Reading Proofs for Validation and Comprehension: An Expert-Novice Eye-Movement Study

Anja Panse*, Lara Alcock and Matthew Inglis****

*Institut für Mathematik, Universität Paderborn, Germany;

apanse@math.upb.de; +49 5251 60-2620

**Mathematics Education Centre, Loughborough University, UK

Keywords: *mathematical texts; reading behavior; proof validation*; *undergraduate mathematics; experimental methods*

Author Note: Raw data and analysis scripts associated with this paper can be found at https://doi.org/10.6084/m9.figshare.5213326.v1

Abstract

Does reading a mathematical proof for validation engender different behaviors from reading it for comprehension? Experts and novices each read two mathematical proofs under different sets of instructions: they were asked to understand one proof, and to assess the validity of the other. Their eye movements were recorded while they read and were analyzed to investigate possible differences in attention allocation, in cognitive demand and in the mathematical reading process. We found negligible differences in reading behaviors under the two sets of instructions, and we discuss the implications of this for theoretical development, research methodology and pedagogical practice.

Reading Proofs for Comprehension and Validation: An Expert-Novice Eye-Movement Study

Introduction and Theoretical Background

Proving is a core element of professional mathematical practice (Rav, 1999), and mathematicians communicate in part via proofs both among themselves and in their interactions with students (Lew, Fukawa-Connelly, Mejía-Ramos & Weber, 2016). Consequently, proof is a central concern among mathematicians (Thurston, 2000), philosophers (Larvor, 2016), and mathematics educators (Stylianides, Bieda & Morselli, 2016) interested in enculturating novices into the mathematical community. It is generally accepted—indeed, it is codified in policy documents (National Council of Teachers of Mathematics, 2000)—that students at all levels should be offered opportunities to construct and evaluate mathematical proofs, and to develop their understanding of which arguments are considered valid and rigorous by expert mathematicians (Ball & Bass, 2000). Therefore, mathematics education researchers seek to study and promote effective reasoning among all students from elementary school (Yackel & Cobb, 1996) to advanced undergraduate level (Yackel, Rasmussen & King, 2000), as well as among their teachers (Knuth, 2002).

In reflecting upon these investigations, mathematics educators have identified both strengths and weaknesses in research-based knowledge. In a 2009 review of studies on proof in mathematics education, Mejía-Ramos and Inglis (2009) documented the fact that published research was dominated by work on proof construction. Clearly proof construction is a valuable skill: one shared aim of mathematics educators is to help students engage in valid reasoning (Stylianides, 2007). But Mejía-Ramos and Inglis argued that focusing on construction had led researchers to overlook other aspects of engagement with proof, and that reading in particular deserved more attention. This is self-evidently true when considering the realities of mathematical education, at least for many undergraduates. In traditional lectures, the dominant pedagogy assumes that students will learn about specific proofs and proof in general by understanding proofs written by others, so that reading is crucial (Lew, Fukawa-Connelly, Mejía-Ramos & Weber, 2016; Selden & Selden, 1995). In more student-centred pedagogies, students spend more time attempting to construct and present proofs. But the central point holds, because such classes rely on students understanding and evaluating arguments constructed by peers, sometimes informally in small-group or whole-class conversations (Larsen, Johnson & Bartlo, 2013; Rasmussen, Wawro & Zandieh, 2015), and sometimes more formally in specific critiquing assignments (Jones & Alcock, 2014; Kasman, 2006).

When Mejía-Ramos and Inglis (2009) made their observations, research on reading proofs had begun. Selden and Selden (2003), for instance, had focused on *validation*, the process of deciding whether a proof is valid.

Using a teaching interview method, they had found that undergraduates performed at chance when first asked to evaluate a proof, but improved with opportunities to reflect. Work had also begun on proof *comprehension*, where early discussion was driven by mathematicians dissatisfied with traditional assessment that asked students to reproduce proofs. Both Cowen (1991) and Conradie and Frith (2000) had argued that it is not clear what such reproduction tests. Does a student who produces a perfect proof really understand it? And might such testing in itself promote undesirable behaviors like rote memorization? In response, Conradie and Frith discussed their approach to constructing and using proof comprehension tests, describing both test items and practical issues surrounding their implementation.

Research on both validation and comprehension has since progressed in a number of directions. First, key results on validation have been replicated. There is no doubt that many undergraduates are unreliable validators: studies using Selden and Selden's items among others have consistently found that students often fail to notice serious logical errors (Inglis & Alcock, 2012; Weber, 2010). Second, we now understand better why this might be. Interview studies have indicated that undergraduates who are good general readers might nevertheless fail to understand mathematics texts because they pay insufficient attention to detail and do not respond effectively to confusion and errors (Shepherd, Selden & Selden, 2012); similarly, an eye-movement study has suggested that, compared with experts, undergraduates might attain poorer understanding of a proof's logical structure because they focus proportionately less on the words and on logical relationships (Inglis & Alcock, 2012). Third, we know that assessing validation judgement is not as clear-cut as researchers had tended to assume. Inglis and Alcock's (2012) expert-novice study revealed that while experts all agreed about arguments with clear reasoning errors (e.g., proving the converse of a statement rather than the statement itself), they disagreed for arguments that were poorly expressed or that had a 'gap'. A follow-up study demonstrated that this finding was robust among a larger sample of 109 mathematicians: 29 judged a purported proof valid, and most were unwilling to change their minds when confronted with information about the majority view (Inglis, Mejía-Ramos, Weber & Alcock, 2013). This makes research on proof validation challenging—if there are no 'right answers' about validity, then 'correct' validity judgements cannot be used as a straightforward measure of understanding. But it also makes related research potentially more authentic. In the longer term, a better understanding of expert behavior should improve our ability to support students in learning to approach validation in expert-like ways (Weber, Inglis & Mejía-Ramos , 2014).

Work on comprehension has developed along related lines. Task-based interview studies have illuminated the detail of reading behaviors in graduate students and experts, suggesting that experienced readers often 'read the meaning' rather than literal symbols, and that they perform frequent comprehension checks and make thorough attempts to rectify failures (Shepherd & van de Sande, 2014). Theoretical work has explicated

what proof comprehension might entail, in geometry (Yang & Lin, 2008) and in a general framework designed to apply to any area of mathematics (Mejía-Ramos et al., 2012). This work has led, via robust psychometric methods, to short-form multiple-choice proof comprehension tests for three commonly encountered proofs (Mejía-Ramos, Lew, de la Torre & Weber, in press). Other work has used interventions to examine ways in which reading for comprehension might be positively influenced. Self-explanation training has been adapted for undergraduate proof reading, and shown to promote better comprehension test performance and more expert-like reading behavior (Hodds, Alcock & Inglis, 2014). Finally, both mathematicians and educators continue to experiment with ways to incorporate critiquing activities into both single and more systematic instantiations of inquiry-based learning (Hayward, Kogan & Laursen, 2016; Larsen, Johnson & Bartlo, 2013).

One thing not studied directly, however, is the relationship between validation and comprehension. As noted by Mejía-Ramos and Inglis (2009), these are at least potentially different. Reading for comprehension positions the reader as a learner in the traditional sense, as one who cedes evaluative authority to the argument's source. This means that the reader can devote attention exclusively to interpreting the argument in light of prior knowledge, hopefully thereby integrating its ideas with that prior knowledge (cf. Roy, Inglis & Alcock, in press). Reading for validation is at least potentially more complex. It positions the reader not as learner but as critic, as one who takes evaluative authority and judges the argument. This certainly requires comprehension: prior knowledge must be mobilized to judge validity. But it also requires active decisions at multiple levels: a validator must check individual inferences and reach a global judgement. Global validity judgements are known to be difficult even for direct proofs (Selden & Selden, 2003), and are arguably more so for complex indirect proofs (Brown, 2014).

Thus it could be that validation and comprehension require different processing, although it is not obvious how this would manifest in reading behavior. One might hypothesize that reading for validation involves more processes so would take longer. But time to judgement could interact with features of given arguments, because error detection might occur early and lead to quick rejection of a proof. One might hypothesize, based on evidence from eye-movement studies, that validation would prompt more back-and-forth movements to check the validity of individual deductions (Inglis & Alcock, 2012). But the same evidence indicates that this behavior might vary systematically with mathematical experience, and comprehension could involve precisely the same kind of checking. One might hypothesize that more experienced mathematical readers would be more influenced by task instructions, because they are better able to employ their skills in differentiated ways depending on strategic goals. But one might just as easily argue that experts would have a practised approach to mathematical reading that is always critical and less influenced by task demands.

So, while potential differences between validation and comprehension have rightly been taken seriously—researchers have clarified which they are studying, and avoided conflating the two or generalizing results about one to claims about the other (e.g., Weber, 2010)—they remain theoretical. Differences have not been empirically established by direct investigation or by indirect inference: results from studies on validation and comprehension are not obviously inconsistent. This uncertainty hampers both pedagogical design and theoretical progress. In the classroom, if validation and comprehension engender different reading behaviors, perhaps both should be included in student activities. If they engender essentially the same reading behaviors, perhaps instructors could safely prioritize according to other considerations. In research, if validation and comprehension involve different behaviors, then researchers should take care to avoid conflating the two. If they are essentially the same, then we could use findings from validation studies to reach conclusions about comprehension and vice versa. The plausibility of the opposing viewpoints therefore leads us to ask a straightforward research question: do proof validation and proof comprehension engender different reading behaviors?

Methodology

The structure of our research question places two key demands upon study design. First, in common with all empirical studies, it requires operationalizing the relevant theoretical constructs—in this case 'reading behaviors'—and deciding upon measures that can meaningfully capture relevant evidence. Second, it requires an analytical approach capable of treating 'reading behaviors are different' and 'reading behaviors are the same' as symmetric possibilities. In this section, we discuss our decisions with respect to each demand and provide detail on the approaches we used.

Study Design Decision: Eye-Movement Approach

Studies of mathematical reading have operationalized 'reading behaviors' in various ways, using designs that render different aspects of reading observable. Questionnaire studies (e.g., Weinberg, Wiesner, Benesh & Boester, 2012) have typically focused on macro aspects of mathematical reading, including decisions about what to read when and the intentions behind those decisions. This approach provides information about global habits and memorable reasons for conscious decisions, though not about the detail of behaviors during specific reading attempts. Interview studies (e.g., Shepherd and van de Sande, 2014) have typically focused on micro aspects of mathematical reading, including the speed and order in which individuals read mathematical texts. This approach permits direct observation of reading attempts, including behaviors such as pausing to check understanding. With concurrent or retrospective verbal reporting, it also permits access to conscious aspects of cognitive processes. However, interpretations of verbal reports are vulnerable to concerns about reactivity and veridicality (Russo, Johnson & Stephens, 1989), and research has shown that even expert mathematical readers do not always report accurately on behaviors such as attention shifting (Inglis & Alcock, 2012). Eye-movement studies (e.g., Hodds, Alcock and Inglis, 2014) have typically focused on physical reading behaviors, recording positions and durations of eye *fixations* in order to capture behaviors such as shifts in attention location. This approach provides no information on conscious thought processes, but permits very detailed analyses of aspects of reading that are only partially under conscious control (Rayner, 2009).

For our research question, these considerations make an eye-movement approach the most appropriate. Behavior differences in response to validation and comprehension tasks could well be both subtle and of types that are not open to accurate verbal reporting or are not fully under conscious control. We would not, for instance, expect a participant to be able to articulate why they felt compelled to re-examine an earlier line of a proof, or even to recall later that they did so. Because of this, and because the literature does not provide compelling hypotheses about expected differences between validation and comprehension, we also considered it important to improve our

chance of detecting differences by using a wide variety of eye-movement measures. We discuss eye-tracking methodology in general in this section and specify the measures we used in the Method section.

Eye-movement studies use eye-tracking technology to monitor and record participants' eye movements in real time. Contemporary eye trackers permit participants to sit comfortably in front of what looks like an ordinary computer screen. Tracking requires initial calibration: each participant is asked to follow a dot around the screen with their eyes in order to check for adequate recording (e.g. Tobii Technology, 2010). But, once this is done, material can be displayed on the screen and the participant can interact with the computer with no awareness of the ongoing tracking. Eye-movement studies have long been used to illuminate cognitive processes that underlie reading and more general visual search (e.g. Just & Carpenter, 1980; see Rayner, 2009, for a more recent review). With improvements in technology, eye-movement studies are now becoming more common in research in education in general (Was, Sansosti & Morris, 2017), and in mathematics education in particular (Hodds, Alcock & Inglis, 2014; Obersteiner and Tumpek, 2016).

Quantitative eye-movement analyses are facilitated by the fact that, when reading static materials, eye movements are not 'smooth'. Although readers might experience their eyes as sliding smoothly across a page, eye movements in fact form a series of short stops called *fixations* linked by very rapid movements called *saccades*. In ordinary reading in English, fixations typically last around 225-250 milliseconds, saccade lengths typically take in around 7-9 letter spaces, and around 10-15% of saccades are *regressions*, moves right to left or to previous lines (Rayner, 2009).

Fixation positions and durations provide information that can be analyzed in various ways. Fixation positions can be used to measure relative attention to different features of a screen. Inglis and Alcock (2012), for instance, set up areas of interest (AOIs) to distinguish words in proofs from mathematical symbols. They found that, compared with mathematicians, undergraduate students devoted proportionately less attention to words and more to symbols. Consecutive fixation positions can be used to infer information about reading or reasoning strategies. Obersteiner and Tumpek (2016), for instance, investigated adults' strategies for fraction comparison tasks. They found that for fraction pairs with common numerators or denominators, adults were more likely to compare components across the fractions, and for pairs without they were more likely to use a holistic strategy. Finally, fixation durations, pupil dilation, and number or proportion of regressive saccades can be used as proxy measures for the processing demand associated with a task. This is because it is well established that for any individual reader, all three increase when tasks are more demanding, for instance as text becomes more conceptually difficult (Jacobson & Dodwell, 1979). These differences can be quite substantial. Roy, Inglis and Alcock (in press) reported average mean fixation durations during proof comprehension of 250-350ms; the top

end of this is considerably higher than would be expected for ordinary written English, reflecting the complexity of this task.

The fact that eye-movement studies use real-time behavioral measures means that they circumvent some unavoidable problems associated with think-aloud protocols. When participants report concurrently on thought processes, reporting demands cognitive resource and so necessarily interferes with reading; when participants read as normal and subsequently reflect, they might not accurately recall their thoughts and intentions (Russo, Johnson & Stephens, 1989). Eye-tracking can be considered a complementary method in that it provides no information on conscious experience, but an accurate real-time record of effects of cognitive activity. Eye movements are not under detailed conscious control: a reader might deliberately return her attention to the top of a page, but she does not decide precisely where her eyes will 'land' or how long the landing fixation will be. Nevertheless, eye movements are under *cognitive* control. For instance, readers do not consciously decide their saccade lengths, but saccade lengths vary in response to fixation-by-fixation variations in the size of a 'window' showing only a restricted amount of text (Rayner & Pollatsek, 1981). Similarly, although readers might be aware that a text seems challenging, they do not consciously alter their fixation durations. Nevertheless, fixations durations are systematically influenced by word frequency, word predictability, number of meanings, age of acquisition, phonological properties, semantic relations, and word familiarity (Rayner, 2009, p.1472). Eye movement data can thus be used to infer information about the cognitive processing prompted by different tasks.

Study Design Decision: Bayesian Analyses

The second demand upon study design arises because our research question is open in the sense that we did have strong pre-existing hypotheses about differences we might find. This restricts the utility of traditional null hypothesis significance testing (NHST) because a 'no difference' outcome cannot be meaningfully interpreted: if NHST finds no significant difference, this could mean that there is no difference to be found or that there is a difference but its effect is too small (relative to noise in the data) to be detected by the test. In these circumstances, we did not wish to test the hypothesis that reading behaviors differ in specific ways; rather, we wished to evaluate whether the evidence generated was more consistent with the hypothesis that reading behaviors are different or the hypothesis that they are the same. Because of this, we adopted a Bayesian approach to our analyses. Unlike the logic forced by NHST (given the hypothesis, how likely is this evidence?), the logic of a Bayesian approach is that which researchers typically want to employ (given this evidence, how likely are the null and alternative hypothesis?).

Bayesian analyses thus permit symmetric treatment of the hypotheses that differences do and do not exist. One way in which this can be accomplished is by using *Bayes factors* to quantify the relative evidence provided

by some data for one hypothesis compared to another. Bayes factors are commonly notated B_{10} and B_{01}, where 0 refers to the null hypothesis and 1 to an alternative hypothesis. They provide information not on how likely a hypothesis is in an absolute sense, but on the extent to which a set of data supports each of the hypotheses. For instance, if an analysis returns a Bayes factor $B_{10} = 5$, then the evidence is five times more supportive of the alternative hypothesis than of the null, and we should update our prior beliefs—whatever these are—accordingly. This particular outcome, naturally, is equivalent to a Bayes factor for the null of $B_{01} = 1/B_{10} = 0.2$. If an analysis finds that $B_{01} = B_{10} = 1$, then the data do not provide evidence either way, and we should retain our prior beliefs (e.g., Dienes, 2016; Morey, Romeihn, & Rouder, 2016).

One challenge is that the Bayesian approach requires the researcher to specify two hypotheses, the null and the alternative. The null will typically be the hypothesis that the level of a condition factor does not predict the dependent variable, which means that a model including the condition as a predictor would offer a worse fit than a model not including the condition. The alternative hypothesis is often harder to specify. Rouder, Morey, Verhagen, Swagman, and Wagenmaker (2016) recommended defining the alternative hypothesis using a distribution on the effect size associated with the condition parameter for whichever statistical model is used. This means that the alternative hypothesis is that condition has some effect on the outcome, and that the effect is from some distribution, defined by the so-called *prior width*. This prior width captures the range of expected effects under the alternative hypothesis (see Rouder et al.'s discussion for detail), thus permitting meaningful interpretations of whatever difference or lack thereof is found.

Method

Design

Having decided on the overall methodological approach, we also designed our study to account for known variations in reading behaviors. Because there are typically individual differences on eye-movement measures (Rayner, 2009), we used a within-subjects design in which participants engaged with both validation and comprehension. Because behaviors might vary systematically with mathematical experience, we used an expert-novice design in which some participants were undergraduate students and others were professional mathematicians. Because there might be interference between tasks—completing a validation task might influence behavior on a subsequent comprehension task, or vice versa—we used a counterbalanced design in which participants completed both tasks in controlled orders, and we report not only the within-subjects analysis but also a separate between-subjects analysis of behavior on participants' first tasks.

Participants, Data Collection and Materials

We recorded the eye movements of 16 mathematicians (5 female and 11 male) and 16 undergraduate students (7 female and 9 male). The mathematicians were all active researchers in the Department of Mathematical Sciences at Loughborough University. The undergraduates were second- and third-year students at the same institution; all had taken courses on analysis and number theory covering the mathematical material they needed for this study. Each participant received an £8 inconvenience allowance for taking part, and data were collected in a quiet Eye Tracking Laboratory using a Tobii T120 eye tracker set to record at 120Hz. Participants attended individually; they were welcomed, asked to sign an informed consent form, and informed that after the eye-tracker was calibrated they would see screens showing two different proofs. They were informed that prior to each proof, an instruction screen would appear, and that they could move to the next slide by clicking the left mouse button. The researcher then left the room and sat outside near the door.

All participants read the same two proofs P_a and P_n. Proof P_a was analytic and identical to that used by Inglis and Alcock (2012) and by Inglis, Mejía-Ramos, Weber and Alcock (2013). Proof P_n was number theoretic and was based on exercise I.2.15 in Scheid (2002), a book written for prospective primary mathematics teachers. We selected these proofs because they did not have outright errors but did have gaps or nontrivial assertions that were not explicitly justified. This was important for our design because we wanted proofs that could reasonably be used in a comprehension task, but that were not obviously valid and that would thus be expected to prompt authentic validation behavior. Both proofs were formatted so as to be likely to yield reliable eye-movement recordings, using a standard font size but larger space than usual between lines. Both can be found in the appendix;

they also appear in their original format, along with the eye-movement data, at https://doi.org/10.6084/m9.figshare.5217418.v1.

For each proof, participants received instruction I_v or I_c. Instruction I_v said:

'On the next screen you will see a proof that was submitted to the Mathematical Gazette, a recreational mathematics journal. The editor has asked you to review the proof. Your task is to read it and decide whether or not you think it is valid. Take as long as you need to reach this decision. When you are ready to state your response, press the mouse button.'

After reading the corresponding proof, participants saw answer screen A_v asking them to click one of two buttons to state whether or not the proof was valid. Instruction I_c said:

'On the next screen you will see a proof that was published in the Mathematical Gazette, a recreational mathematics journal. A student has said that they do not understand the proof and that they need your help. Your task is to read and understand the proof so that you can explain it to them. When you have finished reading you will be asked a couple of questions about it. Take as long as you need. When you are happy that you have understood the proof, press the mouse button.'

After reading the corresponding proof, participants saw answer screen A_c asking them to click one of five buttons to indicate what kind of proof they had just read, where the choices were: A proof by induction; A contradiction proof; A direct proof; All of the above; None of the above. The tasks were ordered using a counterbalanced design in which the proofs, instruction screens and answer screens were arranged to give four different sequences:

(1) I_v-P_a-A_v-I_c-P_n-A_c; (2) I_c-P_a-A_c-I_v-P_n-A_v; (3) I_v-P_n-A_v-I_c-P_a-A_c; (4) I_c-P_n-A_c-I_v-P_a-A_v.

Four mathematicians and four undergraduates saw each sequence.

Dependent Measures and Data Processing

We used eight dependent measures: dwell time (raw), dwell time (%), mean fixation duration (raw), mean fixation duration (%) pupil dilation (%), number of regressive saccades, time to first fixation (%) and number of between-line saccades. Because prior research has demonstrated that reading behaviors vary depending on local aspects of a proof (Inglis & Alcock, 2012), we calculated each measure per participant per line of each proof. This means that for each participant, we had 16 datapoints per dependent variable.

Dwell time captures the total time the participant spent fixating on a given line, where we analyzed both raw time (measured in seconds), and percentage of the participants' total reading time for that line. Dwell time is normally used to compare how participants allocate their attention (e.g., Mello-Thoms, Nodine & Kundel, 2002).

Mean fixation duration captures the average length of all of a participant's individual fixations on a given line. Again, we analyzed both a raw figure (measured in milliseconds) and, since there may be individual

differences in fixation durations, percentage of the individual's overall average fixation duration. As noted in the Methodology section, longer mean fixation durations indicate increased difficulty in processing information at a location, or that the location is more engaging in some other way (e.g., Just & Carpenter, 1976).

Pupil dilation captures the size of the participants' right pupil while fixating on the given line. Since pupil dilation can vary depending on, for example, coffee intake, we considered pupil dilation on each line only as a percentage of the participant's overall pupil dilation during the study. Pupil dilation provides another index of cognitive effort, where higher dilations indicate higher effort (e.g., Pomplun & Sunkara, 2003).

Percentage of regressive saccades is a count of the number of saccades with direction between 177 and 183 degrees (where 0 degrees represented a horizontal saccade to the right). This measure was designed to capture the number of times a participant moved their attention backwards along a line. Regressive saccades are believed to indicate that the participant has had difficulty in processing the text, or that some misunderstanding has taken place (e.g., Sibert, Goturk, & Lavine, 2000).

Number of between-line saccades is a count of the number of times a participant moved their attention between different lines of a proof. In mathematical reading, experts and more successful readers make greater numbers of between-line saccades (Hodds et al., 2014; Inglis & Alcock, 2012).

Finally, *time to first fixation* captures the time at which a participant first fixated on a given line, expressed as a percentage of the participant's whole reading attempt. When appropriately combined, times to first fixation capture the extent to which participants read through the text in its printed order (Inglis & Alcock, 2012).

Inadequate eye-movements recording (commonly due to exaggerated head movements and here meaning that fewer than 70% of samples were recorded accurately) meant that we excluded four mathematician and two student participants from the analyses; two of the mathematicians and one of the students completed the same task sequence, but the others were all from distinct task sequence groups. Because our goal was to establish whether validation and comprehension involve different reading behaviors, our primary independent variable was condition (validation, comprehension). However, because prior research has also demonstrated that mathematicians and undergraduates read differently (Inglis & Alcock, 2012), we also included group membership as a factor in all of our analyses.

Results

Descriptive Statistics

Figure 1 shows the averages across all participants for six of our eight dependent measures (for dwell times and mean fixation durations it shows raw times only; graphs for the percentage versions of these measures show similar patterns). Each graph shows the mean figure across participants for each line of the proofs in the two conditions. These graphs show minimal differences between the validation and comprehension conditions.

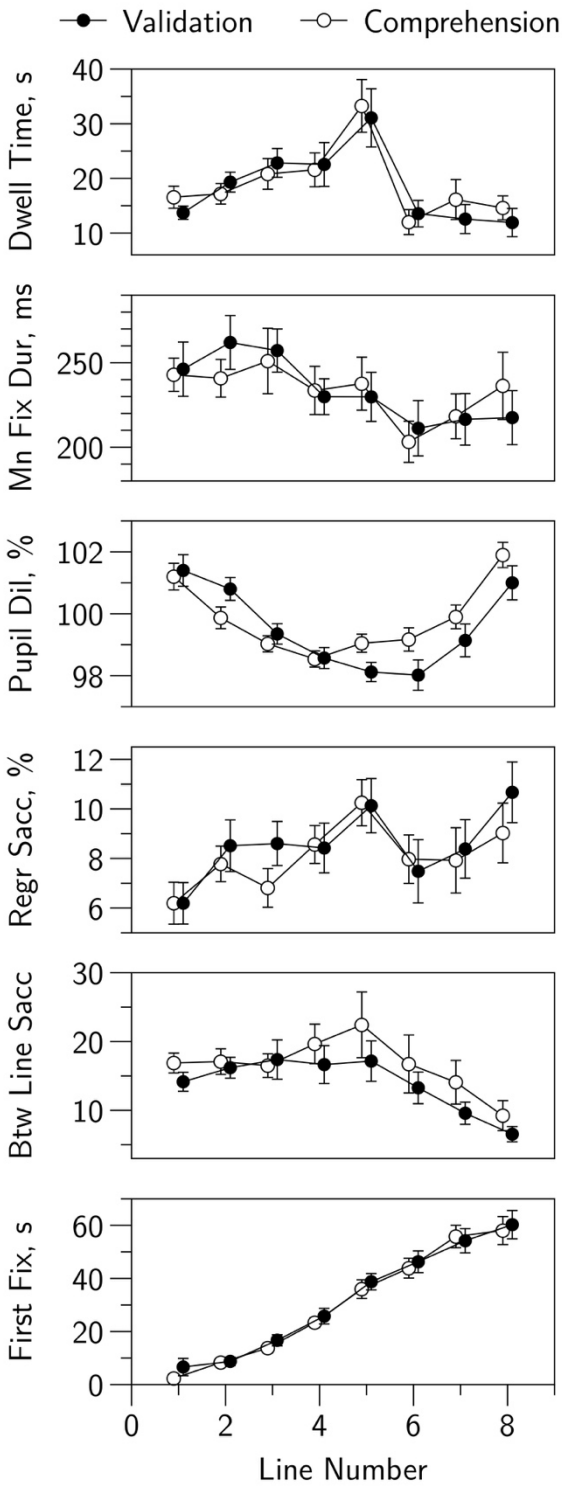

Figure 1: The means for six of the eight dependent measures, for each line, separately for each condition. Error bars show ±1 SE of the mean.

Within-Subjects Analyses

Our Bayesian analyses comprised a series of Bayesian ANOVA model comparisons for the eight different dependent measures. Conducting this analysis requires that alternative models to the null are specified using a distribution of possible effect sizes for each predictor. More concretely, the width of these distributions must be specified in advance by the researcher. Rather than specify these widths subjectively, we used Rouder et al.'s (in press) default prior width of $r = 0.5$ throughout. This captures the range of effect sizes typically found in the behavioral sciences, and ensures that the results from a two-groups Bayesian one-way ANOVA would be identical to that from a Bayesian *t* test with its default prior width of $1/\sqrt{2}$ (Wagenmakers et al., in press). However, our results and interpretations are not strongly sensitive to this choice of prior, within a reasonable range (we also conducted analyses for prior widths of $r = 0.3$ and $r = 0.7$ and found an almost identical pattern of results).

Consider the first dependent measure, Dwell Time (raw). We set up a null model that predicted each participant's raw dwell time on each line of both proofs using three predictors: Line (the line number within the proof), Group (mathematician or undergraduate) and the Line × Group interaction effect. We then compared this null model to a series of alternative models, each of which included Condition (validation or comprehension) as a predictor. The first comparison was between the null model (Line, Group, Line × Group) and the null model with Condition (Line, Group, Line × Group, Condition). Table 1 shows that this comparison yielded a Bayes factor in favor of the null of $B_{01} = 8.710$, meaning that our data were nearly nine times more consistent with the null model than with this alternative model. This provides strong evidence that raw dwell times do not depend on whether participants were reading for validation or for comprehension. We then compared the null model to four further alternative models, which were defined by Condition and various interaction effects that included Condition. These comparisons, along with their equivalents for the other seven dependent measures, appear in Table 1.

Model	Dwell Time (raw)		Dwell Time (%)		Mean Fix Dur (raw)	
	B_{10}	B_{01}	B_{10}	B_{01}	B_{10}	B_{01}
Null model	1	1	1	1	1	1
Condition	0.124	8.710	0.120	8.944	0.110	9.369
Condition + Condition × Line	0.002	636.2	0.002	532.0	0.004	250.8
Condition + Condition × Group	0.046	19.54	0.015	56.50	0.017	50.63
Condition + C×G + C×L	0.001	1438	0.000	3598	0.001	1596
Condition + C×G + C×L + C×G×L	0.132	7.763	1.065	0.954	0.000	8335

Model	Mean Fix Dur (%)		Pupil Dilation (%)		Regressive Saccades (%)	
	B_{10}	B_{01}	B_{10}	B_{01}	B_{10}	B_{01}
Null model	1	1	1	1	1	1
Condition	0.113	9.743	0.161	6.206	0.213	4.694
Condition + Condition × Line	0.003	370.7	0.132	7.554	0.004	276.1
Condition + Condition × Group	0.017	66.81	0.032	31.14	0.041	24.61
Condition + C×G + C×L	0.000	2563	0.024	41.76	0.001	1483
Condition + C×G + C×L + C×G×L	0.000	19k	0.001	1031	0.000	35k

Model	Between Line Sacc		Time to First Fix (%)	
	B_{10}	B_{01}	B_{10}	B_{01}
Null model	1	1	1	1
Condition	3.346	0.307	0.038	3.543
Condition + Condition × Line	0.050	19.53	0.002	422.8
Condition + Condition × Group	0.564	1.971	0.045	22.00
Condition + C×G + C×L	0.008	130.0	0.000	2700
Condition + C×G + C×L + C×G×L	0.001	722.4	0.000	77k

Table 1: Results of eight within-subjects Bayesian ANOVA analyses, with dependent variables dwell time (raw), dwell time (%), mean fixation duration (raw), mean fixation duration (%), pupil dilation (%), regressive saccades (%), between-line saccades, and time to first fixation (%). The null model includes within-subjects factors Line, between-subjects factor Group, and Line × Group interaction effect.

Table 1 shows that our data were more consistent with the hypothesis that reading for validation and comprehension are similar than with the hypothesis that they differ (at least as captured by our eight dependent measures). Of the 40 Bayes factors calculated, only two had $B_{10} > 1$ and thus indicated more support for the alternative hypothesis than the null. One of these had Bayes factor B_{10} = 1.065, indicating that the data did not substantially favor either the null or the alternative hypothesis. The other had Bayes factor B_{10} = 3.346, indicating evidence that reading condition influenced the number of between-line saccades. Inspecting the graph in Figure 1 indicates that participants made slightly more between-line saccades when reading for comprehension than for validation. Other than this difference, we found strong support for the hypothesis that reading for validation and comprehension engender similar reading behaviors. Moreover, greater support for the null hypothesis was found across those models involving interactions. If there had been evidence for a model including a Condition × Group interaction, this would suggest a difference between validation and comprehension for one group—experts or students—but not the other. We thus found evidence against such a difference.

Between-Subjects Analysis

The above analyses were conducted on both validation and comprehension reading attempts for each participant. But we were concerned that the study design might have led participants to read in artificially similar ways across the conditions: perhaps participants who first read for comprehension went on to behave similarly when reading for validation (or vice versa). To investigate this possibility, we conducted a second analysis considering only the first reading attempt for each participant (validation for half of the participants, comprehension for the other half). Our analytical approach was identical to that reported above, except that this time Condition was a between-subjects factor. The results of these analyses appear in Table 2.

Model	Dwell Time (raw)		Dwell Time (%)		Mean Fix Dur (raw)	
	B_{10}	B_{01}	B_{10}	B_{01}	B_{10}	B_{01}
Null model	1	1	1	1	1	1
Condition	0.886	1.129	0.201	4.964	1.081	0.925
Condition + Condition × Line	0.160	6.254	0.026	38.40	0.142	7.037
Condition + Condition × Group	0.316	3.166	0.052	19.26	1.147	0.872
Condition + C×G + C×L	0.051	19.48	0.007	148.3	0.134	7.445
Condition + C×G + C×L + C×G×L	0.012	85.89	0.002	532.1	0.045	22.10

Model	Mean Fix Dur (%)		Pupil Dilation (%)		Regressive Saccades (%)	
	B_{10}	B_{01}	B_{10}	B_{01}	B_{10}	B_{01}
Null model	1	1	1	1	1	1
Condition	0.203	4.917	0.241	4.146	0.435	2.301
Condition + Condition × Line	0.025	39.78	0.014	69.45	0.016	64.16
Condition + Condition × Group	0.051	19.74	0.084	11.96	0.196	5.098
Condition + C×G + C×L	0.007	150.5	0.008	121.4	0.007	146.1
Condition + C×G + C×L + C×G×L	0.002	572.1	0.000	1279	0.001	1220

Model	Between Line Sacc		Time to First Fix (%)	
	B_{10}	B_{01}	B_{10}	B_{01}
Null model	1	1	1	1
Condition	0.458	2.182	0.316	3.163
Condition + Condition × Line	0.018	55.41	0.013	77.63
Condition + Condition × Group	0.338	2.960	0.179	5.595
Condition + C×G + C×L	0.013	78.00	0.007	150.1
Condition + C×G + C×L + C×G×L	0.003	290.4	0.001	1520

Table 2: Results of eight between-subjects Bayesian ANOVA analyses, with dependent variables dwell time (raw), dwell time (%), mean fixation duration (raw), mean fixation duration (%), pupil dilation (%), regressive saccades (%), number of between-line saccades, and -time to first fixation (%). The null model includes within-subjects factor Line, between-subjects factor Group, and Line × Group interaction.

As with the within-subjects analyses, the between-subjects analyses provided strong support for the hypothesis that reading for validation and comprehension engender similar behaviors. Just two returned Bayes factor with $B_{10} > 1$, and both of these were very close to 1. Importantly, unlike the within-subjects analyses, these analyses provided evidence in favor of the null model for the between-line saccades measure.

Discussion

We began with the research question: do proof validation and proof comprehension involve different reading behaviors? The answer based on the analyses above is a rather clear 'no'. In almost all of our 80 model-building Bayesian analyses, the evidence was more consistent with the hypothesis that reading for validation and comprehension are similar than with the hypothesis that they differ. This applied across measures of dwell time as an indicator of attention allocation; mean fixation duration, pupil dilation and regressive saccades as indicators of cognitive demand; and between-line saccades and time to first fixation on each line as indicators of order in the mathematical reading process. It also applied across models involving a Condition × Group interaction, which means that we found no evidence of a difference in validation and comprehension for one group (experts or novices) but not the other. The one exception to this overall picture was that our within-subjects analysis provided more support for the hypothesis that comprehension elicits more between-line saccades. Taken at face value, this might be considered surprising: because validation involves checking deductions, we might have expected the opposite. However, considered alongside our between-subject analyses, we believe that this result should be treated with caution. The absence of an effect in the between-subjects analysis, together with the lack of evidence for other differences, means that we would not be confident that this effect would replicate. Overall, while this study does not allow us to conclude in an absolute sense that comprehension and validation use identical moment-by-moment reading processes, the Bayesian analyses do tell us to increase our confidence that this is the case.

From a theoretical standpoint, this lack of difference is perfectly plausible. Perhaps (at least for readers at these levels) the majority of validation effort is actually directed at comprehension, because comprehension must be attained before a validity judgement can be made. Or perhaps a sincere comprehension attempt provides validation 'for free', because good comprehension would flag up invalid inferences. This does not mean that comprehension or validity judgements will necessarily be accurate, because they depend on existing knowledge and on ability to link the new argument to that knowledge; even if reading behaviors are identical in experts and novices, less experienced readers would be less likely to attain good comprehension and more likely to miss logical errors. Alternatively, it could be that regardless of the instruction, validation and comprehension are intermingled. Perhaps most or all mathematical reading attempts (at this level) involve switching back and forth between comprehension and validation, so that readers routinely do both even within individual lines of a proof. In any of these cases, we see no theoretical surprise in the finding that reading for validation and comprehension engender the same behaviors.

From a methodological standpoint, this conclusion could be considered more debatable, in ways that might depend upon a researcher's philosophical position. As noted in the Methodology section, eye-movement

analyses render some aspects of reading observable by focusing on quantitative measurements of physical behaviors. But they provide no direct information about conscious experience, meaning that researchers interested in that experience—in readers' rationales for reading in certain ways or judgments about their own understanding—will wish additionally or alternatively to use complementary methods such as questionnaires and interviews. At least some researchers have explored this option, asking participants to retrospectively view their own eye-tracking scanpaths and to report on what they recall of their thinking (e.g., Bax, 2013). Moreover, eye tracking conceivably fails to pick up other relevant non-conscious cognition. It thus remains possible that although all outwardly observable behaviors are identical, validation and comprehension engender different cognitive activity. We consider this account unlikely and certainly unparsimonious: the argument would be that two activities that involve essentially identical attention allocation, cognitive demand and order of processing nevertheless result from significantly different underlying activity. But that account is not refutable based on these data, so it remains possible that one or other task leads to different outcomes. If so, then creative use of an alternative methodological approach could demonstrate this—perhaps, for instance, groups given validation or comprehension tasks for learning purposes could be tested on subsequent transfer tasks.

For the time being, the apparent lack of difference between validation and comprehension provides encouraging information for both researchers and instructors. In research terms, if Mejía-Ramos & Inglis (2009) were to be believed, then we would not be able to use studies of validation to draw conclusions about comprehension or vice versa. But if our conclusions here are correct, then it seems reasonable to infer that consistent results from validation and comprehension studies arise due to these tasks engendering similar cognitive processes, so that such generalizations are sensible. For example, studies of validation (Selden & Selden, 2003; Weber, 2010) have found that students often judge apparently deductive arguments to be valid because they do not recognize their logical flaws; it seems safe to infer that similar students would not recognize logical flaws in comprehension tasks either. Similarly, studies of comprehension have found that students pay less attention than experts to comprehension failures (Shepherd & van de Sande, 2014). It seems safe to infer that they would behave similarly in validation tasks, and indeed studies of validation have revealed that students are sometimes willing to make validation judgements even when they know that they do not fully understand a proof (Weber, 2009). This provides researchers with a potential methodological advantage because comprehension tests are difficult and time-consuming to construct (Mejía-Ramos, Lew, de la Torre & Weber, in press). When studying comprehension is the main goal, this might be time well spent. When a comprehension task is used in service of some other goal, then a validation task—requiring a simple valid/invalid judgement—might do just as well.

In the classroom, similar practical implications follow. Instructors might well wish to promote effective mathematical reading (Conradie & Frith, 2000; Kasman, 2006), and to use specific tasks both to teach reading skills and to test the success of their efforts. Our results suggest that they could choose between validation and comprehension tasks and be relatively confident that these will not lead to different reading behaviors, though the approaches would incur different preparation costs. A stand-alone comprehension task could be set with minimal preparation and run in a responsive way, perhaps involving group discussion. But in many cases an instructor might want to assess comprehension more formally, which involves decisions about what specific questions to ask (Conradie & Frith, 2000) or creative approaches to marking open-ended comparison or critiquing tasks (Jones & Alcock, 2014; Kasman, 2006). A stand-alone validation task could be used with similarly minimal preparation, but a sequence of such tasks would need to include invalid arguments so that the task was genuine; such arguments are not commonly offered in textbooks so would need to be constructed. Moreover, in many cases an instructor might wish to know not just that a validation judgement was incorrect, but whether and why the reading attempt failed. Again, this could be addressed informally in a classroom situation, but presents more challenges for formal evaluation of individual students' reading.

Thus there remain numerous questions regarding whether and how instructors and researchers might profitably build on the work presented here. But our study is the first to investigate directly whether validation and comprehension engender different processes. The field had, in response to suggestions made by Mejía-Ramos and Inglis (2009), acted with care upon the reasonable assumption that they might. Our study presents evidence that, in fact, they do not.

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Ball, D. L., & H. (2000). Making believe: The collective construction of public mathematical knowledge in the elementary classroom. In D. Phillips (Ed.), *Yearbook of the National Society for the Study of Education, Constructivism in Education* (pp. 193-224). Chicago, IL: University of Chicago Press.

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing*, *30*, 441–465.

Brown, S. A. (2014). On skepticism and its role in the development of proof in the classroom. *Educational Studies in Mathematics*, *86*, 311–335.

Conradie, J., & Frith, J. (2000). Comprehension tests in mathematics. *Educational Studies in Mathematics*, *42*, 225–235.

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

Hayward, C. N., Kogan, M., & Laursen, S. L. (2016). Facilitating instructor adoption of inquiry-based learning in college mathematics. *International Journal of Research in Undergraduate Mathematics Education*, *2*, 59–82.

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

Hodds, M., Alcock, L., & Inglis, M. (2014). Self-explanation training improves proof comprehension. *Journal for Research in Mathematics Education*, *45*, 62-101.

Inglis, M., & Alcock, L. (2012). Expert and novice approaches to reading mathematical proofs. *Journal for Research in Mathematics Education*, 43, 358-390.

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

Jacobson, J. Z., & Dodwell, P. C. (1979). Saccadic eye movements during reading. *Brain and Language*, *8*, 303–314.

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education*, *39*, 1774-1787.

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology, 8*, 441-480.

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

Kasman, R. (2006). Critique that! Analytical writing assignments in advanced mathematics courses. *Problems, Resources, and Issues in Mathematics Undergraduate Studies*, *16*, 1–15.

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

Larsen, S., Johnson, E., & Bartlo, J. (2013). Designing and scaling up an innovation in abstract algebra. *Journal of Mathematical Behavior*, *32*, 693–711.

Larvor, B. (2016). *Mathematical Cultures: The London Meetings 2012-2014*. Basel, Switzerland: Birkhäuser.

Lew, K., Fukawa-Connelly, T. P., Mejía-Ramos, J. P., & Weber, K. (2016). Lectures in advanced mathematics: Why students might not understand what the mathematics professor is trying to convey. *Journal for Research in Mathematics Education*, *47*, 162–198.

Mejía-Ramos, J.-P., Fuller, E., Weber, K., Rhoads, K., & Samkoff, A. (2012). An assessment model for proof comprehension in undergraduate mathematics. *Educational Studies in Mathematics*, *79*, 3–18.

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.

Mejía-Ramos, J. P., Lew, K., de la Torre, J., & Weber, K. (in press). Developing and validating proof comprehension tests in undergraduate mathematics. To appear in *Research in Mathematics Education*.

Mello-Thoms, C., Nodine, C. F., & Kundel, H. L. (2004). What attracts the eye to the location of missed and reported breast cancers? In *Proceedings of the Eye Tracking Research and Applications Symposium 2002* (pp. 111-117). NY: ACM Press.

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology, 72,* 6-18.

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

Obersteiner, A., & Tumpek, C. (2016). Measuring fraction comparison strategies with eye-tracking. *ZDM Mathematics Education*, *48*, 255–266.

Pomplun, M., & Sunkara, S. (2003). Pupil dilation as an indicator of cognitive workload in Human-Computer Interaction. In *Proceedings of HCI International 2003* (Vol 3, pp. 542-546). Mahwah, NJ: Lawrence Erlbaum Associates.

Rasmussen, C., Wawro, M., & Zandieh, M. (2015). Examining individual and collective level mathematical progress. *Educational Studies in Mathematics*, *88*, 259–281.

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

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

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, *62*, 1457–1506.

Rouder, J., N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E.-J. (in press). Bayesian analysis of factorial designs. *Psychological Methods.* http://dx.doi.org/10.1037/met0000057

Roy, S., Alcock, L., & Inglis, M. (in press). Multimedia resources designed to support learning from written proofs: An eye-movement study. To appear in *Educational Studies in Mathematics*.

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

Samkoff, A., & Weber, K. (2015). Lessons learned from an instructional intervention on proof comprehension. *Journal of Mathematical Behavior*, *39*, 28–50.

Scheid, H. (2002). Elemente der Arithmetik und Algebra. *Heidelberg, Berlin, Spektrum, Akad. Verl.*, 4. Auflage.

Selden, J. & Selden, A. (1995). Unpacking the logic of mathematical statements. *Educational Studies in Mathematics*, *29*, 123–151.

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 Edu- cation*, *34*, 4–36.

Shepherd, M. D., Selden, A., & Selden, J. (2012). University students' reading of their first-year mathematics textbooks. *Mathematical Thinking and Learning*, *14*, 226–256.

Shepherd, M. D, & van de Sande, C. C. (2014). Reading mathematics for understanding—from novice to expert. *Journal of Mathematical Behavior*, 35, 74–86.

Sibert, J. L., Gokturk, M., & Lavine, R. A. (2000). The Reading Assistant: Eye gaze triggered auditory prompting for reading remediation. In *Proceedings of the Thirteenth Annual ACM Symposium on User Interface Software and Technology* (pp. 101-107). NY: ACM Press.

Stylianides, A. J. (2007). Proof and proving in school mathematics. *Journal for Research in Mathematics Education*, *38*, 289-321.

Stylianides, A. J., Bieda, K. N., & Morselli, F. (2016). Proof and argumentation in mathematics education research. In Á. Gutiérrez, G.C. Leder, & P. Boero, (Eds.), *The second handbook of research on the psychology of mathematics education: The journey continues*, pp. 315–351. Rotterdam: Sense Publishers.

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

Tobii Technology (2010). *Tobii Eye Tracking: An Introduction to Eye Tracking and Tobii Eye Trackers*. Stockholm, Sweden: Tobii Technology AP.

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., ... & Meerhoff, F. (in press). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*.

Was, C., Sansosti, F., & Morris, B. (2017). Eye-tracking technology applications in educational research. Hershey, PA, USA: IGI Global.

Weber, K. (2009). Mathematics majors' evaluation of mathematical arguments and their conception of proof. In *Proceedings of the Twelfth SIGMAA on RUME Conference on Research in Undergraduate Mathematics Education*, http://mathed.asu.edu/crume2009/proceedings.html.

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

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

Weinberg, A., Wiesner, E., Benesh, B., & Boester, T. (2012). Undergraduate students' self-reported use of mathematics textbooks. *PRIMUS*, *22*, 152–175.

Yackel, E. & Cobb, P. (1996). Sociomathematical norms, argumentation, and autonomy in mathematics. *Journal for Research in Mathematics Education*, *27*, 458–477.

Yackel, E., Rasmussen, C., & King, K. (2000). Social and sociomathematical norms in an advanced undergraduate mathematics course. *Journal of Mathematical Behavior*, *19*, 275–287.

Yang, K.-L., & Lin, F.-L. (2008). A model of reading comprehension of geometry proof. *Educational Studies in Mathematics*, *67*, 59–76.

Zienes, D. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology, 72,* 78-89.

Appendix

Proof P_n

Theorem: For $n, k \in \mathbb{N}$, the equation $(n-k)! = n^k - 1$ has exactly three solutions:

$(n_1, k_1) = (2,1)$, $(n_2, k_2) = (3,1)$ and $(n_3, k_3) = (5,2)$.

Proof: Let $n \in \mathbb{N}$ and assume that there exists a solution with $n > 2$ and n even.

Because n is even, $(n-1)!$ is also even.

But $n^k - 1$ is odd for all $k \in \mathbb{N}$, so $(n-k)! \neq n^k - 1$ for all even $n > 2$.

Now assume that there exists a solution with $n > 5$ and n odd.

We know that $(n-1)|(n-2)!$, so it follows that $(n-1)^2|(n^k - 1)$.

By factorization of $(n^k - 1)$ we have $n^k - 1 = (n-1)(1 + n + n^2 + \ldots + n^{k-1})$.

On the one hand the second factor is divisible by $(n-1)$.

On the other hand by dividing this factor by $(n-1)$ we get the remainder k.

Thus we have shown that $(n-1)|k$, but this contradicts $n^{n-1} - 1 > (n-1)!$.

Proof P_a

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 $x = (1 + ym)^{\frac{1}{m}} = \left(1 + \frac{y^n}{n}\right) \to e^y$ 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 $\int x^{-1}dx = \ln(x) + c'$.