A Department-Level Protocol for Assessing Students' Developing Competence with Proof Construction and Validation

Tabitha Mingus
Western Michigan University

Mariana Levin
Western Michigan University

Abstract: *This methodological paper describes a protocol for assessing the development of students' competence with proof, created by the assessment committee within the Department of Mathematics at Western Michigan University. The assessment protocol we describe evolved over a period of 20 years and aims to collect information that is meaningful and actionable for improving mathematics instruction within the department. While there are several unique features of Western Michigan University that have created a context in which such work can be undertaken at the level of the department, we believe that this case will be of interest to mathematics departments seeking to find ways to measure their students' developing competence with proof.*

Keywords: Proof construction, Proof validation, Assessment

Introduction

The role of mathematical reasoning and proof in the undergraduate major's mathematical training is crucial. Becoming skilled at constructing mathematical proofs requires the mastery and coordination of a number of different skills and the creative ability and mental dispositions to bring all those skills and the appropriate content knowledge to bear on a particular statement. These skills and abilities can be roughly divided into two categories – comprehension of an argument and construction of an argument, with validation of a purported proof involving skills from both categories.

Early work on proof construction and comprehension established different classification schemes students used to understand mathematical proofs and typologies to classify student generated proofs (Balacheff, 1988; Harel and Sowder, 1998). More recently, Mejia-Ramos et al. (2012) developed a multidimensional model of that revealed the complexity of proof comprehension. They present seven aspects of understanding a proof and then generated possible items that could enable teachers/researchers to assess students' understanding of these facets of proof comprehension. The Mejia-Ramos et al. model incorporated many of the aspects identified by the Selden and Selden (1995, 2003) in their work on student validation of purported proofs.

Successful construction of a proof requires the coordination of skills related to logic, proof structure and types of argumentation, content knowledge, and creativity. Atwood (2001) identified seven obstacles in writing proof – three related to beginning the process and four to completing the process. Thus far, much work on proof construction has focused on course level interventions to improve student abilities to construct proofs (e.g., Selden, Benkhalti, & Selden, 2014). In terms of work on assessing student proof attempts, Andrew (2009) describes a "Proof Error Evaluation Tool (PEET)" that would help instructors provide consistent feedback to students regarding typical errors. Andrew identified two main categories of difficulties students had in writing proofs – eight types of errors related to proof structure and six related to conceptual understanding. He suggests that the tool be used as a rubric for the instructor and as a way of making the assignment of writing a proof more transparent to the student by providing them with the PEET as a reference when writing proofs and when examining the feedback provided by the instructor. Some of the "errors" identified relate less to the validity of the student's argument and more to the superficial issues that influence the readability of the argument (e.g. legibility of the writing, the lack of a diagram to guide the reader, lack of succinctness in argument).

Despite the growing literature on the difficulties students face in constructing and validating arguments, a challenge in the literature remains creating a way to define and track the development of proof competence over longer periods of time and experience, such as over the course of one's undergraduate program of study (Stylianides, Stylianides & Weber, to appear). It is this challenge that the protocol we present in this paper aims to address.

Development of a Protocol to Assess the Longitudinal Development of Proof Competence: The Case of Western Michigan University

Western Michigan University Mathematics Department is a medium-sized department within a public university with Carnegie classification of High Research Activity. Assessment has long been a priority for the university and the department. Several years ago, in response to calls at the university level, all departments were asked to identify one student learning objective that they would focus on collecting data about and improving with respect to. The mathematics department coalesced around the learning objective

"Students will have the ability to detect invalid arguments and construct different types of valid mathematical arguments at an appropriate level of sophistication."

In constructing an assessment protocol around students' development of proof competence, guided by the literature on proof, the assessment committee chose to approach the task of tracking students' developing proof competence by focusing on these two, distinct, but related competences of proof construction and proof validation.

Task Criteria

The first step in putting our department-wide assessment protocol in place was to find candidate tasks. We had several criteria for the sort of tasks we needed to create. The tasks needed to be suitable for students at many levels (i.e. the content of the task should engage high school math and not content students would not have experience with until later in their programs). At the same time, the task needed to allow us to see growth in sophistication of both proof approach and growth in sophistication of ability to communicate mathematically. We were particularly interested in tasks that would submit to several different approaches. Finding tasks that worked well for our purposes took several iterations and pilots. In constructing items to assess our dual objectives of proof construction and validation, we used the data from our proof construction task to generate sample arguments that became the basis of our proof validation task.

Implementation Cycle

We collect data from all students in our program every semester. In the fall semesters, we administer a proof construction task and in the spring semester students work on a proof validation task. Our current protocol involves a two-year cycle of tasks. We knew that we did not want to give the *same* exact task every year because students would begin to know the task and be less generative in their methods. Our data collection efforts are aimed at students in our program (e.g., math majors). However, in many cases, a student may not declare a mathematics major until later in their undergraduate career. Thus, in order to collect data on math majors' proof development, we also sample our lower division courses broadly and collect data from all students enrolled in our courses. The timing of the assessments is also deliberate. We aim to conduct the fall assessments early in their program before having the influence of particular coursework. Thus, we aim to administer the fall assessment within the first weeks of the fall semester. We made a conscious decision to administer the spring assessment as late as possible in the spring semester to allow the maximal time for growth over the course of a single year.

Assessing Performance: The Development of Proof Construction Rubric

Generating the data on the proof construction and validation tasks led to the need to have a systematic way to score and interpret the data. The validation tasks are multiple choice (students evaluate several arguments that are given to them) and are thus easy to score and interpret. The proof construction task, however, presents many more issues with respect to scoring and interpretation. Discussions at the department level about what was valued in assigning codes to student work led to discussions about how to handle/interpret issues such as how students communicated their arguments (symbolic versus with words) or whether written work revealed that students had the kernel of the idea of a proof, but were not yet able to complete it. Engaging the faculty in looking at students' proof work was extremely generative in that it revealed interesting and significant differences among faculty in their expectations and interpretations of student work. Thus, the effort to articulate what was valued by faculty in students' proof work and to develop a common language for making consistent decisions about student-generated proof led directly into the generation of the proof rubric. After many iterations and sessions that involved testing the emerging rubric against student data collected in past iterations, as of Fall 2015 we had a working prototype of the protocol.

The rubric that the department is currently using can be found in the appendix. As an overview, we point out that the rubric can be thought of in three main sections. The first section of questions concern whether or not the proof was valid and whether or not it was communicated at a very high level. We consider an answer of yes to these two questions to be the "high bar" that should be reserved only for the papers that meet our top expectations. If the paper receives a yes to these two questions, the following questions on the rubric do not need to be answered. The next section of the rubric contains questions that pertain to establishing whether markers of minimal growth towards proof writing competence exist in the student work sample being scored. In contrast to the first section, we see this section as a kind of "low bar" that lets us see whether the student's proof attempt contained initial chains of argument or evidence of mathematical reasoning. The final section of the rubric pertains to specific challenges or hurdles to writing a valid proof that are common reasons why the proof sample did not rise to a "yes, yes" profile. These include questions related to whether the proof sample contained an algebra error, or whether students made unwarranted assumptions, or did not provide sufficient evidence for the claims they were making in the argument they were presenting.

Faculty Involvement in Scoring

Collecting longitudinal data on mathematics majors' developing proof competence is a large endeavor. Our aim was to create a data collection and analysis protocol that would be sustainable over the long term in the department and that would inform conversations about how to improve curriculum and pedagogy. Thus, a high amount of faculty involvement across the entire department is necessary for the ongoing success of the protocol. As we geared up to administer the assessment to our focal classes in Fall 2015, the assessment committee held faculty development sessions to train faculty on the interpretation and use of the common rubric. The faculty development sessions involved the use of sample student work generated in the pilot phase and scoring it using the rubric (then coming together and discussing discrepancies in interpretation). The assessment committee created open times for faculty to meet and discuss scoring with committee members. Such opportunities were important as qualitatively analyzing and coding student work is a practice that is not familiar to many mathematics faculty members (who may be more familiar with grading student work solely for correctness and assigning points).

Assignment Protocol and Resolving Discrepancies in Scoring

Early on in our assessment work, the task of scoring student work samples fell on the department assessment committee. As mentioned above, in order to move towards a more sustainable model, we have sought ways to fairly distribute the work of scoring. In our most recent iteration of the two-year cycle, we used the following assignment procedure for scoring the tasks. Lower level courses were assigned to be scored by one person (not the instructor). This is because we were aware that we were casting a wide net in collecting this data and that much of the student work generated would not be by students who ended up pursuing mathematics majors. For any course aimed specifically at mathematics majors, we decided to have two independent scorers (not the instructor). Our general practice was to have a member of the assessment committee be one of the two independent scorers for papers from upper level courses. Once assignments had been made of which faculty would be scoring (roughly an equal number distributed to each faculty member, with the committee having more responsibility), we sent faculty members a link to the scoring rubric in google forms. This allowed for a master spreadsheet of all proof scores to be generated automatically.

Having multiple people score the upper level student papers opened up the possibility that individuals will diverge in their scoring of the student work. For each person marking the paper, the assessment committee generated a spreadsheet (directly from google forms) that included the student's identifier, the person's name who marked the paper, and then entries for every element of the rubric (displayed in 1's for yes and 0's for no). Then, for each statement in the rubric, the assessment committee reviewed differences among the scores. Electronic copies of all exam papers by class were saved to a common workspace and the committee went directly to the pdfs of the student work in order to discuss the meaning of discrepant scores. When two committee members were involved in marking the paper, it was possible to have an immediate discussion that resulted in resolution of the discrepancy. If a scoring discrepancy involved a faculty member not on the committee and did not appear to be an obvious error, the committee assigned .5 to each score given instead of assigning a 1 or a 0.

Reporting the Findings

Currently, analyses presented to the faculty of student results on the assessment (usually presented the semester after data is collected) have included a cross sectional analysis of students' performance on particular competence within the rubric (e.g., "X% of students in Course Y or below were able to produce valid arguments on the assessment item.") Over time, we are interested in reporting patterns in longitudinal for students in our program.

Discussion

In the above sections, we have described in detail the protocol that we use to assess the development of students' competence in generating and validating proofs. There are a number of issues that are under ongoing discussion. While the development of proof protocol that we have put in place and are implementing is powerful, there are some limitations of the choices that we have made in developing our protocol. For example, one thing that we would like to learn more about is how students work to prove statements that are more closely related to the content that they are learning in their upper division coursework. By design, the statements that students engage with in our longitudinal assessment needed to be independent of the coursework that students take in our program. We are exploring the possibility of augmenting our protocol with course-embedded assessments that would be asked of students as part of their normal coursework but collected as part of the departmental profile of their developing proof competence. Other future directions include conducting follow-up interviews to establish alignment between scores on students' written work and what we assess as their understanding in an interview setting in which we can probe their thinking. With respect to faculty development, interview data may be useful for engaging faculty in analyzing student work. The topic of math faculty development around the pedagogical implications of what can be learned from student work (including the student work generated by our proof assessment protocol) is a topic for ongoing investigation.

References

Atwood, P. R. (2001). Learning to construct proofs in a first course on mathematical proof. *Unpublished Doctoral Dissertation.*

Balacheff, N. (1988). Aspects of proof in pupils' practice of school mathematics. *Mathematics, teachers and children*, *216*, 235.

Harel, G., & Sowder, L. (1998). Students' proof schemes: Results from exploratory studies. *Research in collegiate mathematics education III*, 234-283.

Mejia-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*(1), 3-18.
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*, 4-36.
Selden, J., & Selden, A. (2009). Understanding the proof construction process. In *Proceedings of the ICMI Study 19 Conference: Proof and proving in mathematics education* (Vol. 2, pp. 196-201).
Selden, J., Benkhalti, A., & Selden, A. (2014). An analysis of transition-to-proof course students' proof constructions with a view towards course redesign. In *Proceedings of the 17th Annual Conference on Research in Undergraduate Mathematics Education* (pp. 246-259).
Stylianides, G.J., Stylianides, A.J., & Weber, K. (2017). Research on the teaching and learning of proof: Taking stock and moving forward. In J. Cai (Ed.) *Compendium for research in mathematics education* (pp. 237-266). Reston, VA: National Council of Teachers of Mathematics.

Appendix: Proof Construction Rubric

Rubric Section 1

Question 1 -- Is the proof valid? A valid proof should be free of algebraic errors, unjustified claims, missing cases, and the imposition of additional hypotheses. Students who use the fact that a^2-a+1 has no real roots must justify this claim.

Question 2 -- Does the student sufficiently communicate their ideas (whether correct or incorrect)? Is it easy to follow their line of thinking or interpret what they have done? For example, if a student wrote "I don't remember what a reciprocal is, but if I did, I would assume that there is a real number such that the sum of it and its reciprocal equals one," it would be coded "yes." If the student is only reiterating or clarifying the problem, then code as "No."

Rubric Section 2

This section should be marked only if the student did NOT write a valid proof ("No" on Question 1).

Question 3 --- The student appears to understand that the problem is about a + 1/a compared to 1 The rest of their work, if any, can be at any level. A student who writes something like a * 1/a = 1 did not correctly interpret the problem.

Question 4 -- The student engaged with the problem, doing some work (possibly incorrect) beyond just interpreting or rephrasing the statement There is some evidence that the student engaged with the problem and persisted in their attempt. For example the student tried out specific numbers, sketched a graph, or did some algebraic fiddling (beyond rote steps like adding or simplifying fractions without a comparison to 1).

Question 5 -- Some mathematical reasoning exists (possibly built from incorrect assumptions or definitions)

Question 6 -- The student used words to express at least one complete idea in support of their argument If the student is only reiterating or clarifying the problem, then code as "No."

Rubric Section 3

This section should be marked only if the student correctly interpreted the problem ("Yes" on Question 3).

Question 7 -- The student made an algebraic error, or their use of language or notation interferes with progress These errors could range from a simple sign error to conceptual errors such as extending the zero product property to another integer.

Question 8 -- The student ignored cases or imposed additional hypotheses (explicitly or implicitly) For example, a student may successfully complete the proof under the assumption that a is strictly positive. Or a student may cover the case $a < 0$ and after this implicitly assume that a is nonnegative, for example stating that if $a < 1$ then $1/a > 1$. Also code proof by example(s) as "yes."

Question 9 -- The student made claims that they did not attempt to justify Examples of "yes": a student wrote "the equation $a^2 - a + 1 = 0$ clearly has no real solutions and so there cannot be any such real number" or arrived at the equation $a^2 - a + 1 = 0$ and then simply stated "therefore there cannot be any real number such that the sum of it and its reciprocal is 1." Attempting to justify a claim, but doing so incorrectly or incompletely, would be coded "no."

Question 10 -- The student introduced a framework that *could* be used to write a successful proof Manipulating the expression $a + 1/a$ (vs. the equation $a + 1/a = 1$) is exploratory work and not a framework. Introducing a multivariable scheme is unlikely to lead to a successful proof, so also rate this as "No." For a "yes", unresolved issues or gaps in the proof (including missing cases) could be resolved using skills that the student has either demonstrated already, or which are easily accessible to a calculus student (e.g., testing trivial cases). Simply testing the values $a = 1, 2, 3$, etc. would be coded "No."