Attention to Detail: Norms for Proof Evaluation in a Summer Mathematics Program

Cody L. Patterson
The University of Texas at San Antonio

Xiaowen Cui
Texas State University

In this study, we explore the norms by which students and undergraduate mentors in a summer mathematics program evaluate proofs of theorems in number theory. By utilizing cognitive interviews during which students and mentors evaluate number theory proofs written by a hypothetical student, we find that for students as well as mentors, “rigor” is a dimension of mathematical acceptability of proofs distinct from, though related to, proof validity. Additionally, we find that both students and mentors frequently adhere to strict unwritten norms that govern how they believe proofs should be constructed and presented, and that these norms may be more rigid than the intended proof-writing norms of the mathematicians who teach in the summer program. This study suggests some potential challenges associated with the growing practice of asking undergraduate student graders to evaluate proofs written by students in introduction-to-proof courses.

Key words: Proofs, Proof Validation, Cognitive Research, Informal Mathematics Education

One of the goals of a typical undergraduate program in mathematics is to instill in students an understanding of mathematical proof: the purposes for which mathematicians use proof, the process by which mathematicians prove results, and the ways in which we decide whether an argument is valid and should be accepted by the mathematical community. However, students learn about disciplinary norms governing proofs in varied and sometimes idiosyncratic ways. An instructor in an undergraduate mathematics course may rule a proof produced by a student acceptable or unacceptable according to rules that have not been made fully explicit for students. For example, an instructor in an introduction-to-proof course may suggest that a proof of the formula for the sum $[1 + 2 + 3 + \ldots + n]$ using induction is more “rigorous” than the classic reordering argument that pairs this with the sum $[n + (n - 1) + (n - 2) + \ldots + 1]$. An instructor who says this may intend to signal to students that in formal mathematics, we are reluctant to accept reasoning that is hidden “behind” the ellipses in these expressions without careful investigation; however, students may infer from the instructor’s choice that in formal mathematics, rigid proof schemata such as induction are more “rigorous” than arguments that use flexible reasoning. As a result of exchanges such as these, students may develop the belief that the acceptability of a proof is determined more by the manner in which an argument is written rather than by the logical coherence of the argument itself. In this study, we investigate ideas about “rigor” and proof acceptability co-constructed by faculty, mentors, and students in a summer mathematics program, and document some cases in which students demonstrate norms for proof evaluation that may be inconsistent with those that the faculty who teach in the program intend to convey.

Background

There is a growing body of literature on proof validation, the process of deciding whether a proof is valid, as performed by mathematics students and instructors. Validation of proofs is an intellectually complex process requiring many different types of reasoning, including the construction of formal and informal deductive arguments and example-based reasoning (Weber, 2008). Harel and Sowder (1998) suggest that many students, by the time they begin learning
about formal mathematical proofs in college, have developed *external conviction proof schemes* in which validity derives from the “ritual of the argument presentation”; that is, degree to which a proof structurally resembles a formal argument scheme. Additionally, many students have *empirical proof schemes* and may be convinced of the validity of a conjecture by testing specific cases. These non-analytical proof schemes may interfere with students’ ability to assess the validity of nonstandard arguments. Selden and Selden (2003) suggest that exclusive attention to surface features of a proof, or failure to attend to the global logical structure of a proof, may cause students to make incorrect judgments about the proof’s validity.

Mathematicians use a variety of socially-determined criteria when evaluating a proof, such as whether they understand the concepts embedded in the proof, whether the argument is convincing, and whether the theorem being proven is consistent with the existing body of accepted mathematical results (Hanna, 1983). Courses that introduce students to proof-writing in mathematics often have the dual aims of helping students develop the ability to identify and produce valid arguments, and enculturating students into practices and disciplinary norms regarding the writing and evaluation of proofs that are typical among professional mathematicians. While several studies have addressed the mental processes involved in distinguishing valid proofs from invalid ones, relatively few have studied how students’ judgments of proof validity interact with their (possibly separate) assessments of whether a proof is acceptably presented. While mathematicians do not share a single common standard for evaluating the validity and presentation of proofs (Inglis et al., 2013), we view the instructor of a proof-writing course as an exemplar of a particular set of norms and practices that the course is intended to transmit, in part or in full, to its students, and are interested in the various pedagogical forces that influence how students assimilate these norms themselves.

In this study, we aim to address the following research questions:

1. What criteria, other than proof validity, do students who are learning to read and write formal mathematics use to judge proofs?
2. To what degree are novice proof writers able to distinguish flaws that make a proof less readable or less complete from those that render a proof invalid?
3. To what degree do novice students’ and their mentors’ norms for evaluating proofs conform to those of the community of professional mathematicians, as embodied by the instructors teaching their courses?

**Setting of Study**

Our study took place at a summer mathematics program for talented high school students in the United States. Students in the program are recruited from all geographic regions of the United States; while most enter the program with no prior formal experience in undergraduate mathematics, a few have learned to write proofs by participating in high school mathematics competitions.

During the program, students learn how to write mathematical proofs while taking a course in number theory; some students return to the program in subsequent summers and take other courses in undergraduate mathematics. Students are assigned to study groups, with each group supervised by a mentor who has attended the program for several years. Most mentors are undergraduate STEM majors at leading research universities, though some mentors are senior high school students.

Students in the program attend classes each morning and afternoon; during the evenings, they participate in extended study sessions with their groupmates, supervised by their mentors. They
write proofs of theorems in number theory and submit these proofs for “grading” by their mentors. Although these proofs do not receive numerical grades, mentors provide feedback on the proofs’ correctness and style, and sometimes recommend that students revise work that is incorrect or incomplete.

**Method**

We conducted a survey of the first-year students in the program, their mentors, and the mathematician faculty member who teaches the number theory course to determine whether participants in the program were familiar with the use of the word “rigor” as a criterion used to evaluate proofs, and to what extent participants view “rigor” as distinct from proof validity. When asked to agree or disagree with the statement, “When we talk about proofs, the word ‘rigorous’ has the same meaning as the word ‘valid,’” 6 respondents strongly disagreed, 13 moderately disagreed, 13 moderately agreed, and 4 strongly agreed. The instructor of the number theory course was among those who strongly agreed. When asked to explain the distinction between validity and rigor, if they believed such a distinction exists, many respondents suggested that validity points to whether a proof is correct, while rigor deals with the level of detail in a proof and the extent to which the argument explicitly states and justifies each step.

Based on the survey results, we selected two study groups in the summer program for follow-up interviews. We invited students and mentors from these two study groups, along with the mathematician who taught the number theory course, to participate in cognitive interviews in which they evaluated several proofs written by a hypothetical student. Because we selected interview subjects from study groups in which multiple students shared the view that a proof is rigorous if and only if it is detailed, we make no claim that the interview results are representative of the proof-evaluation norms of the entire student body of the summer program. However, we conjecture based on the results of the preliminary survey that the views of rigor and proof validity demonstrated by the students and mentors we interviewed were shared by many other students and mentors in the program.

In the interviews, we asked six students, their two mentors, and the mathematician instructor to review the first theorem shown in Table 1 on the next page, either produce a proof of the theorem or write some notes on the main ideas of the proof, and then evaluate three hypothetical student proofs of the theorem on three dimensions – validity, rigor, and understandability – each on a scale of 0 to 3. We then repeated this routine with the second theorem below. In asking interview subjects to rate these proofs, we hoped to gain additional insight about whether students and mentors made distinctions between validity and rigor in their evaluations of specific arguments, and about whether students’ and mentors’ overall evaluations of proof quality were consistent with the norms the instructor intended to set.

**Results**

Students’ and mentors’ responses to the proofs given in the interview, and in particular their assessments of the rigor of these proofs, varied considerably. Table 2 on the next page shows the ratings the interview subjects gave the three proofs of Theorem 1 for Validity, Rigor, and Understandability, along with the proof each subject rated the best overall and the most rigorous.
Table 1: Descriptions of Hypothetical Student Proofs

<table>
<thead>
<tr>
<th>Theorem</th>
<th>Proof</th>
<th>Description of Proof</th>
</tr>
</thead>
<tbody>
<tr>
<td>If $a$, $b$, and $n$ are integers such that $a &lt; b$ and $n &gt; 0$, then $an &lt; bn$.</td>
<td>1A</td>
<td>Detailed argument that omits essential step demonstrating that difference between $bn$ and $an$ is a natural number</td>
</tr>
<tr>
<td></td>
<td>1B</td>
<td>Correct argument that omits names of algebraic properties that justify steps</td>
</tr>
<tr>
<td></td>
<td>1C</td>
<td>Correct and detailed argument by induction on $n$</td>
</tr>
<tr>
<td>If $a$, $b$, $q$, and $r$ are integers with $b &gt; 0$ and $a = bq + r$, then $GCD(a, b) = GCD(b, r)$.</td>
<td>2A</td>
<td>Correct argument carefully using definition of GCD</td>
</tr>
<tr>
<td></td>
<td>2B</td>
<td>“The common divisors of $a$ and $b$ are the same as the common divisors of $b$ and $r$” argument</td>
</tr>
<tr>
<td></td>
<td>2C</td>
<td>Argument that makes incorrect inference</td>
</tr>
</tbody>
</table>

Table 2: Interview Subjects’ Ratings of Proofs of Theorem 1

<table>
<thead>
<tr>
<th>Subject</th>
<th>Proof 1A</th>
<th>Proof 1B</th>
<th>Proof 1C</th>
<th>Best</th>
<th>M Rig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student X1</td>
<td>V: 3</td>
<td>R: 2</td>
<td>U: 2</td>
<td>1A</td>
<td>1C</td>
</tr>
<tr>
<td>Student X2</td>
<td>V: 3</td>
<td>R: 2</td>
<td>U: 2</td>
<td>1A</td>
<td>1A</td>
</tr>
<tr>
<td>Student X3</td>
<td>V: 3</td>
<td>R: 2</td>
<td>U: 3</td>
<td>2B</td>
<td>2B</td>
</tr>
<tr>
<td>Student X4</td>
<td>V: 3</td>
<td>R: 2</td>
<td>U: 3</td>
<td>2C</td>
<td>2C</td>
</tr>
<tr>
<td>Mentor X</td>
<td>V: 2</td>
<td>R: 3</td>
<td>U: 1</td>
<td>1A</td>
<td>1A</td>
</tr>
<tr>
<td>Student Y1</td>
<td>V: 3</td>
<td>R: 3</td>
<td>U: 3</td>
<td>1A</td>
<td>1A</td>
</tr>
<tr>
<td>Student Y2</td>
<td>V: 3</td>
<td>R: 2</td>
<td>U: 1</td>
<td>2B</td>
<td>2B</td>
</tr>
<tr>
<td>Mentor Y</td>
<td>V: 2</td>
<td>R: 2</td>
<td>U: 3</td>
<td>1B</td>
<td>1B</td>
</tr>
<tr>
<td>Instructor</td>
<td>V: 2</td>
<td>R: 2</td>
<td>U: 3</td>
<td>1B</td>
<td>1B/1C</td>
</tr>
</tbody>
</table>

We asked each subject to explain his or her ratings after evaluating each proof. Several themes emerged during these explanations, suggesting discrepancies between proof-evaluation norms of students and those of their mentors, or discrepancies between norms of students and mentors and those of the mathematician teaching the number theory course.

Level of Detail Versus Essential Reasoning: The Case of Proofs 1A and 1B

We constructed Proofs 1A and 1B to differ in two essential ways. The first is that Proof 1A obfuscates its argument slightly by introducing a new variable $t$ that is equivalent to $n$, while Proof 1B is more direct. The second is that while Proof 1A is more generous in including names of algebraic properties used in the argument, such as the substitution principle, Proof 1B is the only argument that uses the class’s definition of “less than” (that $x < y$ if there exists a natural number $k$ such that $x + k = y$) to establish that $an < bn$. (Proof 1A states that because $at + kt = bn$, we have $at < bn$, without stating or showing that $kt$ is a natural number.) Thus while Proof 1A includes an overall greater level of detail, Proof 1B arguably gives more attention to the most crucial step of the argument.

While both mentors and the instructor noted the key omission in Proof 1A and rated the proof’s validity accordingly, all six students rated the proof’s validity a 3, and five of the six rated Proof 1A more rigorous than Proof 1B. Most of these five students stated that they rated
Proof 1A more rigorous because it more consistently stated reasons for the algebra steps in the proof and named the algebraic properties used. However, Student X3 mentioned specifically that while Proof 1B left out a greater number of details overall, Proof 1A left out reasoning that seemed more essential. This suggests that Student X3 had a way of thinking about proof quality that took into account the relative importance of various details.

Rigidity in Methods and Styles of Proof: The Case of Proof 1C

We constructed Proof 1C in order to see how interview subjects would respond to the hypothetical student’s use of induction to prove a theorem that can be proven directly using the definition of “less than” and the distributive property. The ratings and interviews suggested that while some students found Proof 1C to be the most rigorous of the three (despite preferring one of the other proofs), others considered the proof less rigorous than others in part because of the specific induction language used in the argument. Student X4, who gave this proof a rating of 2 for rigor, said, “It’s kind of like, why are you using induction for something that probably could save you some time?” She later clarified that the use of induction was not the reason for her rating, but rather that the author of the proof might have written a better argument had he or she chosen a different method. Mentor X also took issue with some of the language used in the induction proof; most of her concerns revolved around the fact that the argument, at various times, specified a value for \( n \) (e.g., for the base case) and proceeded to make statements based on that assumption rather than writing these statements in conditional form. For example, the base case assumed that \( n = 1 \) and later said that \( an < bn \); of this, Mentor X said, “That’s not like… that’s just not true… you would say, maybe, ‘thus, when \( n = 1 \), \( an < bn \),’ right. That language sort of implies that they don’t understand what a base case does, what the significance of a base case is.” It seems that in this case, Mentor X read the claim that “\( an < bn \)” at the end of the base case as a general claim rather than as an instantiated claim about the case \( n = 1 \). These comments, along with her other markings on the proof, suggested that Mentor X had adopted a rigid way of reading induction arguments that had some difficulty accommodating differences in how students might represent the inductive logic of a proof.

In addition, Student X3 expressed concern that parts of Proof 1C had been written as indented chains of algebraic steps rather than in paragraph form. When pressed on this, subjects suggested that some mentors in the summer program encouraged students to write proofs in paragraph form, rather than block-indenting sequences of calculations. This is one of several stylistic choices about proof-writing that students in the program may infer to be norms of the mathematical community, but that in fact are artifacts of the program culture and of the way mentors grade proofs. We have collected scanned images of the graded proofs of students in the summer program and hope to perform a more in-depth investigation of the norms for mathematical writing that mentors in the program convey to students.

“Logic” Versus Chains of Axiomatic and Definitional Reasoning: The Case of Proof 2B

We constructed Proofs 2A and 2B to differ in only one significant way. While Proof 2A explicitly uses the formal definition of greatest common divisor given in class (that \( d \) is the GCD of \( a \) and \( b \) if \( d \) is a natural number; \( d \) divides both \( a \) and \( b \); and if \( c \) divides both \( a \) and \( b \), then \( c \) is less than or equal to \( d \)), Proof 2B proves Theorem 2 by showing that \( a \) and \( b \) have precisely the same common divisors as \( b \) and \( r \), and thus concluding that these two pairs must have the same GCD. Under the surface lay a structural difference that, in our preliminary investigations, seemed to have some import for students and mentors in the program: while Proof 2A proceeds through a sequence of steps, most of which can be directly justified using axioms or definitions.
from the number theory course, Proof 2B ends with a verbal argument that appeals to the notion of GCD as the maximum of the set of common divisors of two integers, and attempts to persuade the reader that if two pairs of integers have the same set of common divisors, they must have the same GCD. The latter approach departs slightly from the norm of precise definitional reasoning to which students become accustomed early in the number theory course, though the number theory course itself includes a number of theorems (such as Fermat’s Little Theorem) whose proofs contain steps that appeal to nonlinear, *ad hoc* reasoning rather than axioms.

Of the five students and one mentor who finished reviewing the proofs of Theorem 2, only two students rated Proof 2B as entirely valid. Two students rated the proof’s validity a 2, and Student X4 and Mentor X rated the proof’s validity a 1. Student X1 rated the proof a 3 for validity but only a 2 for rigor; when asked to explain his ratings, he said, “some of the steps aren’t explained clearly, such as the end, where it says that… it’s sort of just explaining logic instead of using any theorem or previous axiom.” In the preliminary survey, two other respondents, including Student Y1, mentioned the use of “logic” as a possible source of disagreement among students and mentors over whether a proof is rigorous; one respondent stated that “sometimes our [mentors] tell us to use more algebra instead of just logic and assuming.”

Student X3, whose interview responses suggested the least commitment to rigid, style-based norms for proof acceptability among those students interviewed, rated the proof a 3 for both validity and rigor, but rated it a 2 for understandability. She suggested that the author of the proof should have constructed the set of common divisors of $a$ and $b$ and the set of common divisors of $b$ and $r$, and shown that these two sets are equal. Responses to Proof 2B suggest that the students interviewed were not entirely comfortable with arguments that depart from chains of reasoning based on axioms and previously proven theorems and that include steps for which the justification is fluid and unfamiliar.

**Norms of a Professional Mathematician: The Case of the Course Instructor**

After we completed interviews of students and mentors, we interviewed the instructor of the number theory course using the same protocol. Prior to rating any proofs, the instructor stated his position that “validity” and “rigor” are not separate constructs, and expressed concern that students might believe that rigor is a separate or stronger standard for proofs than validity. When rating each proof, he assigned the same rating for both validity and rigor.

The instructor initially rated each of the three proofs of Theorem 1 a 3 in all three categories; he changed his validity and rigor ratings for Proof 1A to 2 only after considering that the author of the proof made, in his view, two mistakes: failing to use the substitution principle correctly, and failing to observe that $kt$ is a natural number. He expressed little hesitation in giving ratings of 3 for Proofs 1B and 1C, and showed little concern about the justification steps skipped in Proof 1B.

When the instructor was given the opportunity to review Theorem 2 prior to seeing the proofs, he noted that he is familiar with two different approaches; he then described the approaches used in Proofs 2A and 2B. When shown Proof 2B, he expressed amusement that he had foreshadowed the argument given, and rated the proof a 3 in all three categories, as he had with Proof 2A.

We can envision two different hypotheses to explain the fact that the instructor awarded higher ratings for most of the proofs in the interview than the students and mentors did. One possible explanation is that the instructor’s extensive experience teaching number theory has familiarized him with the arguments provided and desensitized him to possible errors in logic
that may reveal unclear thinking on the author’s part. Another hypothesis is that the instructor, when rating a proof, was more interested in whether the author had a complete chain of reasoning linking the hypotheses to the desired conclusions than in issues of how the proof is written. At various points in the interview, the instructor pointed out instances of writing that he did not consider optimal, but openly dismissed these as inessential to the central issue of whether the proof under review was valid, in contrast with mentors and students who considered some of these issues (such as the explicit naming of axioms about the algebraic structure of $\mathbb{Z}$) crucial in evaluating the validity and especially the rigor of proofs. This contrast suggests a disparity between the proof-evaluation norms of students and mentors in the summer program and those of professional mathematicians as embodied by the instructor of the course.

**Discussion and Implications**

The interview results suggest that some students in the summer mathematics program had criteria for “rigor” or completeness of proofs distinct from the criteria they used to evaluate the validity of proofs. In some responses to interview questions, students rated proofs as valid, but indicated that they believed that their mentors may not find these proofs entirely acceptable. In some cases, students indicated that they themselves did not consider certain proofs (such as Proofs 1A and 1B) totally acceptable despite finding them entirely valid.

We conjecture that in some introduction-to-proof courses, instructors pay explicit attention to issues of style and detail in proof-writing in the belief that strict attention to these issues supports the development of students’ ability to read and write proofs. We make no claim that our study sheds light on the validity of this particular belief. However, we observe that in most of the interviews conducted in this study, subjects did not always make clear distinctions between flaws in a proof that render it more difficult to read or follow and flaws that render a proof invalid. Although the instructor of the number theory course gave little direct instruction on norms for writing and evaluating proofs in class, mentors provided regular feedback on students’ proofs and, based on a preliminary analysis of their markings, sometimes advised students to rewrite proofs in ways that would have yielded at most negligible improvements in their readability. We hope to ascertain through follow-up interviews of the subjects of this study whether the mentors’ feedback instilled in students norms for writing and reading proofs that are more rigid than those shared by most of the community of professional mathematicians. We also hope to conduct a more thorough analysis of the graded proofs we collected over the duration of the program to develop a more precise inventory of the norms that governed the mentors’ evaluations of proofs.

We believe that our study has implications for the teaching of undergraduate mathematics in light of the growing practice of appointing relatively inexperienced students to grade papers in courses that involve proofs, borne of funding decreases that have forced mathematics departments to shift some of this intellectually challenging work from graduate teaching assistants to undergraduate graders. At minimum, we suspect that some undergraduate graders tasked with grading proofs may benefit from professional development that encourages flexibility in evaluating different approaches and styles of proof presentation. We acknowledge, however, that such professional development may not mitigate the intellectual demands associated with evaluating whether unfamiliar arguments are mathematically valid.
References