Mathematics Professors' Evaluation of Students' Proofs: A Complex Teaching Practice

Robert C. Moore¹

Published online: 12 May 2016

© Springer International Publishing Switzerland 2016

Abstract This article reports on an exploratory study of mathematics professors' proof evaluation practices and the characteristics they value in good proof writing. Four mathematicians were interviewed twice. In the first interview, they evaluated and scored six proofs of elementary theorems written by undergraduate students in a discrete mathematics or geometry course, assigned each proof a score out of 10 points, and responded to questions about the characteristics they value in a well-written proof and how they communicate these characteristics to students. In the second interview, they reevaluated three of the earlier proofs after reading the marks and comments written by the other professors, and evaluated and scored a seventh proof. The results reveal that proof grading is a complex teaching practice requiring difficult judgments about the seriousness of errors and the student's cognition that led to errors or other flaws. For five of the seven proofs the scores varied by at least 3 points, and the article discusses reasons for this variation. These professors agreed that the most important characteristics of a well-written proof are logical correctness, clarity, fluency, and demonstration of understanding of the proof. Although the professors differed in the attention they gave to fluency issues, such as mathematical notation, layout, grammar, and punctuation, they agreed in giving these characteristics little weight in the overall score.

Keywords Undergraduate mathematics · Proof evaluation · Proof grading · Proof writing · Teaching proof · Mathematics professors · Practice of mathematicians

Introduction and Research Questions

Mathematical proof is a fundamental component of the undergraduate mathematics curriculum for mathematics majors. To begin the process of teaching students the

✉ Robert C. Moore
moorer@andrews.edu

¹ Department of Mathematics, Andrews University, 4260 Administration Drive, Berrien Springs, MI 49104-0350, USA

deductive nature of mathematics and how to read and write proofs that meet an acceptable level of rigor and clarity, many colleges and universities offer transition-to-proof courses. Such courses commonly include basic logic and standard proof methods (e.g., direct, contradiction, contraposition, mathematical induction), along with material from a variety of mathematical areas, such as set theory, number theory, discrete mathematics, abstract algebra, and real analysis (e.g., Chartrand et al. 2013; Epp 2011; Smith et al. 2011). Early on in this transition to proof, students are asked to write short proofs based primarily on definitions, and over time they learn to construct longer, more complex proofs in courses like abstract and linear algebra, real analysis, and geometry. Furthermore, as students progress through transition courses and subsequent upper-level courses, they are expected to move beyond the notion of informal proofs that simply explain or convince someone to a more rigorous standard of proof in which proofs are clearly written deductive arguments that generally use conventional proof methods and are based on valid argument forms, specified axioms, precise definitions, and previously proved theorems.

Throughout this journey a standard means of teaching students to construct and write proofs is to require them to write proofs for homework, tests, and examinations. The professor grades these proofs by writing comments or other feedback on the papers along with a score and then returns the papers to the students. Thus, to a large extent by this process of writing proofs and receiving feedback from their teachers successful students gradually learn to write clear, correct, rigorous proofs. Surprising, given the importance of grading proofs and giving feedback as students develop their notions of proof and the ability to write proofs, virtually no research has examined this aspect of teaching, in particular the criteria mathematics professors have in mind as they grade students' proofs; the weight they give to logical, notational, grammatical, and other errors; the type of feedback they give students; and the consistency with which different professors grade proofs, assign scores, and write comments for improvement, thereby ensuring that students get consistent messages from their professors as to the nature of proof and the characteristics of a well-written, rigorous proof.

These considerations motivated the present study and the following main research questions:

1. Do mathematics professors agree in their evaluation and scoring of students' proofs?
2. What characteristics do they consider when evaluating students' proofs?
3. In what ways do they attempt to convey the characteristics of good proof writing to their students?

Related Research

Speer et al. (2010) discussed the dearth of research on undergraduate mathematics teaching and called for research studies on actual teaching practices. In particular, they noted that "we know very little about how collegiate teachers design assessments or evaluate students' work on them" (p. 111) and suggested that studies could reveal what teachers value in students' responses and how they try to communicate their values to students. Inglis et al. (2012) also noted that "the level of consistency in mathematicians' grading and instruction on what constitutes a proof is a useful avenue for future

research” (p. 70). The present study is a step toward addressing these calls for research on university mathematics professors’ proof assessment practices.

In addition to contributing to the literature on undergraduate mathematics teaching, the study builds on three more specific areas of undergraduate mathematics education research: proof assessment, proof validation, and proof presentation.

Proof Assessment

While considerable research attention has focused on students’ difficulties and errors in proof construction (e.g., Antonini and Mariotti 2008; Harel and Brown 2008; Harel and Sowder 1998; Moore 1994; Selden and Selden 1987, 1995; Thompson 1996; Weber 2001), little attention has been given to how mathematics professors assess students’ proofs and respond to students’ errors. Specifically, we do not know how professors assess whether a student understands a proof he or she has written, what kind of feedback they give to students about their errors, how they communicate the feedback to the students, and how they assign scores to students’ written work.

Mejía-Ramos and Inglis (2009) developed a framework for classifying proof-related activities. One activity they identified, in which undergraduate students engage, is that of demonstrating mathematical understanding by constructing and presenting an argument to an expert, usually their teacher. A complementary assessment activity the authors identified, which teachers routinely perform, is that of evaluating an argument against a set of criteria and making a judgment about how well the student has demonstrated an understanding of the argument. Both of these activities occur routinely in undergraduate mathematics classes as the students write proofs and the professors grade them. Although many studies have examined students’ proof constructions, research on how teachers evaluate students’ proof constructions is scarce. The present study contributes to the literature in this area because it addresses the assessment of students’ proof writing and errors, identifies ways in which professors provide feedback, and clearly shows that an important aspect of grading students’ proofs involves making judgments about the students’ level of understanding of the proofs and the associated mathematical concepts.

To address the need for instruments to assess students’ understanding of proofs in advanced mathematics courses, Mejía-Ramos et al. (2012) developed a model to assess students’ comprehension of a given proof. But the model was not intended to be used as a tool for efficiently grading and scoring proofs submitted by students for homework or tests, which is the kind of assessment instrument mathematics professors may find useful on an almost daily basis. To that end, Brown and Michel (2010) created an assessment rubric built on three characteristics: readability, validity, and fluency (RVF). The authors claimed that the rubric aids in the efficiency and consistency of evaluating students’ work, serves as a means of communicating to students the characteristics of good writing, and provides feedback for improvement. Since the rubric was based on Brown and Michel’s careful examination of students’ papers, as opposed to research or collaboration with other mathematicians, the question arises as to whether other mathematicians would agree with the rubric, particularly its list of three characteristics of good mathematical writing. In other words, if several mathematicians did not have the rubric and were asked to evaluate a student’s proof, would they evaluate it with the same criteria and score it the same way?

A century ago Starch and Elliott (1913) demonstrated that the answer to this question is “no” for high school mathematics teachers. When 138 mathematics teachers graded a student’s final examination geometry paper, the results showed “extremely wide variation of the grades” (p. 257) for the paper as a whole as well as for individual questions. While the study is a century old and did not focus on proof at the undergraduate level, its findings do raise the question of whether university professors today might also have widely different views on the evaluation of students’ mathematical work. The present study addresses this question.

Proof Validation

Perhaps the most important judgment required in the evaluation of a student’s argument is whether the argument is indeed a valid proof. Inglis et al. (2013) noted that a consequence of the deductive nature of mathematics is that “many in the mathematical community believe that the validity of a proof is not a subjective issue but an objective fact, and they cite the high level of agreement between mathematicians on what constitutes a proof as support for their position” (p. 271). If this claim is true, then mathematicians should show a high degree of consistency in judging the correctness of a student’s argument. On the other hand, others disagree that the validity of mathematical proofs is an objective fact. For example, Auslander (2008) suggested that “standards of proof vary over time and even among mathematicians at a given time” (p. 62). If so, then mathematicians may arrive at different judgments on the validity of a student’s argument and, consequently, very different grades.

Weber (2008) investigated mathematicians’ views on the validity of proofs and found disagreement among mathematicians about the validity of purported proofs, as did Inglis and Alcock (2012), who concluded that “some of these disagreements could be genuinely mathematical, rather than being relatively trivial issues related to style or presentation” (p. 379). In their study of 109 research-active mathematicians, Inglis et al. (2013) found further evidence that mathematicians have no universal agreement about what constitutes a valid proof by showing that many mathematicians judged a particular proof to be valid while many others judged it to be invalid. The authors concluded that mathematicians’ standards of validity differ and questioned whether students are getting a consistent message about what constitutes a valid proof.

Thus, we are led to suspect that mathematicians may differ in their judgments about the correctness of a student’s proof and, consequently, in their grading of a proof. Furthermore, given that they hold different viewpoints on validity, they probably differ on other aspects of proof writing, such as clarity, style, presentation, and the amount of detail students should include to justify steps within a proof. The present study examines these questions.

Proof Presentation

When grading a students’ proof, a mathematician may check the logical structure of the proof and look for mathematical errors, but may also check that it meets other characteristics of quality, such as directness, completeness, readability, or format. One way to get some insight into the characteristics that mathematicians consider important in good mathematical writing is to examine their own proof writing and presentation.

This, in turn, may give us insight into the features to which they attend when grading students' papers and the extent to which they communicate consistent messages to students about good mathematical writing.

Lai et al. (2012) reported on two studies in which they observed mathematicians as they revised two proofs. They found that although mathematicians often agree, occasionally they disagree to a remarkable extent over whether a proposed revision of a proof will improve the proof or make it worse for pedagogical purposes. They found broad agreement that a proof is better when introductory or concluding sentences clarify the proof framework, main ideas are highlighted by formatting, and unnecessary information is omitted so as not to distract or confuse the reader, but they had less agreement on whether a particular justification should be stated or left for the reader. In short, the mathematicians emphasized various ways to make a proof clear and readable for its intended audience. Note that these studies focused on proofs written by mathematicians for presentation to students. While it seems reasonable to expect that mathematicians value the same characteristics in students' writing that they espouse in their own writing and proof presentations, this claim has not been studied.

Given that mathematicians sometimes differ on the question of validity and other desirable features of written arguments, it is reasonable to expect that they may differ in their assessments of students' proofs, which require judgment calls on not only validity but also clarity, readability, the seriousness of errors, justification of steps, and other features. Furthermore, in scoring proofs professors must make judgments about the extent to which a student has demonstrated understanding of the proof and its associated mathematical concepts, and finally must decide on a score, which may involve partial credit. Thus, this evaluation process is more complex than making a judgment call on only the validity of a purported proof and, therefore, offers a rich opportunity for study.

Research Methodology

Participants

In the fall of 2012 I conducted individual interviews that lasted about one hour with four mathematics professors, three women and one man, all from a small private university. (For purposes of confidentiality, I will refer to all of them as though they were female.) Each professor has a Ph.D. in mathematics and is actively involved in research. The general areas of specialty included partial differential equations, differential geometry, dynamical systems and mathematical ecology, and abstract algebra. Each professor had at least 10 years of university-level mathematics teaching experience, including advanced undergraduate courses. Two of them, Professors A and D, have extensive experience teaching upper-division courses that emphasize the reading and writing of proofs, such as advanced calculus and linear and abstract algebra, and one, Professor B, has experience teaching a complex analysis course with relatively light emphasis on the reading and writing of proofs. The fourth professor, Professor C, was first trained in mathematical logic and then dynamical systems and mathematical ecology. She teaches primarily applied mathematics, such as ordinary differential equations and mathematical modeling, but emphasizes proofs in her calculus courses, once taught a transition-to-proof course,

and has extensive experience as a mathematics journal editor and reviewer. Table 1 provides a summary of the participants and their areas of expertise.

Interview Procedures

In the first part of the interview, the professors talked aloud as they evaluated six written proofs (see [Appendix](#)) by (a) marking on the proof what was correct or incorrect, (b) indicating how the proof could have been improved, and (c) assigning a score out of ten points to the proof. In the second part of the interview, they responded to questions about their beliefs and practices in evaluating students' proofs and teaching students to write proofs. These questions included:

- What are the characteristics of a well-written proof?
- When you evaluate and score a students' proof, what characteristics do you look for?
- Which characteristics are most important and carry the most weight in the overall score?
- How do you communicate these characteristics to students?
- Do you score homework proofs differently from test proofs?
- Have you ever used a proof evaluation rubric?

The first four of these questions were aimed directly at answering the main research questions. The fifth question sought to uncover whether the professors would alter their characteristics or grading practices on the basis of circumstances. For instance, a professor may grade homework proofs more leniently than test proofs because she views homework as a means to help students develop their knowledge and skills, whereas she views a test as a means to check for mastery. On the other hand, a professor may grade test proofs more leniently because students are under time pressure and have less time to polish their proofs. The aim of the last interview question was to determine whether the professors had ever written a list of characteristics of good proof writing and used such a list to improve their efficiency and consistency in grading students' proofs, as did Brown and Michel (2010).

Materials

The six proofs, Proofs 1–6, used for the first part of the interview were proofs of elementary theorems, which I labeled as Tasks 1–6 (see [Appendix](#)). The proofs were

Table 1 Summary of the participants' expertise

Professor	Area of specialty	Main upper-level courses taught	Experience teaching transition course
A	partial differential equations	advanced calculus partial differential equations	none
B	differential geometry	complex analysis	none
C	dynamical systems mathematical ecology	mathematical modeling	once
D	abstract algebra	abstract and linear algebra	none

written by my students: five proofs from tests in a discrete mathematics course, which serves as a transition-to-proof course for mathematics majors, and one homework proof from a geometry course. I rewrote the proofs in my own handwriting for three reasons. First, I wanted to conceal the identities of the students who wrote the proofs from the four participating professors. The students had taken classes from these professors, and two students in particular had handwriting that might have been recognized by the professors. I wanted to avoid any potential bias the professors might have when they evaluated the proofs. Second, the proofs were written in pencil, which was not sufficiently dark for scanning and printing on the special Livescribe paper that was needed for the interviews. Third, before this study was initiated I had written marks on the proofs when the students submitted them on their test and homework papers. My handwriting was a bit neater than the students' writing, but I laid out the proofs on the page just as written by the students and attempted to write them as much as possible like the originally written proofs, making only a few minor changes in surface features.

I chose these proofs for three principal reasons. First, they were authentic student productions. Second, they dealt with a variety of mathematical content and proof methods. Third, as undergraduate students' proofs are seldom perfect, often containing minor errors, such as those involving notation or phrasing, as well as more significant errors, such as logical errors, I wanted to see how the professors would respond to various types of errors and presentation styles and how much weight they would give to errors and other aspects of the proofs in their scoring. The six proofs represented a variety of errors and features that I judged to be both good and bad that related to readability, validity, fluency, clarity of reasoning, use of definitions, quantifiers, and surface features such as layout and punctuation (Brown and Michel 2010; Lai et al. 2012; Moore 1994; Selden and Selden 1995).

To be more explicit, here are some of the details I considered in choosing four of the proofs. Tasks 1 and 2, which accompany Proofs 1 and 2, are identical but the proofs are different. Proof 1 is laid out something like a two-column proof rather than in paragraph form. I wanted to know how the professors would react to the format in which it was written. Proof 2 has a number of deficiencies. First, it has capitalization and punctuation errors. Second, the first sentence is a statement of what needs to be shown in the proof rather than a part of the proof. Third, it contains two errors in line 1 (yRz rather than xRz at the end of the first sentence and $\mathbf{Z}$, the set of integers, rather than $\mathbf{R}$, the set of real numbers, at the end of the second sentence), and line 3 should say “for some” integers k and c rather than “let” k and c be integers. Fourth, it lacks readability and is laid out oddly on the page. I was interested to know how the professors would judge a proof that has much room for improvement and, in particular, how they would judge the seriousness of the errors. Proof 5 shows some understanding of the main ideas but lacks fluency in the use of mathematical language and notation. Proof 6 begins with logical errors in paragraph 1—the contrapositive is misstated and the student has confused contraposition with contradiction—but the heart of the proof in paragraphs 2 and 3 is based on a correct statement of the contrapositive. These six proofs provided opportunity for me to observe the features of students' proofs to which the professors attend, how they respond with marks and comments to various types of errors and surface features, and how much weight they give to errors and other features of students' arguments.

Data Collection and Analysis

The interviews were recorded with a Livescribe smart pen that recorded the professors' handwritten marks on the proofs as well as their talk-aloud commentary. I transcribed the interviews and analyzed them using an open coding system (Strauss and Corbin 1990). I began the analysis by making a detailed chart for each proof that summarized the professors' written grading of the proofs: corrections, marks, comments, and scores. As an example, Proof 2 appears in Fig. 1 and its initial analysis appears in Table 2.

Next, I read the transcripts carefully and highlighted sentences and paragraphs that revealed the characteristics to which the professors attended when evaluating the proofs and the rationales for their written marks, comments, and scores. I wrote codes in the margins of the transcript beside the corresponding portions of the transcripts. For example, my initial codes for evaluating proofs were as follows: readability (R), clarity (C), logic (L), beginning of the proof (B), math error (ME), justify (J), notation (N), fluency with math language and notation (FL), student's understanding (S), English mechanics (Eng), format and layout (F), context (C), and missing step (MS). As the analysis progressed, I began to see which codes occurred most often and seemed to be the most important to the professors. I subdivided logic into proof framework (L-pf)

Table 2 Initial analysis of the professors' evaluation of Proof 2

Location of deficiency	Professor A	Professor B	Professor C	Professor D
Sentence 1	Inserts "We want to prove" at the beginning		Writes "This is what we want to prove. It's an aside, not part of the proof."	Writes "need to show"
End of sentence 1		yRz should be xRz		yRz should be xRz
End of line 1	Changes Z to R	Changes Z to R	Changes Z to R	
Sentence 3, beginning at the end of line 2			"We know that ..." should be "We know $x-y=k$ and $y-z=c \exists k,c \in \mathbf{Z}$."	change "let" to "for some"
Lines 4-7	Rewrites: $x-z=k+y-(y-c)$ $=k+c \in \mathbf{Z} \Rightarrow xRz$	Give reasons for each step in the calculation of $x-z$	Calculation of $x-z$ is "okay for scratch work." Rewrites: "Thus, $x-z=(y+k)-(y-c)=k+c$."	
Line 9			Change "an Z " to "in Z "	Change "an Z " to "in Z "
Other			Writes "Proofs should be complete sentences."	

The professors made little comment about capitalization and punctuation, and none of them commented on the loop around $k+c$ and how the proof ran diagonally down the left-hand side of the page

Task 2: Transitive Relation

Define a relation R on the set of real numbers by $x R y$ if and only if $x - y$ is an integer, that is, two real numbers are related if they differ by an integer. Prove that R is transitive.

Proof 2:

if $x R y$ and $y R z$ then $y R z$ Let $x, y, z \in \mathbb{Z}$
 and assume $x R y$ and $y R z$. We know $x - y \in \mathbb{Z}$
 and $y - z \in \mathbb{Z}$ let $k, c \in \mathbb{Z}$

$$x - y = k$$

$$y - z = c \rightarrow y = c + z$$

$$x - (c + z) = k$$

$$x - z = (k + c)$$

since $k + c$
 is an $\mathbb{Z}$ then
 $x R z$ ■

Fig. 1 Task 2 and Proof 2

and line-by-line logic (L-line), and more important, found that the word “clarity” carried a variety of meanings and was confounded with logic, justification, readability, fluency, and understanding.

The next stage was an examination of the transcripts of the professors’ responses to the interview questions that followed the grading of the proofs, from which I developed a table with quotations that were pertinent to the various codes. The written marks on the proofs, the oral commentary during proof grading, and the responses to the interview questions provided a means of triangulating the data and reducing the list of codes to a short list of the most salient ones. The [follow-up study](#), described in the next section, provided further corroboration of the results of the analysis.

Follow-Up Study

As reported below in the Results section, the analysis of the professors’ grading of the six proofs revealed substantial variation in the scores for four of the proofs. In order better to understand why the scores varied, I conducted a follow-up study with the same four professors in the fall of 2013, about a year after the initial study.

Two primary issues motivated this follow-up study. First, I wanted to determine the extent to which the spread in the scores could be attributed to performance errors (Inglis et al. 2013), i.e., to the fact that the professors overlooked flaws or features in the proofs that had been noted by other professors, as opposed to differences in grading practices or views on what constitutes good proof writing.

Second, the data analysis revealed that logical correctness was the most important consideration in the scoring of a student's argument, but none of the six proofs contained a major logical error, except for the error at the beginning of Proof 6. So I wanted to present one more proof to the professors to see how they would respond to a logical error.

For the follow-up interviews I prepared “corrected proofs” for Proofs 2, 4, and 5. These corrected proofs summarized all of the marks the professors had written on the proofs in the earlier study. I began each interview by showing the professor her original scores for these three proofs but not the other professors' scores. Next, I presented the three corrected proofs to her along with her original evaluations of those proofs and asked, “Now that you have seen your comments and marks along with those of the other professors, would you change your scores?” Finally, I asked her to evaluate Proof 7 (see [Appendix](#)), which I wrote for the purposes of this study. The argument begins with the conclusion of the theorem and arrives at a true statement, with no comment about the need to reverse the steps. Also, the argument does not indicate where the assumption that x is positive is needed.

The follow-up interviews lasted about an hour and were audio-recorded with a Livescribe smart pen as in the earlier interviews.

Results

This section begins with an analysis of the professors' evaluation of Proof 2 for two reasons: this analysis is representative of the analysis of the other proofs, and it provides an introduction to the following five observations that emerged from the study:

1. The professors view proof grading as an important part of teaching.
2. When scoring the proofs, the professors made judgments about the seriousness of errors.
3. In deducting points for errors, the professors' judgment about the student's cognition mattered more than the errors themselves.
4. The scores the professors assigned to the proofs varied substantially for some of the proofs.
5. The professors agreed that attending to logical correctness, clarity, fluency, and demonstration of understanding were four important characteristics of well-written proofs, but they noted that these characteristics are sometimes confounded so that they could not always be certain of the reason for a flaw in a proof.

Analysis of Proof 2

In evaluating Proof 2 (see [Table 2](#) and [Fig. 1](#)), all four professors focused primarily on mathematical errors, logical correctness, clarity, and readability. They checked whether the proof began and ended correctly and whether the proof flowed logically and clearly from beginning to end. Three professors said the first sentence should be omitted or prefaced with a phrase like “We want to prove.” The professors marked items for correction: (a) two noticed yRz should be xRz in line 1, which they judged to be a

“typo,” an accidental writing of the wrong letter as opposed to a mathematical misunderstanding; (b) three noticed that $\mathbf{Z}$ (the set of integers) should be $\mathbf{R}$ (the set of real numbers) at the end of line 1; (c) two said the word “let” should be “for some” in line 3; and (d) two changed “an $\mathbf{Z}$ ” to “in $\mathbf{Z}$ ” in the next to last line.

With regard to clarity and readability, Professors A and C said the proof could be clarified by replacing the work on lines 4–7 with “ $x - z = (y + k) - (y - c) = k + c$,” that is, by starting with $x - z$ and showing it equals an integer. Professor C said that the work on lines 4–7 is “okay for scratch work” but is not acceptable for a proof because proofs must “flow logically” from what we know to the desired conclusion and be written in complete sentences. Professors B and D said the proof would be easier to read and the student’s thought process more transparent if reasons were given, as in Proof 1.

The professors paid little attention to punctuation and capitalization and made no comments about the fact that $k + c$ was circled and that the proof ran diagonally down the right-hand side of the page. Upon my questioning, Professor B said the layout could be more organized, but she and Professor D said the layout mattered little to them and normally did not affect the score.

Proof Grading as a Teaching Practice As the professors evaluated and scored the proofs, made corrections, and wrote comments, they clearly viewed themselves as *teachers* and in this role they evaluated the student’s understanding and proof-writing skills, not just the proof as a decontextualized mathematical artifact. They spoke repeatedly about “the student” who wrote the proof and what he or she may have been thinking. The various marks and comments they noted on the proof served to correct the student’s errors, justify the scores, and help the student learn to write a better proof, all of which are acts of teaching.

Judgments about the Seriousness of Errors and the Student’s Cognition The error at the end of line 1 could render the proof invalid. Given that R is a relation on the set of real numbers, the transitive property must be established for all real numbers x , y , and z , whereas the student’s argument assumes x , y , and z are integers. The professors judged the seriousness of this error differently, which was reflected in the scores. Professor A initially judged the proof to be “perfect” and assigned it a score of 10, but later noticed the error and lowered the score to 9.8 and said the proof was “nearly perfect.” Professor C also judged the error as minor, saying it was “probably an oversight-type mistake.”

Professor B, on the other hand, judged the error to be far more serious than Professors A and C and seemed to think it revealed a fundamental misunderstanding on the part of the student.

B: Actually, it’s a very critical one. I didn’t catch it.

I: Real numbers instead of integers.

B: Yeah. Then I’ll change this score – 8 is too high for that [proof] because that is very critical. Like 5.... It’s not simply a mistake, I think.

Professor D did not notice this error and assigned the proof a score of 9, but in the follow-up study interview I pointed out the error and asked whether she would like to

change her score. She lowered the score by two points, one point of which was due to this error, but wrestled with the decision:

D: It could be a small or large mistake. It would depend on if they [the student] generally know where things are or if they consistently don't know where their objects are living. If that was carelessness or confusion, I can't distinguish it out of context, in the context of the class. [She went on to explain that "context" meant her observation of the student's knowledge and performance in similar situations in a course and that she might ask the student what he or she was thinking with respect to this particular error.]

In summary, we see that two professors judged the error to be minor; one judged it as major; and the fourth was unsure whether the error was major or minor because she did not know what the student was actually thinking and, therefore, would tend to give the student the benefit of the doubt, i.e., judge the error to be an oversight, if she knew the student was generally a strong student who normally did not make such errors. Notice that the professors were not simply checking the validity of the argument; rather, they were grading the *student*, and the scores reflected their judgments of the seriousness of the error, based on their judgments of the student's level of understanding.

Variation in the Scores As shown in Table 3, the scores for Proof 2 were 9.8, 5, 8, and 9 out of 10 points, with a mean of 7.95 and a range of 4.8. Professor A said the proof was "nearly perfect," although she noted various ways the proof could be improved. After her initial reading of the proof, Professor B assigned the proof a score of 8, but a few minutes later lowered it to 5 after noting the error at the end of line 1 (*Z* rather than *R*). Professor C said the proof contained "no major mistakes" but "the presentation isn't terribly good." Professor D assigned a score of 9 but said it could be a 7, depending on how much emphasis she had placed on justifying the steps.

In sum, the professors agreed that the overall logical framework of Proof 2 was essentially correct, the student understood the proof and its key ideas reasonably well, the clarity could be improved, and surface features such as layout and punctuation should carry little, if any, weight in the overall score. But they disagreed on the seriousness of the student's error at the end of the first line, the need for additional

Table 3 Scores the professors assigned to the proofs

Professor	Proof 1	Proof 2	Proof 3	Proof 4	Proof 5	Proof 6	Proof 7
A	10	9.8	9.5	10	9.5	9.5	7
B	10	5	6	7	7	5	6
C	8	8	9	8	8	— ^a	7
D	9	9	8	7	7	8	6
Mean	9.25	7.95	8.13	8.00	7.88	7.50	6.50
Range	2.0	4.8	3.5	3.0	2.5	4.0	1.0

^a Due to time constraints, Professor C did not evaluate Proof 6

justifications, and the necessity of writing the proof in complete sentences. Also, the scores they assigned to the proof varied from a low of 5 to a high of 9.8.

We now examine the five observations in more detail.

Proof Grading as an Important Teaching Practice

As noted in the analysis of Proof 2, the professors consistently viewed themselves in the role of a teacher as they evaluated the proofs. The assumption of this role was due in part to the way I began the interviews by asking them to mark and score each proof as if a student had submitted it for homework. But clearly they fell naturally into this role, and it is for this reason that I generally refer to them as professors rather than mathematicians in this paper.

Evidence for the claim that the professors evaluated the proofs from the perspective of a teacher is seen in the way they talked as if they were addressing a student and judged the seriousness of errors by considering the student's cognition. This perspective was expressed cogently by Professor C in reference to Proof 7: "So how am I going to grade such a thing? [It's] perfectly good scratch work. ... I don't understand, how would I divine whether or not the student understands the structure of an if-then statement?"

Further evidence is seen in the types of marks and comments the professors wrote on the papers, which included correcting notation and other mathematical errors, suggesting revisions to improve readability and clarity, asking for justifications, striking out incorrect or unnecessary sentences or paragraphs, correcting punctuation and capitalization, specifying that proofs should be written in complete sentences, and rewriting parts or all of the proof. These marks and comments were not necessary for simply judging the validity of the proofs and scoring them; rather, they were instructive remarks whose aim was to teach the student how to write clear, correct proofs.

I asked the professors how they communicate the characteristics of good proof writing to their students. Two means were prominent in their responses: modeling good proof writing in class and writing comments on students' papers.

A: I model it on the board, actually, when I prove a lot of theorems and properties during class time. After that, when I grade their papers, I make a lot of comments as I did on the papers here today.

In the following quotation Professor D reveals a variety of strategies for teaching the proof-writing process: modeling the process in class, insisting on justifications for the steps in the proofs, writing many comments on students' papers, talking with students individually, verbalizing the underlying logical reasoning, and involving students in the proving process in class.

I: Now, how do you communicate these characteristics of a good proof to students?

D: I rely primarily on examples. I try to illustrate what [a good proof] is, and at the beginning [of a course] insisting that they do give justifications so that I can see they know even the most basic reasoning.

I: So later in the a course as you're doing a proof, do you point out certain aspects of a proof that make it a good proof, that shows it communicates clearly? Or do you just simply do the proofs?

D: I probably don't point it out as much as I should later. ... I write a lot on their papers as far as how to improve on it, and dialogue if they come in to office hours.

I: That would be throughout the course?

D: Right. But as far as within the class, I try to verbalize my logic. You know, I'm thinking, "Why am I making these steps in the proof?" But I don't necessarily point out that it's because this makes it a good proof. I'm not necessarily defining good proof as I am defining logic. ... I try to have them give ideas of how to do the proofs in class, as well, so more interactive proof

Professor C explained that, in addition to modeling good proof writing in class, she explicitly teaches the process of writing a proof by pointing out the logical structure, unpacking definitions and previously-proved theorems, doing the scratch work, and finally writing the proof in proper form.

C: I make a big deal out of the logical structure. Look at this theorem. It says if this is true, then this is true. Because almost none of them have had a transitions course where they learn the structure of proof, I make a big deal out of that. I say, "Well, to prove a statement of this form, we'll assume the first part, and then we'll try to show that the last part follows." And usually that involves unpacking definitions and using previous theorems, and so forth. Each sentence has to follow logically from the previous one, and then we'll end up with the result. I'll usually do a scratch work version first and say, "If I were going to do this, I would do it on scratch paper first." And then I'll actually write on the board scratch work, you know. Some of these proofs we graded here, I would have done that vertically here. I'd say [for Proof 3], "Well, if $2n_1 + 1$ is equal to $2n_2 + 1$," then I'd work down vertically and say, "so that's the guts of the proof. That's what's going to make it work. Now let's write it out in proper form."

Professor C's teaching process is consistent with what Fukawa-Connelly (2010) termed modeling mathematical behaviors. While the other professors reported that they were less explicit and detailed in teaching the proof-writing process, all four devoted considerable time to marking students' papers and writing comments.

The next section lends further support to the notion of proof grading as an act of teaching by demonstrating that the professors consistently evaluated the seriousness of errors in terms of their perception of the students' thinking and the degree to which the students misunderstood aspects of their written arguments. In other words, the professors showed they perceived proof grading as an assessment of students' learning, not simply an evaluation of the correctness of the purported proofs.

Judgments about the Seriousness of Errors and the Student's Cognition

In the analysis of Proof 2 we saw that all four professors agreed that there was an error at the end of the first line of the argument, namely that x , y , and z should be taken as real numbers rather than integers. Two professors decided this error was minor, whereas one decided it was major, and the fourth said she could not know the seriousness of the error without knowing more about the student's general performance in the class or talking with the student about the error. We make two observations here: first, according to the professors, not all errors carry the same weight, i.e., some are more grievous than others, and second, the professors determined the seriousness of an error by the student's cognition rather than by the error itself.

The following evidence supports the claim that the professors judged the seriousness of the error in terms of the student's cognition. Professor D explicitly spoke of judging the seriousness of the error in terms of the student's cognition. Professors A and C also appeared to evaluate the seriousness of the error in terms of the student's cognition. As Professor A talked aloud about Proof 2, she repeatedly referred to the "student" who wrote the proof, and Professor C said "he or she starts out okay except for that problem of putting Z instead of R , and I suspect that is probably an oversight-type mistake." An oversight mistake is something a person does, not a property of a mathematical error apart from the person who wrote it. When Professor B said "it's a very critical one [error]" and "it's not simply a mistake, I think, in my opinion," her meaning was unclear as to whether she was thinking of the error itself or the student's misunderstanding. I inferred, however, that she was thinking about the student for the following reasons: (a) Throughout her oral commentary about Proof 2, she repeatedly referred to "the student" and "the person" who wrote the proof, and (b) when she said "it's not simply a mistake," she was stating that the student made the mistake and the mistake was due to a more serious misunderstanding than a minor slip.

The evaluation of Proof 7 also shows that the professors judged the seriousness of an error in terms of their perception of the student's cognition, arriving at differing conclusions. When I asked the professors to evaluate and score Proof 7, I also asked them to assign the proof a score from 0 to 5 for validity, where 0 meant "very poor" and 5 meant "very good." Although the overall scores were quite consistent (see Table 3), the surprise was in the validity scores: Professors A, B, C, and D assigned validity scores of 3, 1, 2, and 4 out of 5, respectively. Although they all agreed that the proof incorrectly begins with the conclusion of the theorem, they arrived at different conclusions about the seriousness of the error. I probed them for their reasons. After Professor A said that the proof was written in the wrong direction and that it should end with the conclusion of the theorem, I asked, "But you would still say the student understands?" and she said, "Yes, yes." Professor D, who chose a validity score of 4, also gave the student the benefit of the doubt.

D: Validity, um (long pause), that seems to be the main problem. For the validity, they seem to be communicating it one direction rather than equivalency, even if they're thinking equivalence....

I: ... I'm not sure how you interpret what the student is thinking.

D: ... I'm thinking that they've done equivalent steps from this statement, and so this is more like scratch work, and then they need to start from here and go back that direction. I would expect them to write it in that direction.

These excerpts show that Professors A and D considered the student's understanding in their assessment of the validity of the proof and gave the student the benefit of the doubt in understanding the logic of the proof. In contrast, Professor C articulated her struggle to evaluate the seriousness of the error in terms of the student's cognition.

C: They're trying to prove this [inequality], but they're starting with what they're trying to prove, so we have a problem. (pause) ... Okay, so it looks like the main problem is that the implications are all reversed. So how am I going to grade such a thing? [It's] perfectly good scratch work.

... I don't understand, how would I divine whether or not the student understands the structure of an if-then statement? 'Cause sometimes students are just doing scratch work, like on some of these previous proofs we looked at, and it gets really snarled up, but you know the basic logic of those proofs was correct, and you know the student basically understood what he or she was doing. But here, does the student really understand? This could be just scratch work, and then they write it sort of as though it's a proof, or does the student really not understand that you can't start with what you're trying to prove? That's my question. I don't know, it's not clear to me.

In these excerpts we see that the professors focused on assessing the student's understanding as much as evaluating the proof itself, and Professor C explicitly articulated the tension between these two viewpoints, particularly the difficulty in knowing just what the student had in mind.

The professors considered correct logic to be the most important characteristic of a good proof (see the [Characteristics of Good Proof Writing](#) section below) and said it carries the most weight in their scoring of a proof. But the analysis of their grading of Proofs 2 and 7 shows that their assessment of a proof's logical correctness depends on their assessment of the student's understanding. Thus, judgments about the seriousness of errors and the student's cognition associated with the errors play a key role in these professors' evaluation of proofs.

Variation in the Scores Assigned to the Proofs

As reported in Table 3, the scores for four of the first six proofs varied by 3 or more points, prompting me to ask why the scores varied to this extent. One possible explanation is performance errors (Inglis et al. 2013). As noted earlier, a performance error is an error due to overlooking a flaw in a proof. We saw in the analysis of Proof 2 that only two of the four professors noticed the error yRz rather than xRz in the first line, and only three noticed the error Z rather than R at the end of the first line. Furthermore, Professor B initially overlooked the latter error but later noticed it and lowered her score by 3 points. Thus, performance errors may have played a role in the score variation.

A main purpose of the follow-up study was to examine more carefully the reasons for the variation in the scores, particularly the possibility of performance errors. The analysis of the data from the two studies revealed several factors that contributed to the spread in the scores.

Performance Errors

In order to keep the interviews from running too long, I chose to ask the professors to regrade only three of the first six proofs. The data analysis of Proofs 1, 3, and 6 revealed that the professors were largely in agreement and had overlooked few details noted by the others, except that only one professor noted in the second part of Proof 3 that the student needed to show that $(n-1)/2$ is an integer. Thus, I chose Proofs 2, 4, and 5 for regrading because performance errors were more likely to have occurred in the scoring of these proofs. As explained in the methodology section, during this regrading of the proofs the professors were able to see the corrected proofs with all four professors' comments and marks from the original interviews, but not the other professors' scores. With four professors and three proofs, there were 12 scores that could have been changed.

The results of the regrading of the three proofs are shown in Table 4. The professors chose to change only four scores, all of which were changed to lower scores, and only two of which were due to performance errors: (a) Professor A lowered her score on Proof 4 by 0.2 point because in her initial scoring she “didn't take any points off about not having the induction hypothesis,” and (b) Professor D lowered her score on Proof 2 by 1.5 points because she had overlooked an error at the end of line 1 and saw that the layout could be improved. (The other score changes were due to the professors' changing their mind upon reevaluating the proofs.) Although the professors found other deficiencies in the proofs they had not noticed in their initial evaluations of the proofs, they considered these deficiencies to be minor and insufficient to warrant a score change.

Thus, the follow-up study uncovered only two performance errors: one score dropped by 0.2 point and another by 1.5 points. For these three proofs, we may

Table 4 Initial scores and scores after regrading in the follow-up study

Professor	Proof 2		Proof 4		Proof 5	
	Initial	After regrading	Initial	After regrading	Initial	After regrading
A	9.8	9.8	10 ^a	9.8	9.5	9.5
B	5	5	7	5	7	5
C	8	8	8	8	8	8
D	9 ^b	7	7	7	7	7
Mean	7.95	7.45	8.00	7.45	7.88	7.38
Range	4.8	4.8	3.0	4.8	2.5	4.5

^a Score was lowered by 0.2 point due to a performance error

^b Score was lowered by 1.5 points due to a performance error

conclude that performance errors made only a small contribution to the variation in the scores, and therefore the variation was due largely to other factors.

Interestingly, after this reevaluation of three proofs, the range for Proofs 4 and 5 increased from 3.0 to 4.8 and from 2.5 to 4.5, respectively, not because of performance errors but because Professor B assigned a lower score the second time around, leaving five of the seven scores with a point spread of at least 3 points.

Disposition Toward Grading

It is not unexpected that some professors are more lenient in grading while others are stricter. In this study Professor A gave the highest score on all seven proofs, and Professor B gave the lowest score on all but Proof 1. If we omit Professor B's scores and use the scores after the regrading of Proofs 2, 4, and 5, then the score ranges for the seven proofs are 2.0, 2.8, 1.5, 2.8, 2.5, 1.5, and 1.0, respectively, all of which are less than 3. Thus, the professors' disposition toward grading played a significant role in the variation of the scores.

Professor A's high scores can be attributed to her focus on logical correctness and her generous perception of the "student's intention," i.e., general understanding, and that she gave little or no weight to details of layout, clarity, and fluency in the scoring, although she did attend to these factors in her marks and comments. Furthermore, she was reticent to give low grades for fear of discouraging students. Professor C insisted on complete sentences and a rather formal proof-writing style, and she gave these factors some weight in the scoring. Her weighting of these factors was small, however, because she recognized that the students who wrote the proofs were in an introduction to proofs course. In contrast, Professor D was more flexible than Professor C in that she allowed, and even encouraged, students to develop their own proof-writing voice and did not insist on any particular writing style. Although Professor B was more severe than the others in deducting points for errors and said she pays attention to details such as notation and layout in courses that emphasize proof writing, she does not give such details any weight in lower-level courses like calculus.

Professor C said, "I take a more gestalt point of view when I'm grading," indicating that she does not deduct points for each error but instead arrives at a score by judging the overall quality of the proof. While the other professors did not express their proof scoring process in these terms, clearly that they all used this approach to a degree. Thus, some of the variation in the scores was due to the professors' judgments about the overall quality of a proof, which depended on their views about the characteristics of good proof writing and their expectations of students' work.

Judgments about the Seriousness of Errors and the Student's Cognition

The necessity to judge the seriousness of errors and the student's cognition behind the errors appeared often in the professors' proof grading. As discussed earlier, the professors wrestled with the seriousness of the error at the end of the first line in Proof 2. Professor B deducted 3 points for this error, whereas the other professors were more generous. They questioned whether the student made a "typo" or demonstrated a serious misunderstanding and tended to give the student the benefit of the doubt. Professor D said she could not know why the student made the mistake without having

the context of knowing the student's general performance in class, seeing other samples of the student's work, or talking with the student.

In the inductive step in Proof 4, the student did not explicitly state the induction hypothesis, i.e., assume $P(n)$ for some n , and Professor D struggled with the seriousness of this omission and assigned a score of 7 with hesitation.

D: But what do they understand? They didn't write that out. How deep is the error? (pause) Well, I would be concerned that they're just saying we need to show $P(n+1)$ without knowing that it's an if-then, because the if-then's are the guts of proofs and the logical deductions. So that's a significant concern to me. I'm not sure whether it's conceptual or communication skill, but they need to state that in some form. Um, for a score, probably 7 out of 10. And I would probably talk with them more about clarity. It would be from other samples of their work I would know whether they really do understand that if-then or not. If for other problems they're not writing it out, then I'd say that would be a concern. Um, for a score, probably 7 out of 10. And I would probably talk with them more about clarity. It would be from other samples of their work I would know whether they really do understand that if-then or not. If for other problems they're not writing it out, then I'd say that would be a concern.

In addition to having to decide the seriousness of particular errors, all four professors spoke repeatedly of having to judge how well the student understood the proof as a whole. If Professor A thought the student's approach to a proof was essentially correct, she was generous in her judgment of the student's level of understanding and awarded points accordingly, as she did when she awarded a score of 9.5 for Proof 5. As we saw earlier in "[Judgments about the seriousness of errors and the student's cognition](#)" section, Professor C struggled to determine why the student wrote Proof 7 by beginning with the conclusion. Further evidence of the difficulty in judging the overall quality of a proof is the fact that Professor B lowered two scores when she regraded the three proofs in the follow-up study simply because she decided her original scores were too high and not because she found additional deficiencies in the proofs.

Contextual Factors

These four professors had not taught the courses from which the proofs in this study were taken, did not know the proof-writing expectations I had established in those courses, and were not fully aware of the background knowledge that had prepared the students for these proofs. Their assumptions about contextual factors may have affected the scoring. For example, Proof 5 came from a test in an introduction to proof course for which I was the instructor, and one purpose for its inclusion on the test was to assess whether the students had learned correctly to use quantifiers and the set-builder definitions of subset and cross product. I would have deducted points for the many notational errors, whereas the four professors focused on the student's overall reasoning and understanding in their scoring of the proof, while marking notational errors but giving them little weight in the score.

Other contextual factors included the position of the course within the mathematics curriculum (i.e., an introduction to proof course versus an advanced course), the

background knowledge available to the student who wrote the proof, the characteristics of good proof writing the professors value and their expectations of students in a particular course, and for at least one of the professors, Professor D, the professor's familiarity with the student's competence and general pattern of performance.

These observations about context are consistent with other researchers' observations about the importance of context in mathematicians' judgments about the validity of an argument. For example, Weber (2008) found that mathematicians sometimes withhold judgment on the validity of a proof because they lack "sufficient information on the context in which the proof was written" (p. 447). In his study, a step in an argument was unjustified, and the mathematicians struggled to decide whether a student had sufficient knowledge to fill the gap—a decision they found to be difficult and ambiguous, resulting in different conclusions. So context is related to the issue of disagreements about the seriousness of errors because, for instance, a missing justification in a proof may be a simple omission or a serious error, depending on the background knowledge of the writer. Furthermore, Weber found that mathematicians were more likely to accept a proof as valid on the basis of the reputation of the author, as Professor D was more willing to accept a proof written by a generally competent student, even though the proof had a flaw.

Characteristics of Good Proof Writing

The characteristics of a well-written proof that emerged strongly from the data are logical correctness, clarity, fluency, and demonstration of understanding, with logic and clarity being the most important to the professors, as articulated by Professors C and D:

C: A well-written proof? The most important thing is that it's logically correct. If a proof isn't logically correct, I'll often take almost all the points off I give them a pretty low grade if the logic is incorrect. So that's the main thing. I would say the second thing is the readability of the proof. Is it flowing in complete sentences, or is it like vertical scratch work sort of thing?

D: Yes, the logic and clarity are the two principles. It seems like everything falls into those categories.

Logic, Clarity, Fluency, and Understanding

The professors emphasized logical correctness as the most important characteristic of well-written proofs. As discussed in the analysis of Proof 2, the professors checked that the proof began and ended correctly, that the intermediate steps were correct and flowed logically from beginning to end, and that the algebraic manipulations were correct. Thus, we see here that their comments about logic referred to the overall logical structure, or proof framework (Selden and Selden 1995), the line-by-line reasoning, and the correctness of algebraic manipulations and calculations.

Clarity emerged as the second most important characteristic of well-written proofs. The prominence of clarity and readability in the data is highlighted in the preceding quotes by Professors C and D.

For these professors, clarity seemed to encompass a variety of meanings: (a) explicitly stating the reasoning or justifications for the steps in a proof, (b) organizing the proof to make it readable and flow smoothly, and (c) using mathematical language and notation correctly. These meanings of clarity are evident in the analysis of Proof 2. For example, Professor D said that Proof 1 was clearer than Proof 2 because Proof 1 listed justifications for the steps, and two professors said the algebraic manipulations could be clarified by starting with $x - z$ and proceeding toward the integer $k + c$.

Fluency refers to the correct use of mathematical language and notation, as well as English grammar, punctuation, and capitalization, so as to make the proof flow smoothly and be easy to read and understand. Although the professors often marked aspects of fluency on the proofs, they seldom deducted points for such matters. For example, here is an excerpt from Professor C on Proof 5 in which she explicitly spoke about various fluency issues related to the English language and mathematical symbolism:

C: Well, let's see. How would I clean this up as far as how it's written? ... I would want to say "complete sentences." I'm really bad about that. Complete with punctuation. And it can be logical symbols, but those are verbs and all those symbols have parts of speech, and so we can write them with periods and commas and so forth.... There's an arrow coming out of the definition part of the set. So I would say something like "improper syntax" or something. I'm not quite sure. Yeah, "improper syntax, it's OK for scratch work." [writes comments on the proof] This has logical problems. Well, it has syntactical problems, I guess.

I: The next to the last line.

C: Yeah, the next to the last line has, and I would put up here that this has "bad syntactical problems." And I would make a note that "Sets can't imply sets." And I would say "improper use of 'for all' quantifier." But it's not too bad. I would give, probably, an 8 out of 10.

The other professors were less particular about fluency than Professor C, but all of them marked at least some fluency issues on the proofs as a way of instructing students on good proof writing, and they agreed that fluency should carry little, if any, weight in the scoring.

The fourth characteristic of good proof writing is related to understanding. The professors said that a good proof demonstrates that the student understands the proof and the associated concepts. Earlier sections of this paper discussed the professors' focus on students' cognition as they made judgments about the seriousness of errors and the quality of the proof as a whole. Professor D said "my emphasis on the proofs is on looking for the clear logical thinking and that they are *understanding* how they are getting from one place to the next," Professor C said "I think *this person knew what he or she was doing*, logically speaking," and Professor B said "*that person seems to understand* what should be written" (emphasis supplied). Here is an additional excerpt that shows Professor A's focus on students' understanding in her scoring of proofs:

A: Clarification is very important for me, and besides that, first of all I try to catch the student's intention, even if it's not clearly stated and organized. First of all, I

try to catch the student's intention. And then after that, if I'm sure that the students know what they should prove and if their intention is correct, then I try to give full points when I grade. I try to be very generous at that point. But if it [the proof] can have more clarification, that's much better, I believe.

Interactions among the Four Characteristics

Although the professors identified four important characteristics of well-written proofs, actually using them to grade proofs is a difficult, complex activity because the characteristics are inextricably intertwined. In this excerpt Professor D explained that in some cases she cannot distinguish a logical error from a deficiency in clarity.

I: Thinking about the logic and the clarity of communication, does one carry more weight than another when you assign a score?

D: Yes, the logic carries more weight. But I am somewhat dependent on their clarity to know what their logic was. I think there are particular types of clarity where they're going to lose [points], because I think it's logic, and it may or may not be.... They interact.

In fact, this confounding of logic and clarity occurred when she evaluated Proof 4, as we saw in the [Judgments about the Seriousness of Errors and the Student's Cognition](#) section above. In the inductive step, the proof does not explicitly say "assume $P(n)$," and Professor D struggled to decide whether the student mistakenly thought it was necessary to prove $P(n+1)$ rather than the conditional statement "If $P(n)$, then $P(n+1)$ " or whether the student was inept at clearly communicating the details of the proof.

Consider the following statement by Professor B about Proof 5 that illustrates the confounding of the four characteristics:

I: On the set theory proof there was difficulty with the notation. You seem to think the logic was basically right?

B: Yes. That person seems to understand what should be written, but the way the proof is written is not quite readable.

All four characteristics appear in this quote: "That person seems to understand" refers to the student's cognition; "what should be written" refers to correctness, including both logical and symbolic correctness (fluency); "the way the proof is written" refers to a lack of fluency; and "is not quite readable" refers to a lack of clarity. Professor C concurred that the student who wrote Proof 5 seemed to understand what should be written but lacked fluency and clarity.

C: I'm looking at the next to last line of the proof.... First of all, we have one set implying another set, which doesn't make any sense. And the other bad thing is that out front of that, we have a quantifier, for all x and y , and yet x and y are just placeholders in the set. So none of that makes any sense, and yet the person

actually understands what he or she is doing, in some sense, logically speaking. But it's a really terrible way to write it.

These quotes illustrate the complexity of proof grading and the impossibility of segregating the four characteristics they value in good proof writing. This complexity is due in part to the professors' view that *students* wrote these proofs, and therefore the proofs must be evaluated in view of the students' cognition. But to demonstrate understanding the student must write a logically correct proof that flows smoothly from beginning to end, which in turn requires not only logical thinking skills, but also knowledge of mathematical concepts, skills in written communication, and fluency in both natural and mathematical languages. Consequently, the professor faces difficult judgment calls in attempting to ascertain *why* the student omitted an assumption, used mathematical symbols incorrectly, or made other mistakes.

Thus, with respect to the first research question about whether mathematics professors agree in their evaluation and scoring of proofs, the results show commonalities among the four professors in their evaluation of the proofs. They focused on logical correctness, clarity in reasoning, and the student's understanding, while also correcting mathematical errors and writing marks and comments that related to readability, mathematical notation, and punctuation. They differed somewhat in the weight to which they gave these characteristics in the scoring process and sometimes arrived at different conclusions in their assessment of errors and the student's level of understanding.

Discussion and Future Directions

This study contributes to the literature on the teaching and learning of proof in five ways. First, it highlights an area of undergraduate mathematics education that has received little research attention, specifically, mathematics professors' proof grading practices. The participants wrote many detailed marks and comments on the proofs and said this is one of the main tools they use to help students learn to write good proofs. Thus, for these professors, proof grading is an important teaching practice because it plays a role not only in helping students develop their proof-writing skills but also in communicating to students the characteristics of good proof writing that mathematicians value and, to some extent, their views on the nature of proof. As we study the teaching of undergraduate mathematics, we should remember that teaching practices do not occur solely in the classroom but also in the office when no student is present.

Second, this study describes the complexities of proof grading. Earlier studies on proof validation (Weber 2008; Inglis and Alcock 2012; Inglis et al. 2013) found that mathematicians do not always agree in their evaluation of the validity of a purported proof. Weber et al. (2014) expressed the problematic nature of validity judgments:

Whether a deductive argument is a valid proof is (sometimes) not an objective feature of the argument but depends on the mathematician reading it, among other factors. Although some arguments are clearly invalid (e.g., arguments based on blatantly false assertions or containing computational errors), other arguments' validity is more problematic to determine because one cannot be sure how large

of a gap in a proof is permissible. This has an important implication for the teaching of mathematics majors; mathematicians often do not agree on what level of detail should be included in proofs they present to students (Lai et al. 2012) and, we hypothesize, on the level of detail that students should provide on proofs that they submit for credit. (p. 53)

The data from the present study support these authors' position that the difficulty of determining the nature of a gap in a deductive argument may be due to factors other than the proof itself. It also supports the authors' hypothesis that mathematicians may disagree on the level of detail expected in students' proofs. The study contributes to this line of research on proof validation by showing that the evaluation of undergraduate students' arguments involves much more than making judgments about validity and by illuminating some of the "other factors" involved in this process. It explains differences in mathematicians' judgments about validity and errors in proofs by introducing the notion of the seriousness of a flaw in a proof and documents mathematics professors' practice of judging the seriousness of flaws by assessing the student's underlying cognition. In other words, mathematics professors focus on evaluating the student as much as the proof itself. Although they may concur on the importance of logical correctness, clarity, and fluency in good proof writing, they sometimes differ in their evaluation of a student's proof because they differ in their perceptions of what the student was thinking, and consequently they arrive at different judgments on the seriousness of errors, the extent to which the student demonstrated a good understanding of the proof and the associated mathematical concepts, and potentially, the validity of the proof.

Third, the four professors' emphasis on logical correctness, clarity, and fluency as characteristics of well-written proofs is largely consistent with the readability, validity, and fluency categories of the RVF method (Brown and Michel 2010). One difference between the results of the present study and the RVF method is that these professors give little weight, if any, to fluency in the scoring of proofs. Brown and Michel do not specify weights for the three categories, but in an example they suggested that fluency might get up to 2 points out of a total of 8 points. Another difference is that the RVF rubric does not explicitly recognize the role of the professors' judgment of the student's level of understanding in the evaluation and scoring of proofs. Brown and Michel do advise, however, that to achieve consistency and efficiency in the evaluation and scoring of students' proofs the evaluators must be trained in the use of the RVF method and that each of the three categories should be scored "on a scale close to Boolean" (p. 5), specifically, no more than 3 points per category. This advice appears to be consistent with the results of the present study in that the professors struggled when assigning points on a 10-point scale. The difficult judgment calls on the seriousness of errors and the corresponding decisions on the deduction of points would likely be eased when the categories are clarified through training and practice and the point values are small.

Fourth, with respect to the issue of context, this study adds evidence to that of others (Inglis and Alcock 2012; Inglis et al. 2013; Weber 2008; Weber et al. 2014) that context matters when mathematicians write, read, validate, and evaluate proofs. These authors consider the context of the mathematical content (e.g., an introductory course or an advanced one), the author's reputation (e.g., a well-known mathematician or a student), the author's background knowledge or level of expertise (e.g., expert or novice), and where the proof appears (e.g., in a journal or on a students' test or homework paper). These observations are

consistent with Thurston's (1994) assertion that mathematical knowledge, theorems and proofs in particular, exist in a community: "We prove things in a social context and address them to a certain audience" (p. 175). Mathematicians also evaluate proofs with respect to their own expectations, standards, and experience. For example, Inglis et al. (2013) found that pure and applied mathematicians seem to hold different standards of validity. Thus, proofs do not stand alone in isolation; rather, they are artifacts of human activity and are evaluated with respect to the contexts in which they are written and read.

Finally, the study contributes to the literature on the beliefs and practices of mathematicians. Studies (e.g., Inglis et al. 2013; Lai et al. 2012; Weber 2008, 2014) have shown that mathematicians' views and practices are far more varied and nuanced than traditionally thought. Weber et al. (2014) synthesized the research in this area and argued for the value of this line of research for the design of mathematics instruction. The present study adds insight into some of the commonalities and differences in mathematicians' views on the characteristics of well-written mathematics, their proof evaluation practices, and their use of proof grading as an instructional practice. In addition, it reveals that although mathematics professors may generally agree on the characteristics of good proof writing and using these characteristics to provide feedback to students by way of corrections and comments, they may disagree on how to implement these characteristics in scoring the proofs. That is, proof evaluation and proof scoring are two related but distinct aspects of proof grading.

This study lays the groundwork for further investigation. One key result is that in some cases the mathematicians gave very different marks to the same proof. It is natural to wonder how much variance exists in the population of mathematicians as a whole. A much larger sample would be needed to begin to estimate the extent of this variance, so one plausible approach would be to send a survey to mathematicians asking them to evaluate proofs like the ones in this study. Variance in mathematicians' evaluations of proofs has been reported with validating proofs (Inglis et al. 2013) and judging their aesthetic qualities (Inglis and Aberdein 2016). It would be an interesting and important finding if this is often the case with proof grading.

A second key result from this study is that the participants assigned grades, in part, on their view of the student's understanding of the mathematics. Conducting a survey to see the extent that this viewpoint is shared by other mathematics professors could provide an important insight into the factors that mathematicians consider when they grade proofs. Mejía-Ramos and Weber (2014) reported variance in mathematicians' refereeing practices, including the factors that they consider when reviewing proofs, so a more thorough assessment of mathematicians' proof-grading practices, particularly how they judge a student's level of understanding, would add to that earlier work.

This study has only begun to address the ways by which mathematicians evaluate students' mathematical work and assign grades. The study used a small number of proofs in an unnatural setting. Watching mathematicians grade proofs in their actual practice over a period of time or across a variety of courses would provide greater insight into the wide range of contextual factors they consider when they evaluate proofs.

Finally, given the importance the professors in the present study place on writing marks and comments on students' papers, it would be valuable to examine whether proof grading is an effective teaching practice. Do students correctly interpret the comments, and to what extent do they learn from their professors' feedback? Researchers have found that in advanced mathematics, students frequently interpret

comments in different ways from what the professor intends and may not prioritize the things that a professor thinks are important (Lew et al. 2016; Weinberg et al. 2014). Knowing if this is the case with students' reactions to the written feedback they receive on their proofs and other mathematical work would be of practical use to mathematics professors.

The study has several limitations. First is the small number of participants and the fact that only one professor had taught a transition-to-proof course and two of them had rather limited experience teaching upper-division mathematics courses that emphasize the reading and writing of proofs. A second limitation is that only seven elementary proofs were used. Another limitation is that for these professors the proofs were out of context in the sense that they were selected from courses these professors had not taught. Furthermore, one professor mentioned that she had difficulty doing the evaluations while being observed, and thus the interview format may have unduly influenced the professors' evaluation of the proofs. Despite these limitations, we see a rich complexity in the views and practices of these mathematicians in their assessment of students' proofs.

Acknowledgements The author wishes to thank the four professors for generously sharing their time and expertise, Keith Weber for his encouragement and helpful comments on earlier drafts, and the anonymous reviewers for many insightful and helpful comments.

Appendix

Task 1: Transitive Relation

Define a relation R on the set of real numbers by $x R y$ if and only if $x - y$ is an integer, that is, two real numbers are related if they differ by an integer. Prove that R is transitive.

Proof 1:

We must show that for $a, b, c \in \mathbb{R}$, if $a R b$ and $b R c$, then $a R c$.

$$\begin{array}{ll}
 a - b = m & m, n \in \mathbb{Z} \quad (\text{Def. } R) \\
 b - c = n & \\
 b = n + c & (\text{Add } c \text{ to both sides}) \\
 a - n - c = m & (\text{Substitute for } b) \\
 a - c = m + n & (\text{add } n \text{ to both sides}) \\
 \text{so } a R c &
 \end{array}$$

$\therefore R$ is transitive. ■

Task 2: Transitive Relation

Define a relation R on the set of real numbers by $x R y$ if and only if $x - y$ is an integer, that is, two real numbers are related if they differ by an integer. Prove that R is transitive.

Proof 2:

if $x R y$ and $y R z$ then $y R z$ Let $x, y, z \in \mathbb{Z}$
and assume $x R y$ and $y R z$. We know $x - y \in \mathbb{Z}$
and $y - z \in \mathbb{Z}$ let $k, c \in \mathbb{Z}$

$$x - y = k$$

$$y - z = c \rightarrow y = c + z$$

$$x - (c + z) = k$$

$$x - z = (k + c)$$

since $k + c$
is an $\mathbb{Z}$ then

$$x R z \quad \square$$

Task 3: Bijective Function

Let $\mathbf{Z}$ denote the set of integers and let $\mathbf{O}$ denote the set of odd integers. Prove that the function $f: \mathbf{Z} \rightarrow \mathbf{O}$ given by $f(n) = 2n + 1$ is bijective. Base your proof on the definitions of *one-to-one* and *onto*.

Proof 3:

We must show that $f(n)$ is one-to-one and that $\mathbf{Z}$ maps onto $\mathbf{O}$.

1) One-to-one

We must show that $f(n_1) = f(n_2)$ implies $n_1 = n_2$.

$$2n_1 + 1 = 2n_2 + 1$$

$$2n_1 = 2n_2 \quad \text{subtract 1 from both sides}$$

$$n_1 = n_2. \quad \text{divide by 2 on both sides}$$

$\therefore f(n) = 2n + 1$ is one-to-one. $\checkmark$

2) onto

We must show that for $a \in \mathbf{O}$ and $n \in \mathbf{N}$, $f(a) = n$ $b = 2a + 1$
 $\frac{b-1}{2}$

Using our function $f(n) = 2n + 1$, let $a = \frac{n-1}{2}$. Substituting

$$\text{for } a, \quad f(a) = \frac{2}{1} \left(\frac{n-1}{2} \right) + 1$$

$$f(a) = n.$$

$\therefore f(n) = 2n + 1$ is onto. $\blacksquare$

Task 4: Principle of Mathematical Induction

Use the Principle of Mathematical Induction to prove that if a is any real number other than 1, then for all positive integers n , $1 + a + a^2 + a^3 + \cdots + a^n = \frac{a^{n+1}-1}{a-1}$.

Proof 4:

Let a be any real number other than 1 and let $P(n)$ refer to the equation $1 + a + a^2 + a^3 + \cdots + a^n = \frac{a^{n+1}-1}{a-1}$. We must first find $P(1)$.

$$\begin{aligned} 1 + a &= \frac{a^2-1}{a-1} \\ 1 + a &= \frac{(a+1)(a-1)}{a-1} \\ 1 + a &= a + 1 \end{aligned}$$

Clearly these are equal, therefore $P(1)$ is true.

We must show that $1 + a + a^2 + a^3 + \cdots + a^n + a^{n+1} = \frac{a^{n+2}-1}{a-1}$.

First we shall look at the left hand side of the equation:

$$1 + a + a^2 + a^3 + \cdots + a^n + a^{n+1}$$

We know from the basis step that this is equivalent to:

$$\begin{aligned} \frac{a^{n+1}-1}{a-1} + a^{n+1} &\Leftrightarrow \frac{a^{n+1}-1}{a-1} + a^{n+1} \left(\frac{a-1}{a-1} \right) \\ &\Leftrightarrow \frac{a^{n+1}-1}{a-1} + \frac{a^{n+2}-a^{n+1}}{a-1} \\ &\Leftrightarrow \frac{a^{n+1} + a^{n+2} - a^{n+1} - 1}{a-1} \\ &\Leftrightarrow \frac{a^{n+2}-1}{a-1} \end{aligned}$$

After algebraic manipulation we see that both sides of the equation are equivalent. ■

Task 5: Subset

Prove that if A, B, C and D are sets with $A \subseteq B$ and $C \subseteq D$, then $A \times C \subseteq B \times D$.

Proof 5:

Let A, B, C, D be sets such that $A \subseteq B$ and $C \subseteq D$.

This means (by the def. of $\subseteq$) that

$$\forall x, x \in A \rightarrow x \in B$$

$$\forall y, y \in C \rightarrow y \in D$$

By def. of $\times$ $A \times C$ is

$$A \times C = \{x, y \mid x \in A \wedge y \in C\}$$

$$B \times D = \{x, y \mid x \in B \wedge y \in D\}$$

$$A \times C = \{x, y \mid \overset{\nearrow x \in B}{x \in A} \wedge \overset{\nearrow y \in D}{y \in C}\}$$

$$\text{So } \forall x, y \{x, y \mid x \in A \wedge y \in C\} \rightarrow \{x, y \mid x \in B \wedge y \in D\}$$

by def. this $A \times C \subseteq B \times D$ ■

Task 6: Incidence Geometry

Use the axioms below to prove the following theorem: Let l and m be two lines. If there exists a unique point P such that P lies on both l and m , then l and m are distinct and nonparallel.

Axiom 1. For every pair of distinct points P and Q there exists exactly one line l such that both P and Q lie on l .

Axiom 2. For every line l there exist at least two distinct points P and Q on l .

Axiom 3. There exist three points that do not all lie on any one line.

Proof 6:

For the purpose of proof by contrapositive let us prove that the theorem "Let l and m be two lines.

If there exists a unique point P , such that P lies on both l and m , then l and m are not distinct or parallel." will lead to a contradiction. So, let l and m be two lines.

In the case that the lines l and m are not distinct or in other words l is the same line as m , if this is true there would be at least two points where they cross (by axiom 2) as they are the same line and there would be no unique point P .

In the case the lines are parallel by definition of parallel the lines will never cross therefore there is no point P where they cross.

Therefore, through proof by contrapositive we have shown that for there to be a unique point P , lines l and m must be distinct and nonparallel. $\square$

Task 7: Inequality

If x is a positive real number, then the sum of x and its reciprocal is greater than or equal to 2, that is, $x + \frac{1}{x} \geq 2$.

Proof 7:

Multiply both sides by x .

Then $x^2 + 1 \geq 2x$, so $x^2 - 2x + 1 \geq 0$.

This means that $(x-1)^2 \geq 0$,
which is clearly true, so the
theorem has been proved. ▀

References

- Antonini, S., & Mariotti, M. A. (2008). Indirect proof: What is specific to this way of proving? *ZDM*, 40, 401–412. doi:10.1007/s11858-008-0091-2.
- Auslander, J. (2008). On the roles of proof in mathematics. In B. Gold & R. A. Simons (Eds.), *Proofs and other dilemmas: Mathematics and philosophy* (pp. 61–77). Washington, DC: Mathematical Association of America.
- Brown, D. E. & Michel, S. (2010). Assessing proofs with rubrics: The RVF method. In *Proceedings of the 13th Annual Conference on Research in Undergraduate Mathematics Education*, Raleigh, NC. Retrieved from http://sigmaa.maa.org/rume/crume2010/Archive/Brown_D.pdf.
- Chartrand, G., Polimeni, A. D., & Zhang, P. (2013). *Mathematical proofs: A transition to advanced mathematics* (3rd ed.). Boston: Pearson Education.
- Epp, S. S. (2011). *Discrete mathematics: An introduction to mathematical reasoning*. Boston: Brooks/Cole, Cengage Learning.
- Fukawa-Connelly, T. (2010). Modeling mathematical behaviors: Making sense of traditional teachers of advanced mathematics courses pedagogical moves. In *Proceedings of the 13th Annual Conference on Research in Undergraduate Mathematics Education*, Raleigh, NC. Retrieved from <http://sigmaa.maa.org/rume/crume2010/Archive/Fukawa-Connelly.pdf>.
- Harel, G., & Brown, S. (2008). Mathematical induction: Cognitive and instructional considerations. In M. Carlson & C. Rasmussen (Eds.), *Making the connection: Research and practice in undergraduate mathematics* (pp. 111–123). Washington, D.C.: Mathematical Association of America.
- Harel, G., & Sowder, L. (1998). Students' proof schemes: Results from exploratory studies. In A. H. Schoenfeld, J. Kaput, & E. Dubinsky (Eds.), *Issues in mathematics education* (Research in collegiate mathematics education III, Vol. 7, pp. 234–282). Providence: American Mathematical Society.
- Inglis, M., & Aberdein, A. (2016). Diversity in proof appraisal. In B. Larvor (Ed.), *Mathematical Cultures*. Birkhäuser Science (in press).
- Inglis, M., & Alcock, L. (2012). Expert and novice approaches to reading mathematical proofs. *Journal for Research in Mathematics Education*, 43, 358–390. doi:10.5951/jresmetheduc.43.4.0358.

- Inglis, M., Mejía-Ramos, J. P., Weber, K., & Alcock, L. (2012). On mathematicians' different standards when evaluating elementary proofs. In S. Brown, S. Larsen, K. Marrongelle, & M. Oehrtman (Eds.), *Proceedings of the 15th Annual Conference on Research in Undergraduate Mathematics Education* (pp. 67–72), Portland, OR. Retrieved from http://sigmaa.maa.org/rume/crume2012/RUME_Home/RUME_Conference_Papers_files/RUME_XV_Conference_Papers.pdf.
- 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. doi:10.1111/tops.12019.
- Lai, Y., Weber, K., & Mejía-Ramos, J. P. (2012). Mathematicians' perspectives on features of a good pedagogical proof. *Cognition and Instruction*, 30(2), 146–169. doi:10.1080/07370008.2012.661814.
- Lew, K., Fukawa-Connelly, T., Mejía-Ramos, J. P. & Weber, K. (2016). Lectures in advanced mathematics: Why students might not understand what the professor is trying to convey. *Journal for Research in Mathematics Education*, 47(2), 162–198.
- 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: National Taiwan Normal University.
- Mejía-Ramos, J. P., & Weber, K. (2014). Why and how mathematicians read proofs: Further evidence from a survey study. *Educational Studies in Mathematics*, 85(2), 161–173. doi:10.1007/s10649-013-9514-2.
- 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. doi:10.1007/s10649-011-9349-7.
- Moore, R. C. (1994). Making the transition to formal proof. *Educational Studies in Mathematics*, 27, 249–266.
- Selden, A., & Selden, J. (1987). Errors and misconceptions in college level theorem proving. In J. D. Novak (Ed.), *Proceedings of the second international seminar on misconceptions and educational strategies in science and mathematics* (Vol. III, pp. 457–470). Ithaca: Cornell University.
- Selden, J., & Selden, A. (1995). Unpacking the logic of mathematical statements. *Educational Studies in Mathematics*, 29, 123–151.
- Smith, D., Eggen, M., & St. Andre, R. (2011). *A transition to advanced mathematics* (7th ed.). Boston: Brooks/Cole, Cengage Learning.
- Speer, N. M., Smith, J. P., III, & Horvath, A. (2010). Collegiate mathematics teaching: An unexamined practice. *The Journal of Mathematical Behavior*, 29(2), 99–114. doi:10.1016/j.jmathb.2010.02.001.
- Starch, D., & Elliott, E. C. (1913). Reliability of grading work in mathematics. *The School Review*, 21, 254–259.
- Strauss, A. L., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. London: Sage.
- Thompson, D. R. (1996). Learning and teaching indirect proof. *Mathematics Teacher*, 89, 474–482.
- Thurston, W. P. (1994). On proof and progress in mathematics. *Bulletin of the American Mathematical Society*, 30, 161–177.
- Weber, K. (2001). Student difficulty in constructing proofs: The need for strategic knowledge. *Educational Studies in Mathematics*, 48, 101–119.
- Weber, K. (2008). How mathematicians determine if an argument is a valid proof. *Journal for Research in Mathematics Education*, 39, 431–459.
- Weber, K. (2014). *What is a proof? A linguistic answer to a pedagogical question*. In T. Fukawa-Connelly, G. Karakok, K. Keene, & M. Zandieh (Eds.), *Proceedings of the 17th Annual Conference on Research in Undergraduate Mathematics Education*, Denver, CO. Retrieved from <http://sigmaa.maa.org/rume/RUME17.pdf>.
- 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(1), 36–58. doi:10.1080/00461520.2013.865527.
- Weinberg, A., Wiesner, E., & Fukawa-Connelly, T. (2014). Students' sense-making frames in mathematics lectures. *The Journal of Mathematical Behavior*, 33, 168–179. doi:10.1016/j.jmathb.2013.11.005.