Student Validations of Author-free AI-Generated Proofs

Dov Zazkis
Arizona State University

Steven Ruiz
Arizona State University

L. Marizza Stix
Arizona State University

We report on a small-scale clinical interview study of students' validation of AI-generated proofs. During these one-on-one interviews, students were tasked with using ChatGPT to generate proofs and validate those purported proofs. Consequently, neither the interviewer nor the participant knew if the justifications being validated constituted a proof before evaluating it. As such, no two participants validated the same proof. This required the development of novel post-interview analysis methods, which led to some valuable insights into students' proof validation processes. We view the primary contributions of this work to be both the methodology we developed, and our insights gained into how AI-generated proofs provide a different lens for students' validation and error detection.

Keywords: ChatGPT, proof validation, introduction to proof, logic

Proof validation has been a common line of inquiry in proof education research (e.g., Selden & Selden, 2003; Alcock & Weber, 2005; Panse et al., 2018). Typically, participants are presented with some purported proof and asked if it is valid. The justifications utilized within these studies usually fall into one of three categories: 1) researcher-generated justifications that are produced with validation in mind when being constructed (e.g., Selden & Selden, 2003), 2) mathematician-generated justifications that are created for non-validation purposes and then re-purposed by researchers (e.g., Czocher & Weber, 2019), and 3) student-generated proofs that are re-purposed for proof validation work (Moore, 2016). Notably, in some of these studies, the purported authorship of justification misaligns with the actual authorship. For instance, a researcher might generate a justification for participant consideration, but indicate that the justification under consideration has a student author (e.g., Davies et al., 2023).

One notable difference between on-the-spot ChatGPT-generated justifications and the human-generated justification typically found in proof validation studies is the error rate. In many of these studies, the justifications presented by researchers include only relatively few errors. For instance, a twelve-line purported proof presented by a researcher may contain one error, while the number of errors generated by ChatGPT are not dependent on the length of the purported proof and the nature of the errors is not predetermined. It is not uncommon for a short ChatGPT-generated proof to have a dozen errors and for the subsequent proof generated to be error-free. Consequently, when the justifications generated by ChatGPT do not constitute what we would consider a valid proof, there tends to be significantly more opportunities for students to notice a subset of those errors and deem the proof invalid.

In contrast, fewer errors inserted in human-generated justifications create a limited opportunity for students to notice those errors. With these proofs, student validation might hinge on whether a student notices a single error, fabricated by the researcher, that links two lines in a twelve-line proof. In studies that use low error-rate proofs, some researchers have reported that although students do not initially notice an error, when directed to examine a line where an error occurs, students tend to find it without further guidance (e.g., Alcock & Weber, 2005; Selden & Selden, 2003). We argue that expecting students to find a single intentional error in a dozen line purported proof is an unnecessarily high standard. Additionally, noticing such an error is a poor indicator of a student's skill in validating proofs. By using AI-generated proofs, which may

contain significantly more errors, we can better articulate what students view as errors and how adept they are at finding them.

Methods

Six Anonymous State University (ASU) participants participated in one-on-one clinical interviews. Each participant was either currently enrolled in or had completed a transition-to-proof course at the time of the interview. Each interview lasted approximately 60 minutes and was video and audio recorded. Participants were given access to a computer with ChatGPT open to engage with the interview tasks. Students received a printout of the ChatGPT prompt associated with each interview task, but otherwise blank. The handwritten annotations of all participant annotations were scanned to create digital artifacts.

Interview Tasks and Data Collection

Each interview involved the participant using ChatGPT to generate proofs of pre-prepared prompts and subsequently validate them. Each participant received two statements (shown below), one related to number theory and the other to set theory.

1) Show that $7^n - 2^n$ is divisible by 5 for all n.
2) Show that if A and B are subsets of X, then $(A \cap B)^C = A^c \cup B^C$.

Participants were instructed to prompt ChatGPT to generate three proofs to validate—two of statement one and one of statement two. Additionally, we stipulated that the two proofs of statement one use different proof methods. The participants were asked to use their own phrasing for each statement when prompting ChatGPT. Once ChatGPT provided an output, the participant was instructed to study the proof and to think aloud wherever possible. Finally, the participant was invited to speak about their thoughts on the proof's correctness and completeness. They were asked to explicitly cite the features of the proof that led to their decisions. At the end of each interview, the interviewer saved and downloaded the interaction. This downloaded interaction included video recordings of the interview, user inputs entered by the interviewee, and the subsequent outputs generated by ChatGPT. Afterward, these interactions were analyzed.

Data Analysis

ChatGPT generated the purported proofs in this study in direct response to the interviewees' actions. As such, each of the purported proofs generated was unique and dependent on the students' phrasing and its own programming. So, before studying the participants' validation skills, it was necessary to examine each purported proof irrespective of the participant's validation. This timing minimized the influence of the participants' perceptions on those of the expert. This process began with the second author studying each purported proof generated by ChatGPT to glean whether it was valid (from the research team's perspective) and document each error the proof might contain. We documented errors in calculation, grammar, and syntax. Finally, each proof was judged valid (from the research team's perspective) if the logical structure was sound and if it was free of errors that detracted from the strength of the argument. For example, a proof which contained algebraic errors upon which a proof by induction was based was deemed invalid. By contrast, a proof that misattributed a set theory principle to De Morgan's Laws was considered valid because the principle was correct and well-substantiated.

After the research team assessed each purported proof, we examined each participant's engagement with the proofs via the interview recordings. As the participant thought aloud during their assessment of each proof, we specifically noted any instance in which the participant noted a fault in the proof as well as the nature of the error. We also noted the participants' final

decision and justification about each purported proof's validity, correctness, and completeness. After complete annotations of the expert and participant perception of each proof were available, we compared and contrasted them with a particular focus on the errors identified by each and the final decisions on the proof's overall validity. While the expert's search for errors in each proof was meant to be exhaustive, we did not make a similar request to the participants.

Results

In general, students read each ChatGPT-generated proof and noted any errors they found. They commonly terminated their validation process if they found a single significant error or an accumulation of multiple errors that they deemed sufficient to deem the ChatGPT proof invalid. As such, we comment on the proportion of errors found by participants in relation to those found by the research team up until the point the participants stopped reading. The participants identified 25 of the 55 errors – 45% of the errors identified by the research team. This rate is similar to rates reported in validation studies with human-generated proofs (e.g., Alcock & Weber, 2005; Selden & Selden, 2003). As discussed earlier, those human-generated purported proofs typically contain one or zero errors. So, we view that statistic as gesturing toward a parallel between our results and the results of those studies.

Table 1 below compares the research team's validation to the student participants' validation. The instances where the participant and expert disagreed are signified with boxes. In all participant interviews, the expert deemed the first two proofs (of Statement #1) invalid and the third proof (of Statement #2) valid. As can be seen from the table, the participants' final validity assessments were consistent with the research team on 16 of the 18 proofs (89% agreement). Note that on one such disagreement, the participant made their decision based on the weak link between inferences in the proof. The researcher initially noted this lack of detail yet deemed the proof valid. The above agreement rate is notably higher than similar statistics noted in validity studies with human-generated proof tasks (Alcock & Weber, 2005; Selden & Selden, 2003, Panse et al., 2018). Given our small sample size, we cannot make broad statements regarding the statistical significance of this difference. However, we believe this is a result of the higher error rate in the AI-generated proofs. As such, we argue that AI-generated proofs may be more appropriate for studying students' error identification process, whereas human-generated proofs may be more appropriate for studying overall validation.

Table 1. Validation decisions of participants regarding ChatGPT-generated proofs

✓ = Valid Proof	✘ = Invalid Proof	☐ = Disagreement with Expert

	Student #1	Student #2	Student #3	Student #4	Student #5	Student #6
Pf #1a	✘	✘	✘	✘	✘	☑
Pf #1b	✘	✘	✘	✘	✘	✘
Pf #2	✓	✓	☒	✓	✓	✓

In general, participants read proofs looking for any perceived errors and did not appear to be focusing on a particular error type. Their validation processes included highlighting stylistic, logical, and calculation errors. Below is an excerpt from Student #2 investigating an argument based on induction in the first proof of Statement #1. The proof's printout can be found in Figure

1 (OpenAI, 2023). The numbers in square brackets correspond to portions of the proof indicated by Student #2 as he was speaking.

Student #2: It's doing the right process. This [1] is right… It's doing the $m + 1$ case, and then it's going to show that the m^{th} case implies that the $(m + 1)^{\text{th}}$ case is true…
Interviewer: So, where are you?
Student #2: Okay, so I think it lost a term here. I think this completely fell apart. Because we have the 5 + 2 here [2], but then the two term is missing [3]…No, that doesn't make sense because this [4] is a multiple of 7, 7 times 7 times 7 times 2. Those are prime. This [5] is 5 times 2 times 2 times 2 and 2, which are prime, so there's no way that you can make these [5] out of these [2] multiplications. So, I'm going to say that this is an invalid proof.

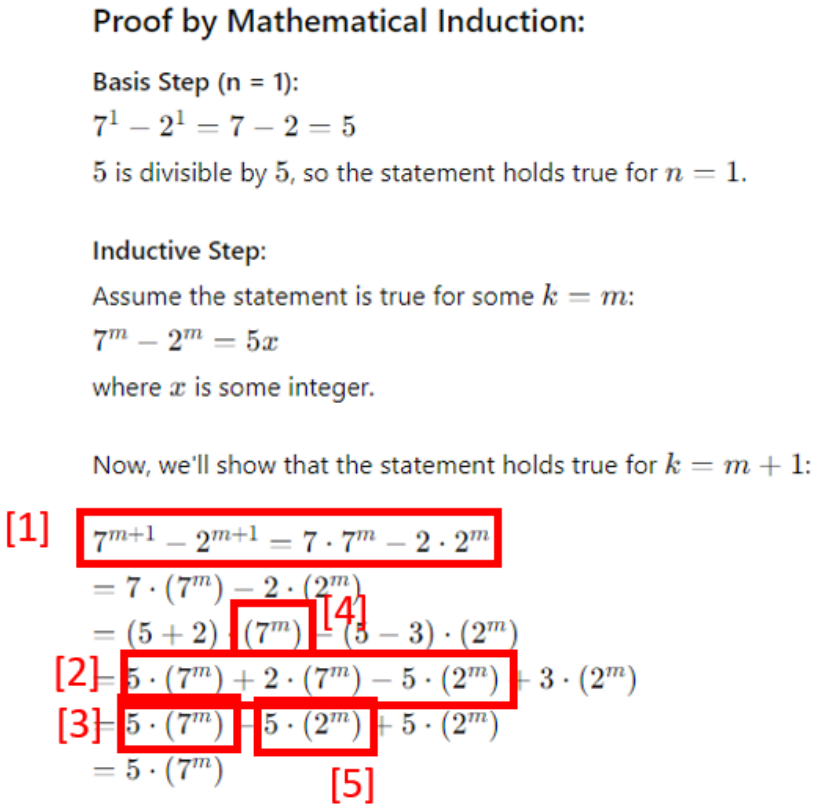

Figure 1. ChatGPT-generated proof #1 from interview with student #2

At the beginning of the transcript, Student #2 attends to the (overall) logical structure of the proof (Selden & Selden, 1995). He affirms that induction is an appropriate proof framework to prove the statement and then proceeds to examine the details of the proof. When he attends to the calculations in the sequence of inferences in the inductive step of the proof, he finds algebraic errors. Since the argument relies on those calculations, Student #2 deems the purported proof invalid. More specifically, since the calculations are what the proof uses to show the "the m^{th} case implies that the $(m + 1)^{\text{th}}$ case," a flaw in that part of the argument causes the purported proof to, "completely fall apart". Notably, the purported proof validly asserts that $7^m - 2^m = 5x$ for some integer x. However, this x, and by extension, the overall induction hypothesis, is not

utilized within the purported proof. This error was noted by the research team but not attended to by the student. Since he stopped reading the purported proof at the end of the transcript, we are unable to comment on whether he would have attended to it if he had been directed to that specific line of the proof (e.g., Selden & Selden, 2003, Alcock & Weber, 2005).

Discussion

We presented students with purported proofs generated on-the-spot by Chat-GPT. This has several consequences: 1) The participants knew that each purported proof was, in a sense, author-free. 2) Neither the participant nor the interviewer knew whether any given purported proof was valid at the time it was generated, and 3) Given the current nature of AI, the purported proofs under consideration tended to have more errors than human-generated proofs used in proof validation studies. The confluence of these consequences creates advantages and disadvantages in studying proof validation.

The method we reported has some notable disadvantages. First, researchers need additional work to assess whether each justification students see is valid before examining students' processes. This contrasts with work that shows the same purported proof to multiple participants across multiple interviews. Additionally, the uniqueness of each AI-generated proof makes comparing work across students more complicated, which is why much of what we discussed here focused on summary statistics rather than cross-interview comparisons.

Despite the potential drawbacks, AI-generated proofs present significant advantages by providing a lens into proof validation that is arguably not colored by authorship, or at least colored in a very different way. It is known that purported authorship of a justification influences validation processes (e.g., Davies et al., 2023; Harel & Sowder, 1998; Weber & Mejía-Ramos, 2011). For instance, a justification purported to be generated by a mathematician is likely to be viewed by participants with less suspicion than a comparable proof purported to be generated by a student. This authorship influences the field's understanding of participants' proof validation processes. For instance, consider what might happen if Student #2 had been given the same proof discussed in the results section but was instead told a mathematician generated it. In that case, it would be reasonable for them not to attend to the proof framework and check for its validity since a mathematician is unlikely to (unintentionally) use an inappropriate proof framework. Instead, we see student #2 determine whether the overall logical structure is appropriate before checking the details. That is, Student #2 examined the proof for various errors that could have occurred, which may be different than the subset of errors they might have looked at with different information about the author of the proof.

Notably, our overall error identification rate is similar to the correct validation rates reported in studies with human-generated purported proofs (Selden & Selden, 2003; Alcock & Weber, 2005; Panse et al., 2018). However, our reported correct validation rate was noticeably higher (ibid). These rates point to the intuitive finding that increasing the number of errors contained within a purported proof increases the likelihood that a student identifies enough of those errors to deem the proof invalid. Additionally, we argue that AI-generated purported proofs have significant potential for studying students' error identification processes, as there are more identifiable errors per line. Such exploration can increase the field's understanding of the relationship between students' overall error identification process and final validity assessments.

References

Alcock, L., & Weber, K. (2005). Proof validation in real analysis: Inferring and checking warrants. *The Journal of Mathematical Behavior, 24*(2), 125-134.

Davies, B., Miller, D., & Infante, N. (2023). The role of authorial context in mathematicians' evaluations of proof. *International Journal of Mathematical Education in Science and Technology*, *54*(5), 725-739.

Moore, R. C. (2016). Mathematics professors' evaluation of students' proofs: A complex teaching practice. *International Journal of Research in Undergraduate Mathematics Education*, *2*, 246-278.

OpenAI. (2023). *ChatGPT* [Large language model]. https://chat.openai.com/chat

Panse, A., Alcock, L., & Inglis, M. (2018). Reading proofs for validation and comprehension: An expert-novice eye-movement study. *International Journal of Research in Undergraduate Mathematics Education, 4*, 357-375.

Selden, J., & Selden, A. (1995). Unpacking the logic of mathematical statements. *Educational studies in mathematics*, *29*(2), 123-151.

Selden, A., & Selden, J. (2003). Validations of proofs considered as texts: Can undergraduates tell whether an argument proves a theorem? *Journal for research in mathematics education, 34*(1), 4-36.

Weber, K. (2008). How mathematicians determine if an argument is a valid proof. *Journal for research in mathematics education*, *39*(4), 431-459.

Weber, K., & Czocher, J. (2019). On mathematicians' disagreements on what constitutes a proof. *Research in Mathematics Education*, *21*(3), 251-270.