understood an introduction to group theory when this was written in words rather than in words and symbols.

The second aim of the current study was to directly determine the aspects of purported proofs to which undergraduate students and mathematicians attend when validating. If the preceding account is correct, we would expect that, during validation attempts, there would be a between-groups difference in the proportion of time spent attending to formulae and to text.

Analyses of Proof Validation

The preceding suggestions might account for errors in students' validation judgments, but researchers also would prefer to address what it takes to successfully validate a proof. Although there have been only limited empirical studies involving expert proof validators (e.g., Weber, 2008), and none directly comparing expert and novice validators, several researchers have given theoretical analyses of what is required to validate successfully. Here, we structure our review of these accounts by adopting Weber and Mejia-Ramos's (2011) terminology.

Weber and Mejia-Ramos (2011) suggested, based on introspective observations by practicing mathematicians, that there are two broad strategies that can be adopted when reading proofs: zooming in and zooming out.

Zooming in. Given the impractical length of strict formal derivations, mathematical proofs necessarily contain logical gaps (Fallis, 2003). Rav (1999) suggested that one of the primary responsibilities of a reader is to fill in these gaps by constructing subarguments. It is this process—of filling in gaps between successive statements in a proof—that Weber and Mejia-Ramos (2011) referred to as the zooming-in strategy. Others have called this method a *line-by-line* strategy (Selden & Selden, 2003).

In Rav's (1999) conceptualization, a proof consists of a series of statements A_1, A_2, . . . A_n, B, within which B is the to-be-proved theorem. Although some statements may be logically independent of preceding statements (they may, for example, introduce new notation or

definitions), most steps $A_i \rightarrow A_{i+1}$ require what Toulmin (1958) called a *warrant*: a justification that allows the reader to conclude that A_{i+1} follows from some subset of A_i, A_{i-1}, . . ., A_1, and known axioms, definitions, or theorems. Weber and Alcock (2005) pointed out that the warrants in mathematical proofs are sometimes implicit: a warrant might not be explicitly written, especially if the author thinks that it is easy to infer.

When zooming in to consider a proof line by line, the reader is faced with a triple task: he or she must (a) accurately determine when a warrant is required, (b) correctly infer the implicit warrant intended by the author, and (c) evaluate the warrant's mathematical validity. How step (b) in this process is conducted has been the subject of disagreement: Rav (1999) suggested that when mathematicians cannot immediately infer an implicit warrant for step $A_i \rightarrow A_{i+1}$ they construct a deductive subproof (essentially they decompose the step into a series of substeps). But in an empirical study of mathematicians' proof validation behavior, Weber (2008) found that on some occasions this did not happen. Instead, he observed mathematicians bridging such gaps using nondeductive methods such as evaluating examples and constructing informal nondeductive arguments (cf. the analogous argument Inglis, Mejia-Ramos, & Simpson, 2007 made in the context of proof construction). Interestingly, this finding suggests that mathematicians do not always attempt step (b) in the three-stage process outlined previously; apparently it sometimes suffices to provide a gap-bridging warrant that is clearly different from that intended by the purported proof's author. Presumably, this alternative strategy can only be adopted when the reader believes that the step is valid (if a proof is to be rejected because an implicit warrant is invalid, it would seem important that the reader has inferred the warrant that the author intended).

Weber and Alcock (2005) proposed that the complexity of this three-step process (determine, infer, evaluate) is one major reason why students find proof validation so difficult; they did not, however, speculate as to which of the three stages causes students to struggle. In this article we will use eye-movement data to address this question.

Zooming out. Weber and Mejia-Ramos (2011) argued that mathematicians do not read proofs only by zooming in. They also identified an alternative, or perhaps supplementary, strategy that they called *zooming out*. When a mathematician zooms out they do not focus on the logical detail of the argument, but on what Rav (1999) called *methodological moves*: encapsulated strings of logical derivations that together form coherent chunks of the whole argument. Weber and Mejía-Ramos suggested that it is possible to validate a proof by decomposing it into methodological moves and evaluating whether these moves fit together to imply the theorem.

Along with introspective comments from the mathematicians they interviewed, various sources of evidence support Weber and Mejia-Ramos's (2011) suggestion that mathematicians sometimes use zooming out as part of a validation attempt. For example, written proofs are often structured to include cues intended to help the reader partition the argument into methodological moves (Konior, 1993), and several philosophers and mathematicians have asserted that proofs are validated by reference to their overall plausibility rather than their detailed logical coherence (Hanna, 1991; Manin, 1977; Thurston, 1994).

One particular point of interest for this article is that when asked to introspect on their proof validation behavior, some of the mathematicians interviewed by Weber (2008) reported that they would adopt a zooming-out strategy specifically at the start of an attempted proof validation. Participants suggested that adopting such a strategy allowed them to gain an overview

of the structure of the proof, before moving on to a zooming-in strategy. For example, when asked how they validate proofs, one mathematician said that he "just read through the proof to just to get the big picture, just to get the feel, to get the flow of it and then go back and get the details" (p. 441). Another suggested that he or she would "first try to understand the structure of the proof, to get an overview of the argument that's being used" (p. 441). However, despite several of the participants in the study making similar claims, there were no examples of participants actually doing this in the validation component of Weber's verbal protocol data (Weber, personal communication). To date, then, the proposal that mathematicians sometimes adopt a zooming-out strategy is supported only by introspective reports, not by empirical observation. Our results allow us to speak to this question, as described subsequently.

Attention location when zooming in and zooming out. If Weber and Mejia-Ramos (2011) are correct that mathematicians use two distinct validation strategies, this would seem to imply that there should be two distinct ways in which readers deploy attention when successfully validating. One could characterize the zooming in strategy by suggesting that it would proceed in an approximately linear fashion, with the reader's attention moving sequentially from line to line. In contrast, a characterization of the zooming-out strategy would involve a substantially more nonsequential reading order: One would expect the reader's attention to regularly move between different sections of the proof as they attempt to link together its various methodological moves. In particular, if mathematicians do begin by zooming out, one would expect their attention to move across the whole proof rapidly at the beginning of a validation. Although these descriptions are caricatures (we would expect genuine data to be substantially less tidy than these descriptions suggest), we nevertheless believe that the two strategies are sufficiently distinct to make different predictions about the frequency of different types of eye movements during proof

validation: We would expect a zooming-in strategy to involve large numbers of shifts of attention between consecutive lines, but few shifts of longer distances, whereas a zooming-out strategy would involve large numbers of shifts of attention of distance greater than one line.

The third aim of the current study was to investigate (a) whether there is empirical support for the existence of two distinct strategies for proof validation; and (b) whether there are differences in the extent or sophistication of the use of these strategies by mathematicians and undergraduate students.

Methodological Issues With Studying Proof Validation

As discussed in the preceding sections, studies investigating proof validation processes typically have relied upon either introspective reports from practicing mathematicians (Weber & Mejia-Ramos, 2011), or verbal-protocol methods in which students and mathematicians have been asked to "think aloud" while validating proofs (Selden & Selden, 2003; Weber, 2008). Each of these methods has substantial drawbacks, especially given Selden and Selden's suggestion that a participant's proof validation may be "difficult to observe because not of all it is conscious" (p. 5). Clearly it is not possible to verbalize nonconscious activity either retrospectively (via an introspective report) or in the moment (via a verbal protocol).

More generally, theorists have proposed two possible drawbacks with verbal protocol methods such as those adopted by Selden and Selden (2003) and Weber (2008). First, asking participants to verbalize their thoughts during a cognitively demanding task may alter their behavior. This threat to validity has been called reactivity (Russo, Johnson, & Stephens, 1989). The concern is that by requiring participants to verbalize their thoughts as they engage in a proof validation, the researcher may reduce the level of cognitive resource that participants can devote to that primary task. This appears to be particularly harmful to the validity of research results

when the primary task involves important nonconscious activity (such as insight problems, e.g., Schooler, Ohlsson, & Brooks, 1993). Intriguingly, however, asking participants to verbalize their thoughts might actually increase the efficacy of their behavior on some tasks, perhaps including reading. Chi, Bassok, Lewis, Reimann, and Glaser (1989) found that those participants who were successful at comprehending explanations given in scientific textbooks were more likely to have "self-explained" (generated explanations to themselves) than those who were less successful. It is thus plausible to suppose that asking participants to validate proofs while thinking aloud might increase the likelihood that they generate self-explanations and thus increase their efficacy. Regardless of whether the verbal protocol method increases or decreases the efficacy of participants' proof validations, it seems highly plausible that the data so generated may not reflect with total accuracy the ways in which participants validate proofs when not thinking aloud.

The second threat to the validity of verbal protocol methods is known as veridicality. This relates to the accuracy of the self-reports offered by participants, and it also applies to retrospective accounts. The data generated from verbal protocol or retrospective accounts may be nonveridical for several reasons. Participants may fail to report crucial components of their strategy (a particular problem if proof validation involves nonconscious thought), or they may report mental events that did not, in fact, occur (Ericsson & Simon, 1984; Nisbett & Wilson, 1977). Russo et al. (1989) found evidence for both these types of nonveridicality when they asked participants to solve a variety of cognitively demanding tasks under different verbal-reporting conditions.

The conclusion from this set of findings is that it is dangerous to rely solely upon self-report methods to understand the processes by which cognitively demanding tasks are

performed. Thus, if a sophisticated understanding of proof validation is to be developed, the hypotheses and theoretical accounts generated by self-report methods need to be checked and investigated using alternative methods.

One alternative way of studying the processes by which participants engage in effortful tasks is by recording their eye movements as they perform the activity. The theoretical justification for this method is the eye-mind hypothesis (Just & Carpenter, 1980; Rayner, 1998, 2009), which states that there is a close relationship between gaze direction and attention location. Although this relationship is clearly not exact (one can, for example, gaze out of the window while pondering how best to structure the next section of one's paper), it is believed to be strongly positive, and especially so in effortful tasks such as reading or reasoning (e.g., Ball, Lucas, Miles, & Gale, 2003; Deubel & Schneider, 1996; Rayner, 1998, 2009).

When people interact with stationary visual information, their eye movements consist of a series of fixations (stationary periods during which information from the environment is processed, which typically last around 150 ms to 500 ms). Longer individual fixations are believed to be associated with a greater processing effort for the fixated-upon area,[1] and more time is spent fixating on areas regarded as task-relevant than those regarded as task-irrelevant

[1] One of the clearest demonstrations of this came from a study conducted by Gould (1973). Gould asked participants to commit a list of letters to memory, and then search a visual display of target letters (systematically in order) with the task of determining whether each target letter was on the memorized list. On some trials, participants were asked to memorize a longer list of letters than on others. Gould found that the length of participants' fixations on target letters was proportionate to the length of the list that they had been asked to memorize. He concluded that participants tackled the task by sequentially comparing each target with every item on their memorized list: When the memorized list was longer, more processing steps were required to complete the comparisons, so participants' fixations were longer compared to trials with shorter lists.

(e.g., Gould, 1973; Kaakinen, Hyönä, & Keenan, 2002; Liversedge, Paterson, & Pickering, 1998). After a fixation, the eyes move to a new location via an extremely rapid movement known as a saccade, during which no information can be processed (e.g., Matin, 1974).

The eye-mind hypothesis is the basis of much research. Eye movements have been used extensively to study participants' behavior in tasks such as visual search (e.g., Watson & Inglis, 2007), numerical cognition (Merkley & Ansari, 2010; Moeller, Fischer, Nuerk, & Willmes, 2009), and logical reasoning (e.g., Ball et al., 2003). In the context of educational research, eye movements have been used widely to investigate how learners engage with multimedia learning environments (see the discussion of the use of eye-tracking to study multimedia learning in Hyönä, 2010). Perhaps the clearest example of the utility of eye-movement data for education comes from the domain of reading. There is now a large body of research that has used eye-movement analyses to investigate the processes involved in skilled reading (for reviews, see Rayner, 1998, 2009). For example, it is known that skilled readers activate phonological codes very early after first fixating on a word, that is, that the "sounding out" of a word begins very rapidly after it is first fixated upon (Pollatsek, Lesch, Morris, & Rayner, 1992). This finding led researchers to develop instruction that explicitly focused on the relationship between graphemes (printed letters) and phonemes (their associated sounds). Such research has fed into the ongoing debate between those who advocate using phonics for the teaching of English in early-years education and those who advocate whole language methods (for a review, see Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001).

Eye movements can be recorded via hidden cameras embedded in a computer screen that looks otherwise normal, so they provide a nonintrusive, real-time measure of where participants allocate their visual attention. This means that eye-movement analyses do not suffer to the same

extent as retrospective reporting or verbal protocol methods in terms of reactivity or nonveridicality. However, the method is not without drawbacks. In particular, although a participant's fixation location gives a good indication of that to which they were attending, it does not provide information about the reasons why they were attending to that piece of information or about the success or failure of their attempt to process the information. We thus believe that the primary value of eye-movement analyses for studying proof validation behavior is in the way they allow us to (a) examine and test hypotheses generated from previous verbal protocol studies; and (b) compare the ways in which different validators allocate their attention.

Summary of Research Questions

We have discussed three main aspects of the literature that pertain to proof validation: the extent to which mathematicians and students agree on the validity of purported proofs, possible reasons why students may find proof validation to be difficult, and different strategies that can be used to successfully validate. We have also discussed methodological issues related to studying the processes of proof validation. Our aim here is to contribute to this literature by reporting the first direct comparison of the proof validation behavior of relative experts (research-active mathematicians) and relative novices (beginning undergraduate students). Specifically, we address three main questions:

1. Do research mathematicians typically agree on the validity of purported proofs as suggested by Selden and Selden (2003) and Azzouni (2004), or do they sometimes disagree, as suggested by the results of Weber's (2008) study?
2. Do experts and novices attend to different parts of purported proofs to different degrees? For example, do novices spend proportionately longer focusing on surface features as hypothesized by Selden and Selden (2003)?

3. Is there any evidence for two distinct strategies for proof validation—zooming in and zooming out—as proposed by Weber and Mejia-Ramos (2011)? Are there expert/novice differences in terms of the frequency with which these strategies are used or the sophistication of their use?

METHOD

Participants

Participants were 18 first-year undergraduate students (9 male) studying either single- or joint-honors mathematics, and 12 academic mathematicians (10 male), all from Loughborough University. Each was paid £8 (approx $13) for participating. The undergraduates had been mathematically successful in their school-level education.[2] At university, they had all completed two semesters of study of proof-based calculus and linear algebra; those studying single-honors mathematics had also studied other mathematics modules (including proof-based modules focused on geometry, differential equations, and number theory). The teaching style on all these mathematics modules was traditional, consisting of two or three large-group lectures and sometimes one smaller-group problem-solving class per week (cf. Weber, 2004). The School of Mathematics at Loughborough is a highly-ranked research-intensive department;[3] consequently, as well as teaching undergraduate- and/or postgraduate-level mathematics courses, the mathematicians in the study were all active researchers.

---

[2] All had met the university's entry requirements of AAB grades at A-Level (18-year-old school leavers in England have typically studied three subjects at A-Level, each of which is graded on a scale from A to F. Thus, the highest possible grade profile is AAA).

[3] Over 50% of the department's research was rated as "internationally excellent" or "world leading" in the 2008 Research Assessment Exercise, the government-administered assessment of UK universities' research quality.

Stimuli, Procedure, and Apparatus

The study took place in a quiet room. Testing was conducted individually. Eye-movements were recorded with a Tobii T120 Eye-Tracker (Tobii Technology, Stockholm, Sweden), which was set to sample at 60Hz. This is a remote eye-tracker that consists of two hidden binocular infrared cameras underneath a 17" TFT monitor. Stimuli are displayed on a screen that participants view (without head restriction) from approximately 60cm away. The accuracy of eye-position tracking is typically 0.5°. For each participant, prior to the start of the study the eye-tracker was calibrated with a 9-point display. This setup is typical for eye-movement studies (e.g., Jepma & Nieuwenhuis, 2011; Merkley & Ansari, 2010).

The purpose of the study was explained to participants, both verbally by the experimenter and onscreen, and the procedure involved three phases. In the first phase, participants were told that they would be shown four student-generated proofs of the same theorem, and that their task was to read each proof and decide whether it was valid. Each proof was displayed onscreen, and participants were encouraged to take as long as they needed to make an informed judgment about its validity. The proof stayed onscreen until the participant indicated that they had made their judgment by pressing the mouse button, at which point the proof was removed from display. On the next screen, participants selected "valid" or "invalid," clicked submit, and were taken to a third screen where they were asked to say how certain they were that their response was correct. This judgment was submitted via a five-point Likert scale from 0 ("It was a total guess") to 4 ("I am completely certain"). After participants had submitted their response and taken a break if desired, the next proof was displayed. The first phase lasted until participants had read and responded to the four proofs.

The second phase was similar to the first, except that participants were asked to read two short nonmathematical passages (taken from newspapers). The purpose of this section was to give participants a break between reading the mathematical proofs.

In the third phase, participants were told that they would be presented with two proofs that had been submitted to a popular mathematics journal such as *The Mathematics Gazette*. As in the first phase, their task was to determine whether the proofs were valid. Responses were submitted following the same procedure as that used in the first phase.

Prior to the presentation of the first proof, full instructions were given onscreen, and a practice proof was displayed to familiarize participants with the procedure. After the practice proof, the researcher asked the participants to carry on with the study at their own pace and left the room. Participants were told that if they wanted to talk aloud as they were reading they should feel free to do so, but that this was not required. Most did not. However, when submitting their decisions about the proofs, several of the mathematicians verbalized their reasons for giving an "invalid" rating (justifying a "valid" rating was rarer). Audio responses were recorded with an external microphone.

The proofs from the first phase (Proofs 1–4) were identical to those used by Selden and Selden (2003) and Weber (2008), and are given in the Appendix. Proof 6 was also used by Weber, but Proof 5 was novel (loosely based on the argument given by Sangwin, 2010); these proofs also appear in the Appendix. Although some of the proofs have clear mistakes (Proofs 1, 4, and 6), the validity/invalidity of the others is contestable (cf. Weber, 2008). Consequently we did not give feedback to participants, and neither will we discuss the number of "correct" responses they gave. Instead, our primary focus in the results section that follows is on

understanding the differences between the approaches adopted by experts and novices when reading for validation.

RESULTS AND DISCUSSION

The eye-movement data were analyzed with Tobii Studio 2.1, configured to use the Tobii Fixation Filter (Tobii Technology, 2010). Prior to conducting the main analysis, we compared the overall reading times and responses of the two groups. There was no significant difference between the mean total reading times (across all six proofs) of the mathematicians and the undergraduates, $t(28) = 1.40$, $p = .17$, suggesting that there was no systematic difference in the time it took the different groups to validate the proofs.[4]

We structure our discussion of the results in three main sections, corresponding to our three research questions. First, we discuss the perceived validity of the purported proofs. Second, we investigate between-group differences with respect to the balance of time spent studying algebraic symbols and text. Finally, we discuss the extent to which participants' behavior fits Weber and Mejia-Ramos's (2011) discussion of zooming in and zooming out, and we evaluate between-group differences in validation strategies.

Validity Judgments

After reading each purported proof, participants were asked to state whether they believed it was valid or invalid, and to rate how confident they were in their judgment. Response frequencies for the two groups are shown in Table 1.

There was a significant difference between the two groups' responses to Proof 1, $p = .004$, Proof 4, $p = .001$ and Proof 6, $p = .010$ (all Fisher's Exact Tests). These were the proofs

[4] Prior to conducting the parametric statistical tests reported in the article, the data were assessed to determine whether they met the required assumptions of normality, homogeneity of variance, and sphericity.

with clear mistakes; the mathematicians (reliably) judged these proofs invalid, but the undergraduates did not. Group differences did not approach significance for any other argument. In particular, there were large disagreements among the mathematicians about the validity of Proofs 2, 3, and 5 (see Table 1). These results are consistent with Weber's (2008) finding that mathematicians do not exhibit uniform agreement about the validity of proofs. To explore this lack of agreement on Proofs 2, 3, and 5 in further detail we conducted an analysis of the verbal explanations given by some of the participants when submitting their responses.

Some of the reasons mathematicians gave for rating proofs as invalid were related to style. For example, regarding Proof 2, one mathematician justified his decision by complaining about the coherence of the proof: "I think it's invalid because um, they had most of the ingredients there for a proof, but they just haven't put them all in the right order and didn't really explain what they were doing." However, some of the reasons involved sophisticated mathematical objections. For example, when discussing her response to Proof 5 (which six mathematicians rated as valid and six as invalid) one mathematician remarked:

> I don't know how they're defining *e* to the *y*. I guess . . .[5] they must be defining *e* to the *y* by this limit, 1 plus *y* over *n* to the *n*. And then, how do they define log? The way that you define . . . usually, usually you start by defining log as this integral. And then you use that to define the exponential function. And then you can prove these limits are true [. . .] and so, doing it, kind of the other way around, I'm . . . um . . . I bet it works, but . . . um, it's not clear what they're taking as the definitions and how they're proving the various things.

[5] In these quotations, an ellipsis (. . .) indicates a pause of more than one second by the speaker, whereas a bracketed ellipsis [. . .] indicates an omission of some of the speaker's words.

Another questioned whether it was reasonable to leave the assumptions made in Proof 5 to be inferred by the reader:

> It's that little line, "By the properties of *e* . . ." that got me thinking. Are you assuming a Taylor expansion? You're assuming a lot of calculus in there. That's where, that's the bit that worries me about it. [. . .] So I . . . it's probably valid, but there's a lot of assumptions sort of hidden under the hood there.

There are two reasonable hypotheses that can account for the mathematicians failing to agree about the validity of the proofs (beyond disagreements on stylistic issues). It may be that the disagreements were the result of performance errors: Some mathematicians may have spotted problems with the proofs that others missed. If this were the case, we might expect that if a mathematician who rated Proof 5 as valid discussed the issue with one who rated it as invalid, they would quickly reach a consensus. Another possibility is that these disagreements are the result of different normative standards of validity: Perhaps there is not as uniform a standard of validity among mathematicians as has been assumed by some philosophers and education researchers (e.g., Azzouni, 2004; Rav, 2007; Selden & Selden, 2003). Instead, it may be that the validity of a proof depends upon both the reader and the social context in which the proof is read and that this has implications for the amount of background and detail that is required for an argument to be perceived as valid (e.g., Manin, 1977; Thurston, 1994; Weber, 2008). This latter account would have significant implications for educational theory. For example, the influential proof-schemes framework (Harel & Sowder, 1998, 2007) defines the goal of instruction with respect to proof as the development of students' proof schemes so that they match those shared by the community of mathematicians. This second account raises the possibility that there is no such shared scheme.

Our primary aim was to study validation processes, so we, like Weber (2008), are vulnerable to the criticism that our sample size is somewhat smaller than would be expected for a study designed to investigate the outcomes of mathematicians' validation attempts. For further discussion regarding these issues, including a report of a study with a substantially larger sample, see Inglis, Mejía-Ramos, Weber, and Alcock (in press).

Dwell Times on Different Parts of the Proofs

Selden and Selden (2003) hypothesized that students focus on surface features of proofs, rather than engaging with their logical structure. In particular, they suggested that students focus mainly on notational and computational aspects of proofs. To investigate this suggestion, we calculated each participant's total dwell time on the "formulae" in each proof (i.e., the total duration of all the fixations on "formulae"). We categorized as *formulae* those parts that had been typeset using LaTeX's mathematics mode (see Lamport, 1994). These data, shown in Figure 1, were subjected to a 2 × 2 Analysis of Variance (ANOVA) with one within-subjects factor (type: formulae or nonformulae) and one between-subjects factor (status: mathematician or undergraduate). There was a significant type × status interaction, $F(1,28) = 8.81$, $p = .006$, $\eta_p^2 = .239$; the students spent proportionately longer fixating on the formulae than did the mathematicians (55% of the students' time was spent dwelling on formulae compared to 45% of the mathematicians' time, $t(13.0) = 2.74$, $p = .017$).

One possible way of accounting for this finding would be to suppose that the undergraduates, more than the mathematicians, found it hard to process the formulae, and so needed to devote more attention to these areas. Support for this proposal comes from an analysis of the mean fixation durations of the two groups. As well as spending proportionately longer than the mathematicians fixating on the formulae in the proofs, the undergraduates had longer

mean fixation durations on these areas, 433 ms versus 343 ms, $t(28) = 4.08$, $p < .001$, $d = 1.54$. Undergraduates' greater fixation duration may suggest a greater cognitive effort on processing the information contained in the formulae than that exhibited by the mathematicians (cf. Just & Carpenter, 1976, 1980). However, if processing difficulty were the primary reason for the observed type × status interaction, we might expect that the undergraduates would have spent longer in total than the mathematicians inspecting the formulae. In fact, although the students spent proportionately longer dwelling on formulae than did the mathematicians (55% versus 45%, as above), the two groups had roughly similar *absolute* dwell times on these areas (370 secs versus 349 secs): The interaction effect was driven by the mathematicians spending longer dwelling on the parts of the proof that were not formulae.

An alternative hypothesis, which accounts for both of these findings, is that the undergraduate students did not spend as long as the mathematicians studying the logic of the arguments; this is plausible because the parts of the proofs that explicate the logical relationships are primarily contained in the text rather than the formulae. We discuss further evidence for this possibility later in the article. (Of course, it is possible that some combination of these two factors may account for the type × status interaction we observed.)

Initial Reading Strategy

To investigate whether mathematicians typically zoom out at the start of a validation attempt, as claimed by participants in Weber's (2008) study, we calculated the time at which participants first fixated on each line of each proof. Expressing this as a percentage of their overall time spent validating that proof allowed us to establish whether they conducted a first, fast read-through before considering the details. The grand means of these first-fixation data are shown in Figure 2. Both mathematicians and undergraduates appeared to show a similar pattern:

the mean proportion of time elapsed prior to the first fixation on proofs' last lines was 50% for mathematicians and 66% for undergraduates. Although this difference was significant, $t(28) = 3.34$, $p = .002$, we note that 50% is nonetheless a higher figure than would be expected with an initial zooming-out strategy. We thus attribute this difference to the mathematicians in the study being quicker than the students at pursuing a zooming-in strategy, rather than to a substantial strategy difference.

This finding contrasts with the claims made by Weber's (2008) participants. It may be that they were referring to proofs that clearly have a global structure that cannot be checked easily using a zooming-in strategy, for example, proofs by contradiction or by cases. To assess this possibility, we repeated the analysis described previously considering only Proof 2, which was a contradiction proof that also had two separate cases (see Appendix). An essentially identical pattern of results emerged: The mean proportion of time elapsed prior to the first fixation on the last line of the proof was 50% for mathematicians and 67% for undergraduates. Thus, we do not believe that these data are consistent with the claims made by the participants in Weber's (2008) study. If either mathematicians or undergraduates had initially zoomed out, we would expect to see them fixate on the proof's last line well before 50% of their total reading time had elapsed.

We see three reasonable ways of accounting for the discrepancies between the claims made by Weber's (2008) participants and the findings of this study. First, it is possible that Weber's participants were conflating proof validation and proof comprehension (cf. Mejía-Ramos & Inglis, 2009). Although we believe that validating a proof might be one way in which mathematicians read for comprehension, it is at least possible that differences in readers' goals lead to differences in their behavior (whether this is the case is an issue that future research could

productively investigate). Perhaps, then, Weber's participants were describing their proof comprehension behavior rather than their proof validation behavior.

Another possibility is that the proofs we used during this study were too simple to require any complex behavior. The proofs we used (like those used by Weber, 2008, and Selden and Selden, 2003) were relatively short, more similar to proofs typically encountered by undergraduates during their courses than to those found in research-level mathematical texts (this was by necessity, given our expert/novice design). It could be that mathematicians read the entire proof as a single methodological move, and thus no zooming out was required. Consequently, we are only able to draw robust conclusions about proofs of the type typically encountered by undergraduates. Whether zooming out is a strategy adopted by research mathematicians when reading research-level mathematics remains an issue for future research.

A third possibility that accounts for the difference between the findings of this study and the claims made by Weber's (2008) participants is that the mathematicians interviewed by Weber were simply wrong. Many research studies have demonstrated that experts (in various domains) are extremely poor at reflecting on the nature of their expertise (Ericsson, 2006; Hoffman, 1992; Nisbett & Wilson, 1977). The classic example given in the literature is that of chicken-sexers: professionals who are able to reliably determine the sex of day-old chicks using subtle perceptual clues. Despite having extremely high accuracy rates (> 98%), such experts are apparently unable to explain how they make their decisions (e.g., Horsey, 2002). Similar claims have been made about chess players, wine tasters, and even doctors making medical diagnoses. In all these cases, experts make reliable decisions without reliably knowing how. As discussed previously, of particular concern with asking participants to retrospectively describe their behavior during a particular activity is not just that they may not know, but that they may offer

invalid post-hoc descriptions which do not match their actual behavior (e.g., Nisbett & Wilson, 1977). For example, in the informal debrief after our experiment, one of the mathematicians endorsed the suggestion made by Weber's (2008) participants, claiming that she would typically scan through a proof prior to reading it line-by-line. This introspective remark was not consistent with her eye movements. Finally, this third hypothesis is also consistent with earlier observations of a gap between introspections about expert mathematicians' behavior and their actual behavior (Inglis & Mejia-Ramos, 2009; Weber, 2008).

General Reading Strategies: Zooming in and Zooming Out Revisited

Although it appears that neither mathematicians nor undergraduates validate proofs by initially adopting a zooming-out strategy, it may be that they do use this strategy to some extent during the course of their proof validation attempts. As discussed in the introduction, the zooming-in and zooming-out strategies result in different paths of attention: Adopting a zooming-in strategy would result in an essentially sequential reading order, whereas a zooming-out strategy would lead to a substantially more nonsequential reading order.

To determine whether mathematicians and undergraduates read the proofs in different orders, we calculated line-transition matrices for each validation attempt. To do this we identified every saccade that started and ended with a fixation on a line of the proof (as opposed to offscreen or in white space), then counted the number of saccades that resulted in a transition to a new line and the number that stayed within a single line; together, these two types constitute all the within-proof saccades.

The mean number of between-line saccades was significantly higher for mathematicians, at 78.8 per proof, than it was for undergraduates, at 53.3 per proof, $t(28) = 2.11$, $p = .044$, $d = 0.80$. As mathematicians had a slightly (but nonsignificantly) higher number of within-proof

saccades in the study overall (an average of 387 per proof compared to 351), we also compared the two groups' frequencies of between-line saccades as a proportion of the total number of within-proof saccades. The same pattern emerged: Mathematicians had a significantly higher mean proportion of between-line saccades, 23%, compared to 18% for undergraduates, $t(28) = 2.18$, $p = .038$, $d = 0.82$.

To explore the nature of these extra between-line saccades, we calculated the number of saccades of distances 1, 2, and 3 or more (a saccade between line n and line n+1 has distance 1, whereas a saccade between line n and line n–3 has distance 3). These data are shown in Figure 3, and were subjected to a 3 × 2 ANOVA with one within-subjects factor (saccade-distance: 1, 2, 3 or more) and one between-subjects factor (status: mathematician, undergraduate). There was a significant saccade-distance × status interaction, $F(1.02,28.7) = 5.22$, $p = .029$, $\eta_p^2 = .157$ (the data failed to meet the assumption of sphericity, so the Greenhouse-Geisser correction was applied). This reflected the fact that the mathematicians made significantly more saccades between consecutive lines than did the undergraduates, $t(28) = 2.24$, $p = .033$, $d = 0.85$, but that the between-groups differences for the other distances considered did not approach significance, all ps > .2.

In summary, the mathematicians in our study seemed to move their attention between lines considerably more than the undergraduates: They made approximately 50% more between-line saccades. This difference seemed to occur primarily because the mathematicians moved back and forth between consecutive lines more than the undergraduates, not because they moved their attention around the proof at a more global level.

One hypothesis that might account for these additional consecutive between-line saccades is that, following Weber and Alcock (2005) and Weber and Mejia-Ramos (2011), the

mathematicians were more frequently seeking to infer implicit between-line warrants, of the type associated with the zooming-in validation strategy. To test this hypothesis, we conducted an in-depth analysis of the participants' reading behavior on Proof 6 (shown in Figure 4), chosen because of its length and because of the large between-groups difference in its perceived validity. This proof, shown in Figure 4, which 55% of the undergraduates and 100% of the mathematicians categorized as invalid, contained nine lines. The logic of the proof flows correctly, except that the statement on Line 5 ("Every number that leaves remainder 1 when divided by 4 is divisible by a prime that also leaves remainder 1 when divided by 4") is false (9 is a counterexample).

To test the hypothesis that mathematicians devote more effort than undergraduates to searching for between-line warrants, we categorized each line-to-line transition in Proof 6 as either requiring a warrant or not requiring a warrant. For example, the transition from Line 2 to Line 3 does not require a warrant, because Line 3 consists of the statement of a definition and thus could be introduced without justification at any point. In contrast, the transition from Line 1 to Line 2 does require a warrant, as Line 2 is a conclusion that follows from the supposition made in Line 1.[6] Notice that the warrant that justifies Line 2 (finitely many things can be written in list form) is not explicitly stated in the proof, and therefore must be inferred and evaluated if a successful validation is to occur.

---

[6] The transitions between Lines 1 and 2, 3 and 4, 4 and 5, 6 and 7, 7 and 8, and 8 and 9 were categorized as requiring warrants; the transitions between Lines 2 and 3, and 5 and 6, were categorized as not requiring warrants. Note that this classification is one of degree: Clearly every line in every proof requires a warrant to some extent. For example, both lines 3 and 6 implicitly appeal to the definition of the p_is, and to definitions of various mathematical terms and symbols.

If it is true that the between-groups difference in the number of between-line saccades of distance 1 was due to the mathematicians' greater propensity to seek warrants, we would expect to see a greater between-groups difference in warrant-seeking behavior for those lines that require a warrant compared to those lines that do not. Before directly testing this prediction, however, we first report the general behavior of the two groups when validating Proof 6.

Proof 6, and those between-line saccades that occurred with mean frequency greater than 1 are shown in Figure 4. We make two comments on this Figure. First, in line with the analysis reported previously, both mathematicians and undergraduates made the vast majority of their between-line saccades in Proof 6 between consecutive lines. Indeed, only three non-consecutive between-line saccades occurred with a mean frequency greater than 1. Second, although mathematicians and undergraduates made the same types of between-line saccades, overall the mathematicians made substantially more (indicated by the size of the arrow heads).

To test our prediction we counted the number of times each participant visited Line x followed by Line x–1 followed by Line x.[7] We believe that such a sequence of eye movements may be associated with an attempt to connect Line x with Line x–1, in other words, to find the warrant that links them. Of course, there may be other reasons for such a sequence of eye movements, for example, general search behavior. However, these other reasons would only add noise to the data analyzed in this section.

We calculated the number of warrant-seeking $x \rightarrow x\text{–}1 \rightarrow x$ sequences made by each participant for the line-to-line transitions in Proof 6 that had been categorized as requiring

[7] For this analysis we used collapsed sequences of fixations, because we were interested in how participants' loci of attention moved between lines rather than the speed at which this process took place. For example, the two sequences of fixations XXXYYYYYXX and XYX would both have been classified as X→Y→X sequences.

warrants and for those categorized as not requiring warrants. These data are shown in Figure 5 and were subjected to a 2 × 2 ANOVA with one within-subjects factor (transition-type: warrant-requiring, non-warrant-requiring) and one between-subjects factor (status: mathematician, undergraduate). In line with our prediction, there was a significant transition-type × status interaction, $F(1, 27) = 4.38$, $p = .046$, $\eta_p^2 = .140$. On average the mathematicians made 14.9 more warrant-seeking sequences when a warrant was required than when one was not, whereas the equivalent figure for the undergraduates was 5.0.

There are several limitations to this analysis. For example, we considered only searches for warrants between neighboring lines. Many proofs, including Proof 6, require the coordination of several lines to produce valid warrants for later lines, and some proofs have statements within lines that require an implicit warrant to be found and evaluated. Notwithstanding these limitations, these data are consistent with our prediction based on the hypothesis that one of the reasons for the greater number of between-line saccades made by the mathematicians is that they were making more eye-movement sequences designed to infer implicit warrants to link consecutive lines in the proof.

There are two possible accounts of this finding. One possibility is that the mathematicians were better than the undergraduates at recognizing when a warrant is required. An alternative possibility takes account of the fact that both groups made significantly more warrant-seeking moves for those lines that required warrants compared to those that did not, undergraduates: $t(17) = 2.47$, $p = .025$, $d = 0.51$; mathematicians: $t(10) = 2.92$, $p = .015$, $d = 0.68$. Perhaps both the mathematicians and the undergraduates recognized when a warrant was required but the undergraduates did not devote as much effort to searching for these warrants. We consider pedagogical suggestions that follow from this latter account in the discussion that follows.

SUMMARY AND CONCLUSIONS

Summary of Main Findings

Our goal in this study was to compare directly the proof validation behavior of research-active mathematicians and beginning undergraduate students. By recording participants' eye movements as they read purported proofs, we were able to gain insights into their real-time attention allocation during validation. Our analysis had three main foci.

First, like Selden and Selden (2003) we found that undergraduate students do not reliably distinguish invalid from valid proofs. Moreover, contrary to the suggestions made by Selden and Selden (2003) and Azzouni (2004), we found that the mathematicians in our sample did not consistently exhibit uniform agreement either. Indeed Proof 2—described as "The Real Thing" by Selden and Selden (p. 15)—was judged to be invalid by 5 out of the 12 mathematicians in our study. An analysis of participants' explanations of their judgments suggested that at least some of these disagreements could be genuinely mathematical, rather than being relatively trivial issues related to style or presentation (see also Inglis et al., in press).

Second, we found that, compared to mathematicians, undergraduate students spend proportionately more time fixating on formulae when validating. This finding provides compelling evidence in favor of Selden and Selden's (2003) hypothesis that a major cause of validation difficulty is the tendency of undergraduates to focus on the surface features of arguments rather than their underlying logical structure.

Third, we investigated Weber and Mejia-Ramos's (2011) suggestion that there are two distinct strategies that can be successfully adopted when validating purported proofs: zooming in and zooming out. We found that a major difference between the undergraduates and mathematicians appears to be the extent to which they zoom in. The mathematicians made nearly

50% more between-line saccades than the undergraduates, and the majority of these additional saccades were between consecutive lines. Evidence from an in-depth analysis of validation behavior on Proof 6 is consistent with the hypothesis that a major cause of these additional consecutive, between-line saccades was that the mathematicians were devoting more effort than the undergraduate students to inferring implicit between-line warrants.

In contrast, we found no evidence that either mathematicians or undergraduates engage in zooming out, either at the beginning of or during an attempted validation. This apparent contradiction between mathematicians' introspective reports (Weber, 2008; Weber & Mejia-Ramos, 2011) and empirical observations of their behavior reflects similar contradictions found by other researchers (e.g., Inglis & Mejia-Ramos, 2009; Weber, 2008). We believe that providing a satisfactory account of these apparent contradictions should be considered an important goal of mathematics education research, and we discuss this issue further subsequently.

Methodological Implications

This study has demonstrated that relying solely upon mathematicians' introspective accounts of their own practice may not lead to a valid understanding of that practice. This is significant for mathematics education more broadly because many theorists believe that a key goal of instruction is to enable students to engage in "authentic mathematics," that is, to adopt practices that resemble those of professional mathematicians (e.g., Harel & Sowder, 1998; Lampert, 1990; RAND Mathematics Study Panel, 2003; Stylianides, 2007; Stylianou, 2002; Weber, 2008). If enabling students to validate proofs like mathematicians is a desirable goal, then we require an accurate understanding of how mathematicians validate. Without such an understanding, instructional designers may inadvertently be developing activities that are

authentic to *perceptions* of mathematical practice, but inauthentic to *actual* mathematical practice. With this in mind, it seems particularly important to determine whether mathematical practice is homogeneous at the research level. Homogeneity is implicitly (and sometimes explicitly) assumed in writings by researchers who advocate the introduction of authentic mathematics into the classroom (e.g., Harel & Sowder, 1998; Lampert, 1990; Stylianides, 2007). Like Weber's findings (2008), however, ours hint that there may not be only one standard for mathematical validity among research mathematicians, suggesting that "authentic mathematics" may consist of a set of norms and practices that is less uniform than is commonly thought. Thus, if we wish to promote authentic mathematical activity, more attention needs to be given to the empirical study of mathematicians' practices, and this could appropriately use a variety of methodological and theoretical approaches. Our own approach complements self-report studies by allowing us to test their findings while reducing problems of reactivity and veridicality. However, it does have other limitations, as discussed next.

Limitations of the Study

The first limitation of this study is that all the participants read the purported proofs onscreen. Two of the 30 participants noted that this was an uncommon reading context for them, and that when reading mathematics they would typically prefer to read printed paper. It is possible that participants might exhibit different reading behavior when reading on paper from reading onscreen. That said, there is a growing trend for universities to put learning resources on Virtual Learning Environments, and for academic journals to be accessed online, so onscreen reading is becoming more common. Furthermore, the use of an eye-tracker allowed us to dispense with the arguably more atypical context used by previous researchers. Whereas students in typical verbal protocol studies are observed in real time by an authority figure with an overt

video camera, our participants moved through the study at their own pace in a private room, observed only by a nonintrusive infrared camera built into the screen.

A second limitation of this study is that, like Weber (2008) and Selden and Selden (2003), we used relatively short proofs. Although these proofs may be reasonably representative of the arguments encountered by undergraduates in their mathematics courses, they are considerably shorter than typical mathematics research texts. Determining whether the behavior of mathematicians is substantially different when engaging with research-level mathematics would be a valuable goal of future research.

A third limitation of the study is that, although we did find significant differences between the behavior of the two groups, our participants were sampled from a single university. Although we believe that the mathematics syllabus delivered at Loughborough is relatively typical (in terms of both content and style) of research-intensive universities, further research with a wider group of participants could profitably investigate whether our findings transfer to other educational contexts.

One final important limitation of this study is that we studied proof validation, not proof comprehension: Each participant had the explicit goal of deciding whether each purported proof was valid or invalid. As Mejía-Ramos and Inglis (2009) have argued, it is possible that proof comprehension (for which the primary goal of the reader is to understand rather than validate) might involve different reading strategies from proof validation. A consequence of this observation is that it may be premature to attempt to design instruction based on these findings if the primary aim is to improve students' proof *comprehension* skills. However, research on reading comprehension in other educational settings (e.g., Chi, de Leeuw, Chiu, & LaVancher,

1994) has tested pedagogical approaches closely related to our findings, thus allowing us to suggest potentially profitable directions for further research.

Educational Implications

Supposing our results are representative of the wider mathematical community, they could have implications for two separate approaches to improving students' proof validation: (a) developing instruction to promote effective student validation strategies, and (b) manipulating how proofs are written.

Developing effective validation strategies via instruction. Drawing implications for educational practice from expert/novice comparisons is not straightforward. It might be that the experts' strategies rely upon knowledge structures that novices do not share, rendering these strategies ineffective if taught without modification. Nevertheless, in some educational domains expert/novice studies have led to pedagogical insights that, after further testing and refinement, have led to substantial learning gains. Here we outline one such research program and relate it to our own work.

Chi et al. (1989) studied students' learning from physics texts and found that students who were more successful at solving problems that were isomorphic to those they had studied spontaneously generated more "self-explanations"—interpretations of what had been read that involved information and relationships beyond those literally contained in the text. An important follow-up question was: If students were taught to adopt the strategies spontaneously used by successful readers, would their learning improve? The answer seems to be *yes*: Chi et al. (1994) found that school students who were asked to self-explain while reading a biology text learned significantly more than students simply asked to read the text twice. Since these initial studies, the self-explanation effect has been robustly demonstrated to lead to learning gains in a variety

of domains, including history (Wolfe & Goldman, 2005), programming (Bielaczyc, Pirolli, & Brown, 1995), and statistics (Renkl, 1997).

The self-explanation effect provides a useful case for understanding how insights from expert/novice studies might eventually lead to improved pedagogy: The first stage was to investigate the strategies adopted by expert and novice readers, the second was to develop training materials that sought to encourage all learners to use the same strategies, and the third demonstrated the efficacy of these training materials. Finally, insights from these training materials were incorporated into genuine instructional materials (for a review, see Roy & Chi, 2005).

The results we have reported in this article focused on the strategies adopted by experts and novices. We found that the experts spent considerably more effort attempting to infer implicit between-line warrants, suggesting that it would be worthwhile to develop materials that encourage all validators to (a) decide whether a warrant is required, (b) infer an appropriate warrant, and (c) evaluate the inferred warrant. Such materials may be relatively simple to produce: The self-explanation training materials used by Ainsworth and Burcham (2007) consisted of a brief training package that defined and illustrated the self-explanation strategy and reported evidence about its effectiveness. Engaging with these short training materials led to a 30% increase in postreading comprehension-test scores; it is plausible that a similarly straightforward manipulation could lead to substantial improvements in students' proof validation skills.

Can proofs be written to aid validation attempts? An alternative approach to improving proof validation could be to write the to-be-validated proofs in a way that encourages effective

validation. This study provides evidence suggestive of two manipulations that might provide such encouragement.

First, we found that undergraduate students spent proportionately longer than mathematicians studying the formulae contained within the purported proofs. We have hypothesized that this focus on algebraic manipulation may be distracting students from engaging in the inference of implicit warrants. Although some proofs clearly cannot be expressed without using a large amount of symbolism, this can be reduced in many proofs (for example "3 | n^2" in Proof 3 could be rewritten "3 is a divisor of n^2"); and it is plausible to hypothesize that reducing symbolism in this fashion could lead students to allocate more of their attention to the logical connections between statements, rather than to what Selden and Selden (2003) called *surface features*. Such a result would be consistent with Österholm's (2005) finding that reducing symbolism in a mathematical text improved students' comprehension of the underlying mathematics being communicated. Of course, manipulating a text in this way would reduce students' experience with symbolism, so the goal of improving validation would need to be balanced against the need to provide this experience.

A second type of manipulation would be simply to state more warrants explicitly rather than leaving the reader to infer them. Thus, the burden on the validator would be reduced from (a) deciding whether a warrant is required, (b) inferring an appropriate warrant, and (c) evaluating the inferred warrant, to only step (c). Arguably it is this role—explicating the implicit warrants contained in a written proof—that is fulfilled by a lecturer when presenting a proof in a traditional lecture (e.g., Weber, 2004).

Although such explication of warrants might seem to be a sensible strategy, judging the extent to which it should be used may be a nontrivial task. If too few warrants are explicated,

then validators may be unable to complete successfully a zooming-in validation, but if too many warrants are explicated, then the validator need not infer anything, and the level of cognitive engagement with the text is likely to suffer. Clearly, the extent to which warrants should be made explicit will depend upon both the complexity of the proof itself and the sophistication of its intended readers. One might even expect an expertise reversal effect: A level of detail that would be helpful for less successful students may be distracting and disruptive for more successful students (cf. Kalyuga, 2007; Tobias, 1989). Further research is thus required to determine the extent to which warrants should be made explicit.

CONCLUDING REMARKS

It is now widely accepted that proof should be central to the study of mathematics, at least at the secondary and university levels. Research has investigated students' conceptions of proof (e.g., Bell, 1976), and pedagogical strategies have been developed to help students develop a deductive proof scheme (e.g., Stylianides & Stylianides, 2009b). However, even if students understand that mathematicians will accept only deductive arguments as proofs—and research indicates that most undergraduate students are in this position (e.g., Weber, 2010)—there is little research that provides insight into how best to develop their ability to successfully engage with such arguments. Here, we have reported the first direct comparison between the proof validation behavior of undergraduate students and research mathematicians. We have demonstrated that mathematicians validate proofs in a substantially different manner from undergraduate students: They appear to expend more effort inferring implicit between-line warrants and attend less to algebraic manipulations. The challenge now is to use these insights to develop, test, and refine pedagogical strategies that will improve students' skills in engaging with, and learning from, mathematical proofs.

REFERENCES

Ainsworth, S., & Burcham, S. (2007). The impact of text coherence on learning by self-explanation. *Learning and Instruction, 17,* 286–303. doi:10.1016/j.learninstruc.2007.02.004

Alcock, L., & Weber, K. (2005). Proof validation in real analysis: Inferring and checking warrants. *Journal of Mathematical Behavior*, *24*, 125–134. doi:10.1016/j.jmathb.2005.03.003

Azzouni, J. (2004). The derivation-indicator view of mathematical practice. *Philosophia Mathematica, 12,* 81–105. doi:10.1093/philmat/12.2.81

Ball, L. J., Lucas, E. J., Miles, J. N. V., & Gale, A. G. (2003). Inspection times and the selection task: What do eye-movements reveal about relevance effects? *Quarterly Journal of Experimental Psychology*, *56A*, 1053–1077. doi:10.1080/02724980244000729

Bell, A. W. (1976). A study of pupils' proof-explanations in mathematical situations. *Educational Studies in Mathematics*, *7*, 23–40. doi:10.1007/BF00144356

Bielaczyc, K., Pirolli, P. L., & Brown, A. L. (1995). Training in self-explanation and self-regulation strategies: Investigating the effects of knowledge acquisition activities on problem solving. *Cognition and Instruction, 13*, 221–252. doi:10.1207/s1532690xci1302_3

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145–182. doi:10.1016/0364-0213(89)90002-5

Chi, M. T. H., de Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439–477. doi:10.1016/0364-0213(94)90016-7

Coe, R., & Ruthven, K. (1994). Proof practices and constructs of advanced mathematical students. *British Educational Research Journal*, *20*, 41–53. doi:10.1080/0141192940200105

Cowen, C. C. (1991). Teaching and testing mathematics reading. *The American Mathematical Monthly, 98*, 50–53.

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research, 36*, 1827–1837. doi:10.1016/0042-6989(95)00294-4

Ericsson, K. A. (2006). Protocol analysis and expert thought: Concurrent verbalizations of thinking during experts' performance on representative tasks. In K. A. Ericsson, N. Charness, R. R. Hoffman, & P. Feltovich (Eds.), *Cambridge handbook of expertise and expert performance* (pp. 223–242). Cambridge, England: Cambridge University Press.

Ericsson, K. A., & Simon, H. A. (1984). Protocol analysis: Verbal reports as data. Cambridge, MA: MIT Press.

Fallis, D. (2003). Intentional gaps in mathematical proofs. *Synthèse*, *134*, 45–69. doi:10.1023/A:1022131513275

Gould, J. D. (1973). Eye movements during visual search and memory search. *Journal of Experimental Psychology, 98*, 184–195. doi:10.1037/h0034280

Hanna, G. (1991). Mathematical proof. In A. J. Bishop (Managing Ed.) & D. Tall (Vol. Ed.), *Mathematics education library: Vol. 11. Advanced mathematical thinking* (pp. 54–61) Dordrecht, the Netherlands: Kluwer.

Hanna, G. (2007). The ongoing value of proof. In W. M. Roth, L. Verschaffel (Series Eds.), & P. Boero (Vol. Ed.), *New directions in mathematics and science education: Vol. 1. Theorems in school: From history, epistemology and cognition to classroom practice* (pp. 3–18). Rotterdam, the Netherlands: Sense.

Harel, G., & Sowder, L. (1998). Students' proof schemes: Results from exploratory studies. In T. Dick (Managing Ed.), A. H. Schoenfeld, J. Kaput, & E. Dubinsky (Vol. Eds.), *CBMS issues in mathematics education: Vol. 7. Research in collegiate mathematics education III* (pp. 234–282). Providence, RI: American Mathematical Society.

Harel, G., & Sowder, L. (2007). Toward comprehensive perspectives on the learning and teaching of proof. In F. K. Lester Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 805–842). Charlotte, NC: Information Age.

Hazzan, O., & Zazkis, R. (2003). Mimicry of proofs with computers: The case of linear algebra. *International Journal of Mathematical Education in Science and Technology, 34,* 385–402.

Healy, L. & Hoyles, C. (2000). A study of proof conceptions in algebra. *Journal for Research in Mathematics Education*, *31*, 396–428.

Hoffman, R. R. (Ed.). (1992). *The psychology of expertise: Cognitive research and empirical AI*. New York, NY: Springer-Verlag.

Horsey, R. (2002). The art of chicken sexing. *UCL Working Papers in Linguistics, 14*, 107–117. http://www.ucl.ac.uk/psychlangsci/research/linguistics/publications

Hyönä, J. (2010). The use of eye movements in the study of multimedia learning. *Learning and Instruction, 20,* 172–176. doi:10.1016/j.learninstruc.2009.02.013

Inglis, M., & Mejia-Ramos, J. P. (2009). The effect of authority on the persuasiveness of mathematical arguments. *Cognition and Instruction*, *27*, 25–50. doi:10.1080/07370000802584513

Inglis, M., Mejia-Ramos, J. P., & Simpson, A. (2007). Modelling mathematical argumentation: The importance of qualification. *Educational Studies in Mathematics, 66,* 3–21. doi:10.1007/s10649-006-9059-8

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

Jepma, M., & Nieuwenhuis, S. (2011). Pupil diameter predicts changes in the exploration-exploitation tradeoff: Evidence for the adaptive gain theory of locus coeruleus function. *Journal of Cognitive Neuroscience, 23,* 1587–1596. doi:10.1162/jocn.2010.21548

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology, 8,* 441–480. doi:10.1016/0010-0285(76)90015-3

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixation to comprehension. *Psychological Review, 87*, 329–354. doi:10.1037/0033-295X.87.4.329

Kaakinen, J. K., Hyönä, J., & Keenan, J. M. (2002). Perspective effects on online text processing. *Discourse Processes, 33,* 159–173.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review, 19,* 509–539. doi:10.1007/s10648-007-9054-3

Konior, J. (1993). Research into the construction of mathematical texts. *Educational Studies in Mathematics*, *24*, 251–256. doi:10.1007/BF01275425

Knuth, E. J. (2002). Secondary school mathematics teachers' conceptions of proof. *Journal for Research in Mathematics Education*, *33*, 379–405.

Lampert, M. (1990). When the problem is not the question and the solution is not the answer: Mathematical knowing and teaching. *American Educational Research Journal, 27,* 29–63. doi:10.3102/00028312027001029

Lamport, L. (1994). *LaTeX: A document preparation system. User's guide and reference manual* (2nd ed.). Reading, MA: Addison-Wesley.

Liversedge, S. P., Paterson, K. B., & Pickering, M. J. (1998). Eye movements and measures of reading time. In G. Underwood (Ed.), *Eye guidance in reading and scene perception* (pp. 55-76). Oxford, England: Elsevier Science.

Mamona-Downs, J., & Downs, M. (2005). The identity of problem solving. *Journal of Mathematical Behavior*, *24*, 385–401. doi:10.1016/j.jmathb.2005.09.011

Manin, Y. I. (1977). *Graduate texts in mathematics: Vol. 53. A course in mathematical logic* (N. Koblitz, Trans.; P. R. Halmos, F. W. Gehring, & C. C. Moore, Series Eds.). New York, NY: Springer-Verlag.

Martin, W. G., & Harel, G. (1989). Proof frames of preservice elementary teachers. *Journal for Research in Mathematics Education*, *20*, 41–51. http://www.jrme.org/publications/jrme.aspx

Matin, E. (1974). Saccadic suppression: A review and an analysis. *Psychological Bulletin, 81*, 899–917. doi:10.1037/h0037368

Mejía-Ramos, J. P. & Inglis, M. (2009). Argumentative and proving activities in mathematics education research. In F.-L. Lin, F.-J. Hsieh, G. Hanna, & M. de Villiers (Eds.),

Proceedings of the ICMI Study 19 conference: Proof and proving in mathematics education (Vol. 2, pp. 88–93). Taipei, Taiwan: National Taiwan Normal University.

Merkley, R. & Ansari, D. (2010). Using eye tracking to study numerical cognition: The case of the ratio effect. *Experimental Brain Research, 206,* 455–460. doi:10.1007/s00221-010-2419-8

Moeller, K., Fischer, M. H., Nuerk, H. C., & Willmes, K. (2009). Eye fixation behaviour in the number bisection task: Evidence for temporal specificity. *Acta Psychologica, 131*, 209–220. doi:10.1016/j.actpsy.2009.05.005

Moore, R. C. (1994). Making the transition to formal proof. *Educational Studies in Mathematics*, *27*, 249–266. doi:10.1007/BF01273731

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231–259. doi:10.1037/0033-295X.84.3.231

Österholm, M. (2005). Characterizing reading comprehension of mathematical texts. *Educational Studies in Mathematics, 63,* 325–346. doi:10.1007/s10649-005-9016-y

Pollatsek, A., Lesch, M., Morris, R. K., & Rayner, K. (1992). Phonological codes are used in integrating information across saccades in word identification and reading. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 148–162. doi:10.1037/0096-1523.18.1.148

RAND Mathematics Study Panel. (2003). *Mathematical proficiency for all students: Toward a strategic research and development program in mathematics education*. Santa Monica, CA: RAND Corporation.

Rav, Y. (1999). Why do we prove theorems? *Philosophia Mathematica*, *7*, 5–41. doi:10.1093/philmat/7.1.5

Rav, Y. (2007). A critique of a formalist-mechanist version of the justification of arguments in mathematicians' proof practices. *Philosophia Mathematica, 15,* 291–320. doi:10.1093/philmat/nkm023

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372–422. doi:10.1037/0033-2909.124.3.372

Rayner, K. (2009). The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology, 62,* 1457–1506. doi:10.1080/17470210902816461

Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest*, *2*, 31–74. doi:10.1111/1529-1006.00004

Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science, 21*, 1–29. doi:10.1016/S0364-0213(99)80017-2

Rowland, T. (2002). Generic proofs in number theory. In C. A. Maher, R. Speiser (Series Eds.), S. R. Campbell & R. Zazkis (Vol. Eds.), *Learning and teaching number theory: Research in cognition and instruction* (pp. 157–184). Westport, CT: Ablex.

Roy, M., & Chi, M. T. H. (2005). The self-explanation principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning*, (pp. 271–286). Cambridge, England: Cambridge University Press.

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory and Cognition, 17,* 759–769. doi:10.3758/BF03202637

Sangwin, C. (2010). Intriguing integrals: Part II. *Plus Magazine, 54.* Retrieved from http://plus.maths.org/issue54/features/sangwin2/

Schooler, J. W., Ohlsson, S., & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General, 122,* 166–183. doi:10.1037/0096-3445.122.2.166

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. http://www.nctm.org/publications/jrme.aspx

Singh, S. (1997). *Fermat's last theorem: The story of a riddle that confounded the world's greatest minds for 358 years*. London, England: Fourth Estate.

Stylianides, A. J. (2007). Proof and proving in school mathematics. *Journal for Research in Mathematics Education*, *38*, 289–321. http://www.nctm.org/publications/jrme.aspx

Stylianides, A. J., & Stylianides, G. J. (2009a). Proof constructions and evaluations. *Educational Studies in Mathematics, 72,* 237–253. doi:10.1007/s10649-009-9191-3

Stylianides, G. J., & Stylianides, A. J. (2009b). Facilitating the transition from empirical arguments to proof. *Journal for Research in Mathematics Education, 40*, 314–352. http://www.nctm.org/publications/jrme.aspx

Stylianou, D. A. (2002). On the interaction of visualization and analysis: The negotiation of a visual representation in problem solving. *Journal of Mathematical Behavior*, *21*, 303–317. doi:10.1016/S0732-3123(02)00131-1

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

Tobias, S. (1989). Another look at research on the adaptation of instruction to student characteristics. *Educational Psychologist, 24,* 213–227. doi:10.1207/s15326985ep2403_1

Tobii Technology. (2010). *Tobii eye tracking: An introduction to eye tracking and Tobii Eye Trackers*. Stockholm, Sweden: Tobii Technology AP.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge, England: Cambridge University Press.

Vinner, S. (1997). The pseudo-conceptual and pseudo-analytic thought processes in mathematics learning. *Educational Studies in Mathematics*, *34*, 97–129. doi:10.1023/A:1002998529016

Watson, D. G., & Inglis, M. (2007). Eye movements and time-based selection: Where do the eyes go in preview search? *Psychonomic Bulletin & Review*, *14*, 852–857. doi:10.3758/BF03194111

Weber, K. (2001). Student difficulty in constructing proofs: The need for strategic knowledge. *Educational Studies in Mathematics*, *48*, 101–119. doi:10.1023/A:1015535614355

Weber, K. (2004). Traditional instruction in advanced mathematics courses: A case study of one professor's lectures and proofs in an introductory real analysis course. *Journal of Mathematical Behavior*, *23*, 115–133. doi:10.1016/j.jmathb.2004.03.001

Weber, K. (2008). How mathematicians determine if an argument is a valid proof. *Journal for Research in Mathematics Education*, *39*, 431–459. http://www.nctm.org/publications/jrme.aspx

Weber, K. (2010). Mathematics majors' perceptions of conviction, validity, and proof. *Mathematical Thinking and Learning, 12,* 306–336. doi:10.1080/10986065.2010.495468

Weber, K., & Alcock, L. (2004). Semantic and syntactic proof productions. *Educational Studies in Mathematics*, *56*, 209–234. doi:10.1023/B:EDUC.0000040410.57253.a1

Weber, K., & Alcock, L. (2005). Using warranted implications to understand and validate proofs. *For the Learning of Mathematics*, *25*(1), 34–38, 51.

Weber, K. & Mejia-Ramos, J.-P. (2011). Why and how mathematicians read proofs: An exploratory study. *Educational Studies in Mathematics, 76,* 329–344. doi:10.1007/s10649-010-9292-z

Wolfe, M. B. W., & Goldman, S. R. (2005). Relations between adolescents' text processing and reasoning. *Cognition and Instruction, 23*, 467–502. doi:10.1207/s1532690xci2304_2

Authors

Matthew Inglis, Mathematics Education Centre, Loughborough University, Loughborough, LE11 3TU, United Kingdom; m.j.inglis@lboro.ac.uk

Lara Alcock, Mathematics Education Centre, Loughborough University, Loughborough, LE11 3TU, United Kingdom; l.j.alcock@lboro.ac.uk

Accepted January 3, 2012

Appendix

Note that the stimuli used in this study have the same content and line breaks as the proofs given below, but were typeset in the Computer Modern font. The original proofs (image files) as seen by the participants are available upon request.

Practice Proof.

Theorem. The product of any two odd numbers is also odd.

Proof. Let n and m be two odd numbers.

In other words $n = 2k+1$ and $m = 2l + 1$ for some $k,l \in \mathbf{Z}$.

Consider $nm = (2k+1)(2l+1) = 4kl + 2k + 2l + 1 = 2(2kl + k + l) + 1$, which is odd.

So nm is odd.

Proof 1.

Theorem. For any positive integer n, if n^2 is divisible by 3, then n is divisible by 3.

Proof. Assume that n^2 is an odd positive integer that is divisible by 3.

That is, $n^2 = (3n+1)^2 = 9n^2+6n+1 = 3n(n+2) + 1$.

Therefore, n^2 is divisible by 3.

Assume that n^2 is even and a multiple of 3.

That is $n^2 = (3n)^2 = 9n^2 = 3n(3n)$.

Therefore, n^2 is a multiple of 3.

If we factor $n^2 = 9n^2$, we get $3n(3n)$; which means that n is a multiple of 3.

Proof 2.

Theorem. For any positive integer n, if n^2 is divisible by 3, then n is divisible by 3.

Proof. Suppose to the contrary that n is not a multiple of 3.

We will let $3k$ be a positive integer that is a multiple of 3,

so that $3k + 1$ and $3k + 2$ are integers that are not multiples of 3.

Now $n^2 = (3k + 1)^2 = 9k^2 + 6k + 1 = 3(3k^2 + 2k) + 1$.

Since $3(3k^2 + 2k)$ is a multiple of 3, $3(3k^2 + 2k) + 1$ is not.

Now we will do the other possibility, $3k + 2$.

So, $n^2 = (3k + 2)^2 = 9k^2 + 12k + 4 = 3(3k^2 + 4k + 1) + 1$ is not a multiple of 3.

Because n^2 is not a multiple of 3, we have a contradiction.

Proof 3.

Theorem. For any positive integer n, if n^2 is divisible by 3, then n is divisible by 3.

Proof. Let n be an integer such that $n^2=3x$ where x is any integer.

Then $3 \mid n^2$.

Since $n^2 = 3x$, $nn = 3x$.

Thus $3 \mid n$.

Therefore if n^2 is a multiple of 3, then n is a multiple of 3.

Proof 4.

Theorem. For any positive integer n, if n^2 is divisible by 3, then n is divisible by 3.

Proof. Let n be a positive integer such that n^2 is a multiple of 3.

Then $n = 3m$, where $m \in \mathbf{Z}^+$.

So $n^2 = (3m)^2 = 9m^2 = 3(3m^2)$.

This breaks down into 3*m* times 3*m* which shows that *m* is a multiple of 3.

Proof 5.

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

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

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

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

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

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

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

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

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

Proof 6.

Note: The lines beginning "Every number . . ." and "So dividing . . ." each appeared as a single line in the onscreen proof.

Theorem. There are infinitely many primes that can be written as $4k+1$ (where $k \in \mathbf{Z}$)

Proof. Suppose there are finitely many primes of the form $4k + 1$.

Then these primes can be listed $p_1, p_2, p_3, \ldots, p_n$.

Define a number *a* as follows. Let $a = p_1p_2p_3 \ldots p_n + 4$.

Note that dividing a by 4 leaves remainder 1.

Every number that leaves remainder 1 when divided by 4 is divisible by a prime that also leaves remainder 1 when divided by 4.

However, for all i such that $1 \leq i \leq n$, p_i divides $p_1p_2p_3 \ldots p_n$ and p_i does not divide 4.

Thus p_i does not divide a.

So dividing a by 4 leaves remainder 1 and a is not divisible by any prime that leaves remainder 1 when divided by 4.

This is a contradiction.

Tables

Table 1

Responses to the Six Arguments, Showing Frequencies of Valid and Invalid Responses, for Each Group

	P1*	P2	P3	P4*	P5	P6*
			Mathematicians			
Valid	0	7	5	0	6	0
Invalid	12	5	7	12	6	12
			Undergraduates			
Valid	9	11	4	11	11	8
Invalid	9	7	14	7	7	10

Note: * indicates that the difference between the responses of the two groups reached significance (Fisher's Exact Test).

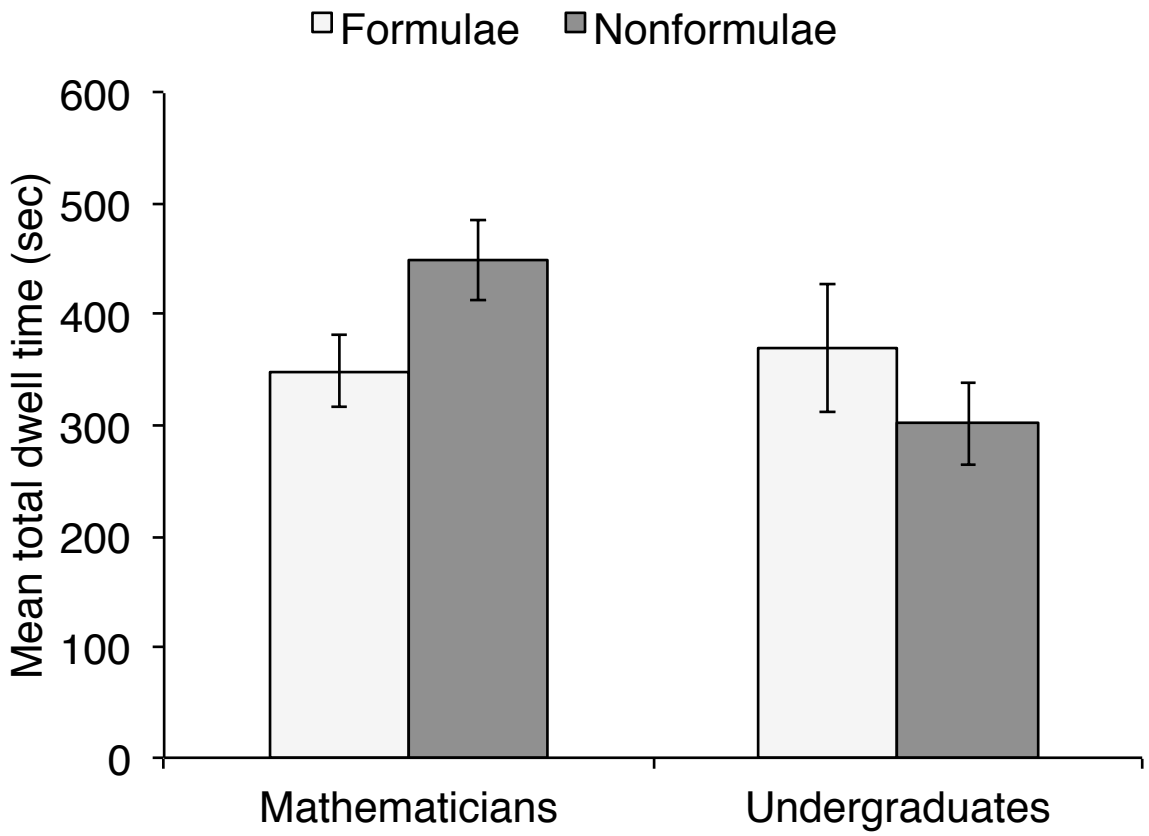

Figure 1. The mean total dwell times of mathematicians and undergraduates on the formulae and nonformulae, across all proofs in the study. Error bars show ±1 SE of the mean.

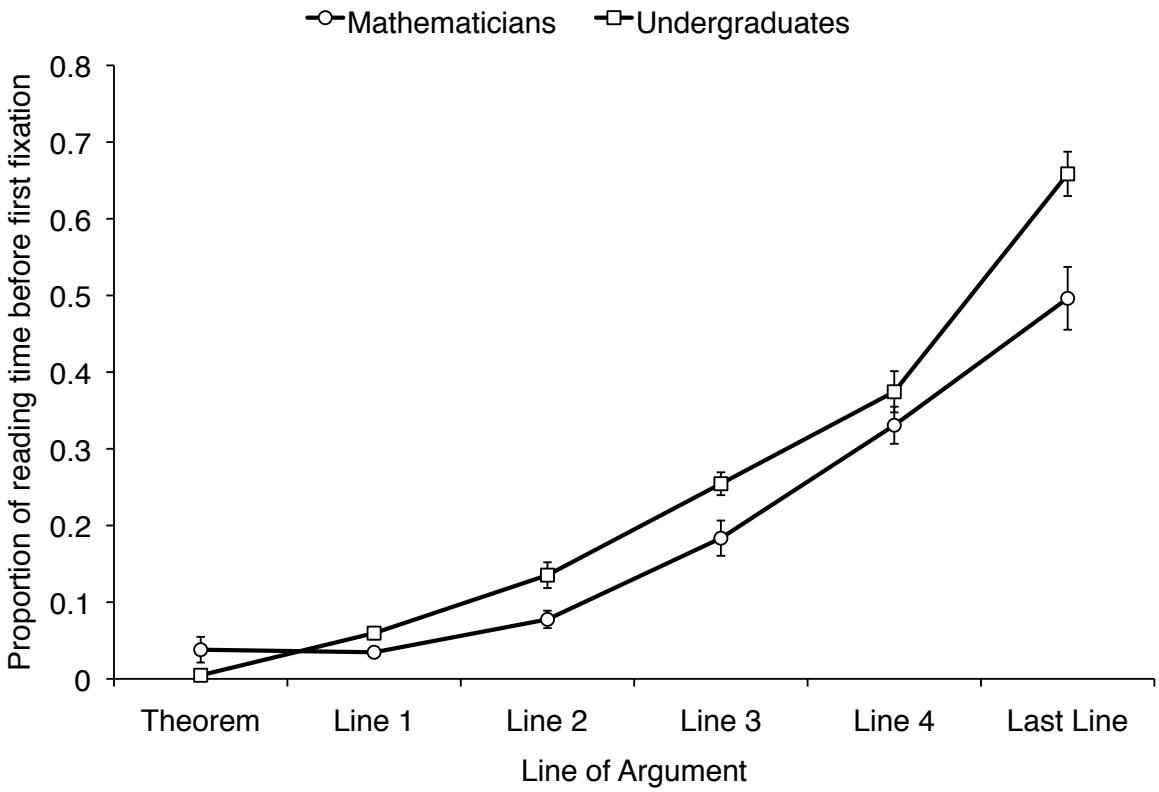

Figure 2. The mean proportion of reading time before participants first fixated on each line, collapsed across arguments. Error bars show ±1 SE of the mean.

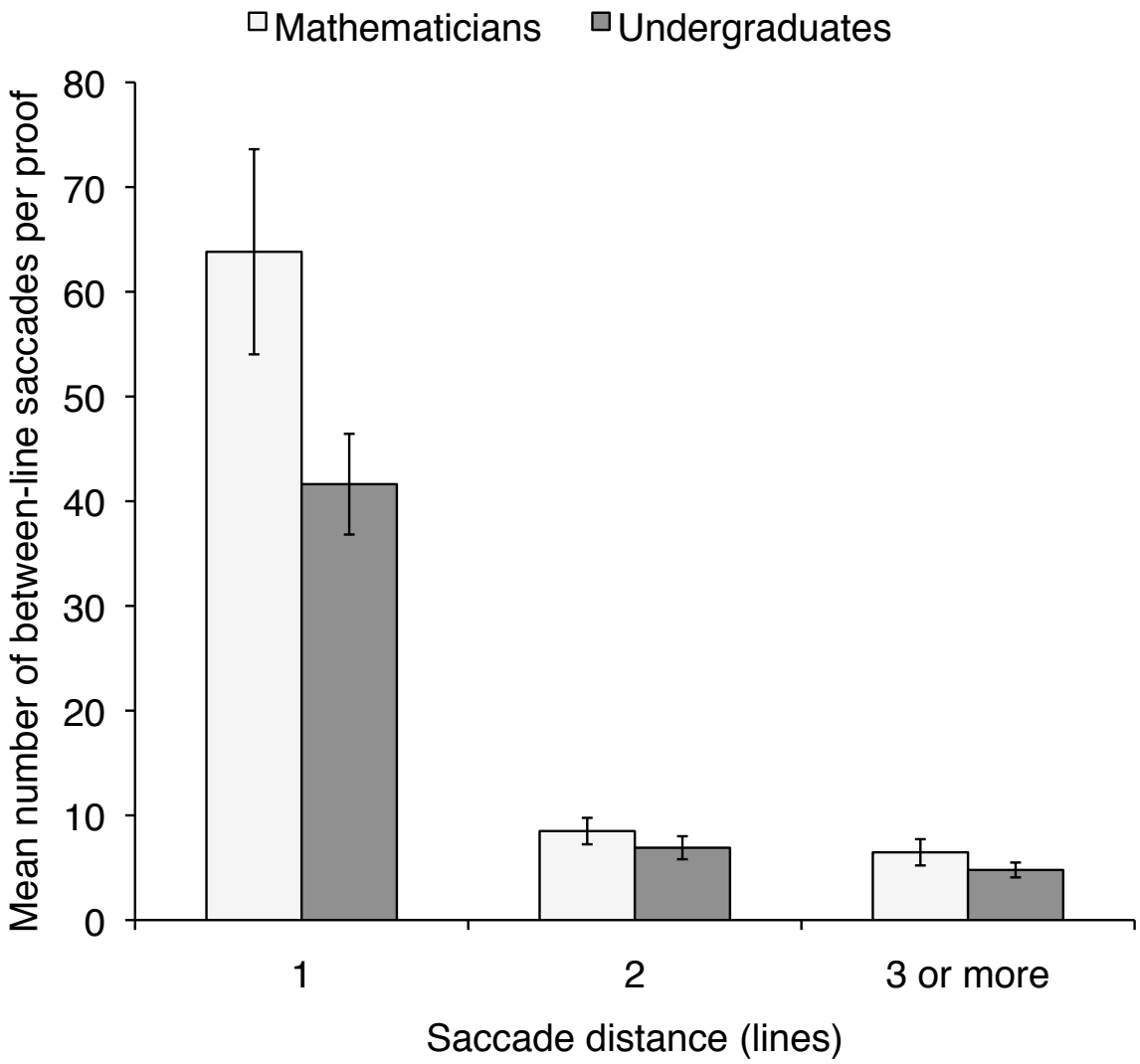

Figure 3. The mean number of between-line saccades of different lengths made per proof by mathematicians and undergraduates. Error bars show ±1 SE of the mean.

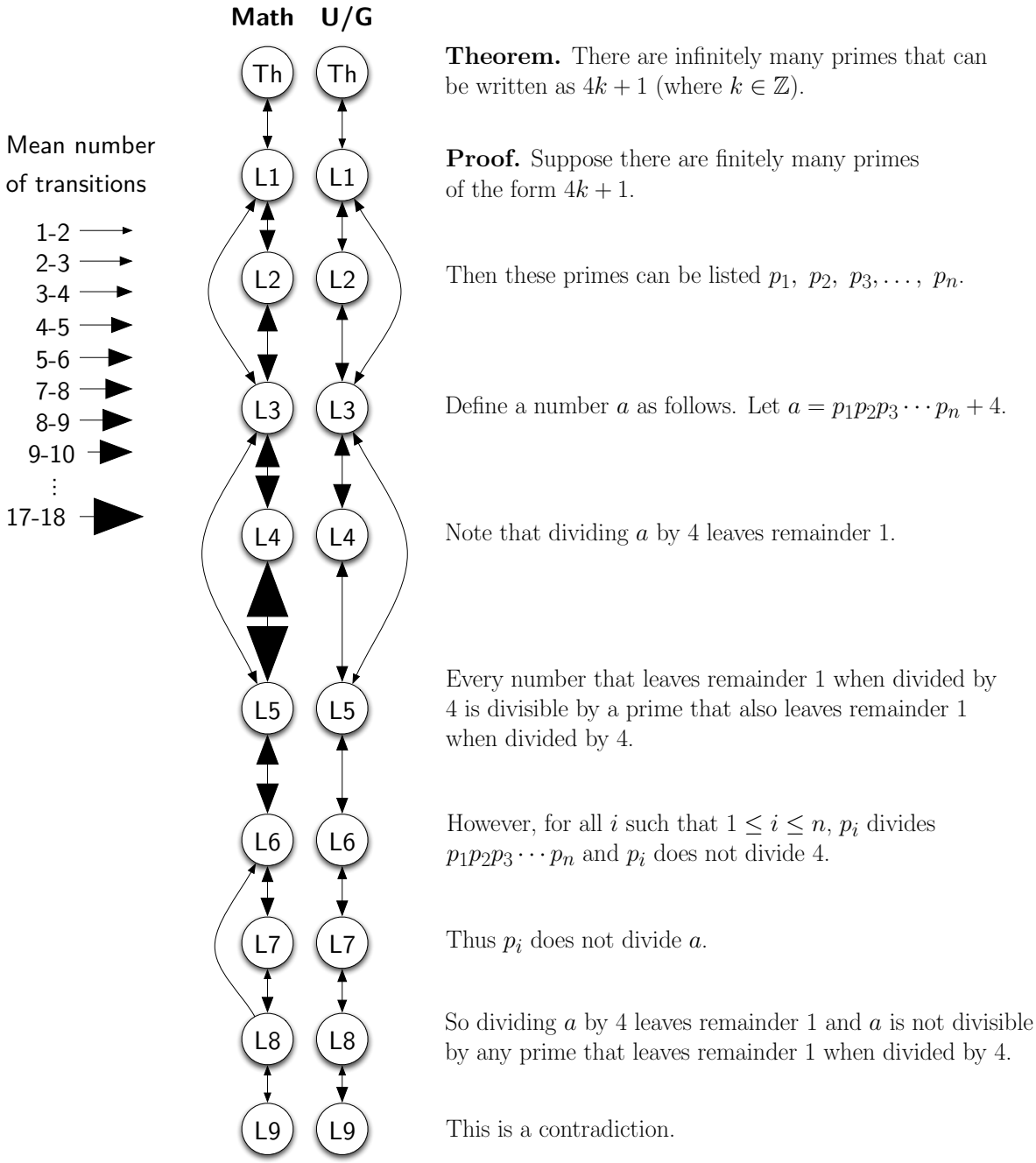

Figure 4. State transition diagram for Proof 6, showing the mean number of between-line transitions made by mathematicians (Math) and undergraduates (U/G). Note that when viewed by participants, each line was presented without line-breaks.

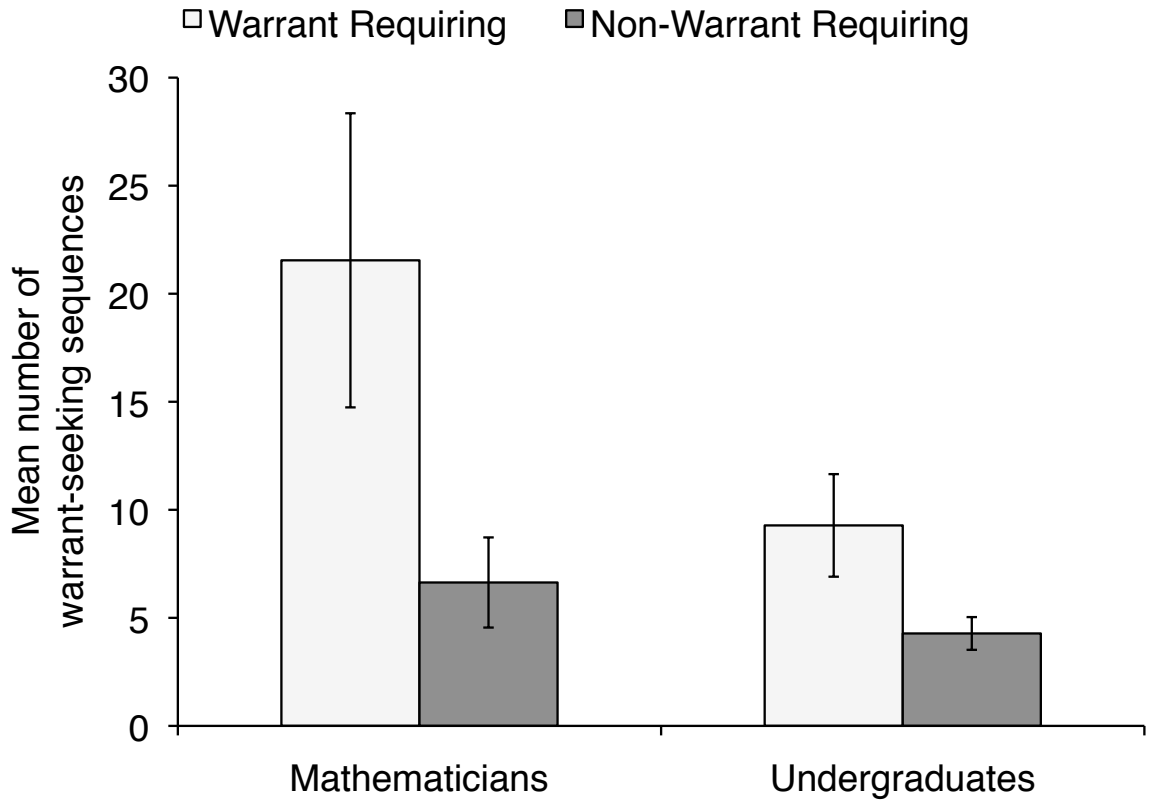

Figure 5. The mean number of warrant-seeking eye-movement sequences (sequences from Line *x* to Line *x*–1 and back to Line *x*) made during the reading of Proof 6, for neighboring line transitions that required warrants and those that did not. Error bars show ±1 SE of the mean.