

Developing and validating proof comprehension tests in undergraduate mathematics

Juan Pablo Mejía-Ramos¹, Kristen Lew², Jimmy de la Torre³, Keith Weber¹
¹Rutgers University ²Arizona State University ³The University of Hong Kong

Abstract

In this article, we describe and illustrate the process by which we developed and validated short, multiple-choice, reliable tests to assess undergraduate students' comprehension of three mathematical proofs. We discuss the purpose for each stage and how it benefited the design of our instruments. We also suggest ways in which this process could be employed by other researchers to develop and validate their own reliable proof comprehension tests.

Key words: assessment, proof comprehension, proof reading

1. Introduction

Justification and proof are central to the discipline of mathematics and the practice of mathematicians. Dreyfus (1990) argued that the practice of proving is the characteristic “that most clearly distinguishes mathematical behavior from behavior in other scientific disciplines” (p. 126). Numerous influential mathematics educators have advocated that justifying and proving should play prominent roles in all mathematics classrooms (e.g. Hanna, 1995; Schoenfeld, 1994). Notably, mathematics educators have argued that proofs should not just be used to convince students of the truth of mathematical statements, but also as a tool to provide explanation, facilitate communication, and promote understanding in the mathematics classroom (de Villiers, 1990; Hanna, 1990; Hersh, 1993; Knuth, 2002; Weber, 2010a). While proof is expected to play a prominent role in mathematics classrooms at the school level (e.g. Knuth, 2000), it assumes even greater importance at the university level, especially in the advanced proof-oriented courses that mathematics undergraduate students take.

In advanced mathematical courses, much emphasis is given to improving students' ability to write proofs about the concepts that were covered in the course. Weber (2001) pointed out that “this ability is often the primary goal of advanced mathematics courses and typically the only means of assessing students' performance” (p. 101). Not only is proving an activity in

which students are expected to engage, proof is also the most common form of instructional explanation that professors provide for their students (Lai, Weber, & Mejía-Ramos, 2012) and is the predominant way that mathematics is communicated in university classrooms. It is widely accepted that advanced mathematics courses are largely taught in a ‘definition-theorem-proof’ format, in which lectures mostly consist of the professor introducing definitions of concepts, stating theorems about those concepts, and then presenting proofs of these theorems (Mills, 2011; Weber, 2004; see also Dreyfus, 1991). Textbooks for advanced mathematics courses present material in a similar fashion (cf. Raman, 2004). An important assumption underpinning this pedagogical practice is that students are successfully able to learn mathematics from the proofs that they read. Yet as mathematicians and mathematics educators observe, students’ understandings of the proofs that they read are rarely assessed in a meaningful way (e.g. Conradie & Frith, 2000; Cowen, 1991). This is due, in part, to the dearth of valid assessments that measure such understandings (Weber, 2012).

In this article, we describe and illustrate the process by which we developed and validated reliable tests to assess readers’ understanding of three proofs, and suggest ways in which this process could be employed by other researchers to develop and validate their own reliable assessments.

2. Literature review

2.1 Research on proof reading in mathematics education

In the last three decades, there has been a vast amount of research on proof in mathematics education. Based on a systematic review of the literature, Mejía-Ramos and Inglis (2009) found that the large majority of studies on proof investigated students’ proof construction and only a small minority focused on students’ reading of proofs. Mejía-Ramos and Inglis’ findings are consistent with the calls from other researchers who remarked upon the scarcity

of research on reading proofs and called for more research into this area (e.g. Alcock & Weber, 2005; Hazzan & Zazkis, 2003; Konior, 1993; Mamona-Downs & Downs, 2005; Selden & Selden, 2003).

A second finding from Mejía-Ramos and Inglis' (2009) literature review was that most articles concerning the reading of proof focused on students' proof evaluations, in which students were typically given an argument and asked to evaluate it in some way (e.g. Is this argument valid? Is this the type of argument you might produce? Are you convinced by this argument?). In general, most studies of this type used students' evaluations of arguments as a lens into their conceptions of conviction and validity (e.g. Healy & Hoyles, 2000; Martin & Harel, 1989; Segal, 1999; Weber, 2010b), with a smaller number of studies seeking to determine if students or teachers could distinguish between valid or invalid arguments (e.g. Alcock & Weber, 2005; Knuth, 2002; Powers et al., 2010; Selden & Selden, 2003). These types of studies have given us reasons to believe that students may learn little from the proofs that they read. For instance, there is a growing body of studies showing that many mathematics majors are unable to determine if a mathematical argument is valid (e.g. Alcock & Weber, 2005; Inglis & Alcock, 2012; Selden & Selden, 2003; Weber, 2010b). If students cannot determine if an argument is valid or not, it is unlikely they understood the argument very well. We also have anecdotal evidence that mathematics majors do not understand or see no value in the proofs that they read in their advanced mathematics courses. Cowen (1991), a mathematician, attempted to convince his audience that students do not understand the proofs they read with the following argument:

If you need evidence that we have a problem, let one of your B students ... explain the statement and proof of a theorem from a section in the book that you have skipped. My students, at least, do not have the innate ability to read and understand what they have read. (p. 50)

If we accept that the ability to read and understand proofs is a skill that mathematics undergraduate students should develop, and we believe Resnick and Resnick's (1992) credo that students will not learn what we do not assess, it follows that it is imperative to develop ways to assess mathematics majors' understanding of the proofs that they read. Based on his interviews with mathematicians, Weber (2012) found that this was not done in a meaningful way. The interviewed professors claimed that when they assessed students' understanding of a proof, they usually had students superficially alter the proof to prove a very similar theorem or they simply asked the students to reproduce the proof by rote; often these professors did not assess students' understanding of the proofs that they read at all (see also Conradie & Frith, 2000; Rowland, 2001). However, the mathematicians interviewed by Weber (2012) also expressed dissatisfaction with their methods of assessment, as these encouraged rote memorization or meaningless symbol manipulation and did not measure understanding in a meaningful way. These mathematics professors claimed that they wanted to assess students' understanding of proof meaningfully, but lacked the tools to do so.

These findings raise the following questions: How should this assessment be done? What types of understanding should be measured? How can we generate items that measure this understanding? And how can we tell if these measures are valid and reliable?

2.2 Research on proof comprehension assessment in mathematics education

Prior to our work in the area, we are not aware of the existence of *systematic* ways of assessing students' understandings of proofs specifically in undergraduate mathematics courses. However, there have been two important contributions in this area in the research literature. Yang and Lin (2008) developed a model of Reading Comprehension for high school Geometry Proofs (RCGP) that defined four levels of understanding: *surface* (i.e. understanding the meaning of terms and statements), *recognizing the elements* (i.e. knowing whether a statement was an axiom, assumption, definition, or deduction), *chaining elements*

(i.e. seeing how new statements are deduced from previous ones), and *encapsulation* (i.e. viewing the proof as a whole to comprehend the higher level ideas in the proof). Yang and Lin developed specific items to assess the first three levels of understanding of a given proof.

From our perspective, three things are noteworthy about Yang and Lin's (2008) model and assessment methods. First, their model was the first to explicitly define and structure what it could mean to understand a proof in a specific area in mathematics, and how one would measure such understanding. Second, since developing this assessment model, these researchers were able to use this model to address important research questions about geometry proofs that had hitherto been ignored, such as what skills and attitudes correlate positively with proof comprehension in geometry (Yang, 2012) and what learning trajectories might students traverse as they come to understand geometry proofs (Lin & Yang, 2007). This illustrates how assessment models of proof comprehension may spawn valuable research. Third, while some of the ideas in Yang and Lin's model are pertinent to proofs at the undergraduate level, we do not believe that the model by itself is sufficient to probe students' understanding of a proof in an advanced mathematics class. For instance, Yang and Lin did not attempt to assess if students achieved the highest level of understanding in their model, which consisted of viewing the proof as a whole to comprehend its higher-level ideas. While this type of understanding might not be a central concern for a high school geometry teacher, we contend that skills such as being able to summarize a proof or being able to flexibly apply the methods of a proof to prove a new theorem are crucial skills for students in advanced mathematics classes. Further, there are also logical nuances that are present in some undergraduate proofs that are not accounted for in Yang and Lin's assessment model, such as how a proof by contradiction or a proof by cases should be understood. From Yang and Lin's perspective, this is not important as such proofs are rare in high school geometry classes, but they are common in undergraduate mathematics classes. As we describe shortly, our

assessment model for proof comprehension builds upon the model of Yang and Lin, adapting it to the context of undergraduate mathematics.

Conradie and Frith (2000) made another important contribution by directly addressing the issue of proof comprehension tests in undergraduate mathematics. In addition to stressing the need to develop these tests, these researchers provided comprehension tests for two different proofs. While their items were intriguing and called mathematics educators' attention to an underrepresented area of research, we note that these tests seemed to be created in a somewhat *ad hoc* manner, and it was unclear how these items were generated or what specific skills or understanding each item was designed to assess.

In earlier work, we engaged in theoretical analysis aimed at determining the types of understandings one would want to assess with proof comprehension tests, and how one could systematically generate items to assess these understandings. We defined *proof comprehension* in terms of the types of understandings that were valued by mathematicians and mathematics educators in the literature. The main product of such work was an assessment model describing seven different types of items one could generate to assess a reader's understanding of a proof in advanced mathematics (Mejía-Ramos et al., 2012). We separated these types of items into two groups. The first group concerns *local* understandings of the proof, meaning that these types of questions can be answered by focusing on a small number of statements within the proof. These questions are concerned with the meaning of these specific statements, their logical status in the proof, and how they are justified within the proof. We addressed local understandings of the proof by adapting the first three components of Yang and Lin's (2008) model for geometry proofs.

The second group concerns *holistic* understandings of the proof. In contrast with local understandings, one would not be able to answer items assessing holistic understandings of a proof by focusing on a small number of statements in that proof. Instead, to answer this type

of items a reader would have to infer the ideas or methods that motivated a large part of the proof, or the proof in its entirety. The holistic understandings relate to the “encapsulation” level in Yang and Lin’s (2008) model (which, as we noted, was not developed in their paper). To develop the holistic components of proof comprehension, we reviewed the literature and interviewed nine mathematicians on how they understood proofs in their own professional practice (see Weber & Mejía-Ramos, 2011), why they presented proofs to their students, and what they hoped their students would learn from the proofs that they read. If a particular facet of understanding was mentioned by multiple mathematicians and was discussed in the mathematics education research literature, we incorporated this type of understanding into our assessment model. The resulting components of our model are summarized below (see Mejía-Ramos et al., 2012 for a complete description of the methodology for generating this model and a thorough description of each component).

The local types of assessment items are:

- *Meaning of terms and statements*: items of this type measure students’ understanding of key terms and statements in the proof.
- *Logical status of statements and proof framework*: these items assess students’ knowledge of the logical status of statements in the proof and the logical relationship between these statements and the statement being proven.
- *Justification of claims*: these items address students’ comprehension of how each assertion in the proof follows from previous statements in the proof and other proven or assumed statements.

The four holistic types of assessment items address students’ understanding of the proof as a whole. These assessment types are:

- *Summarizing via high-level ideas*: these items measure students' grasp of the main idea of the proof and its overarching approach.
- *Identifying the modular structure*: items of this type address students' comprehension of the proof in terms of its main components/modules and the logical relationship between them.
- *Transferring the general ideas or methods to another context*: these items assess students' ability to adapt the ideas and procedures of the proof to solve other proving tasks.
- *Illustrating with examples*: items of this type measure students' understanding of the proof in terms of its relationship to specific examples.

Instead of having to generate test items in an *ad hoc* manner, lecturers, textbook writers, and researchers can use the template questions of our model to generate open-ended items that measure students' proof comprehension along dimensions that are valued by both mathematicians and mathematics educators. Indeed, researchers have used this model to investigate questions about proof comprehension (e.g. Alcock et al, 2015; Fuller et al., 2014; Hodds et al., 2014; Mantini et al., 2012; Weber et al., 2012). However, this way of using the model has two shortcomings. First, these open-ended items can be time-consuming to generate and grade, which may limit their utility in teaching and research situations involving a large number of test takers. Second, the validity and reliability of these tests is uncertain. The purpose of our current project is to address both shortcomings by designing and validating short, multiple-choice, reliable proof comprehension tests for proofs commonly presented in undergraduate mathematics courses.

2.3 The development of assessment instruments in undergraduate mathematics education

The last 30 years have seen a boom in the development and validation of tests designed to assess students' understandings of fundamental concepts in STEM fields at the undergraduate level (Deeds & Callen, 2006). Commonly known as “concept inventories”, these tests focus on measuring students' deep understanding of these concepts, while avoiding conflating it with the computational/procedural aspect of such understanding. Probably the most famous example is the Force Concept Inventory (Hestenes, Wells, & Swackhamer, 1992), a multiple-choice test designed to assess students' understanding of the foundational concepts of Newtonian physics.

Influenced in part by the significant impact that the Force Concept Inventory had in physics education, mathematics educators have engaged in the development and validation of similar types of tests in undergraduate mathematics. In particular, Carlson, Oehrtman, and Engelke (2010) developed the Precalculus Concept Assessment (aimed at assessing students' understanding of concepts that are fundamental for learning calculus, such as the concept of function) and Epstein (2013) developed the Calculus Concept Inventory, “a test of conceptual understanding (and only that — there is essentially no computation) of the most basic principles of differential calculus” (p. 1018).

It is important to note a significant difference between the aim of these concept inventories and that of the proof comprehension tests we developed: while concept inventories address students' understanding of a collection of central concepts in a given area, our proof comprehension tests address students' understanding of specific proofs. That is, at this stage, our proof comprehension tests are not designed to assess the extent to which a student understands proofs in general (or even proofs of a particular type, or in a given area of mathematics), but how the student understands the specific proof being tested. While the competencies involved in understanding these specific proofs may be indicative of more

general proof comprehension competencies (e.g. the ability to understand proofs by contradiction in elementary number theory), further research is required to delineate what are those more general competencies. We believe the kind of tests we have developed can play a crucial role in determining if it makes sense to study a more general notion of proof comprehension as a construct in its own right.

3. Development of the proof comprehension tests

Our research team has so far developed comprehension tests for proofs of the following three theorems:

Theorem 1: The set of prime numbers is infinite.

Theorem 2: Every third Fibonacci number is even.

Theorem 3: The open interval $(0,1)$ is uncountable.

Our choice of theorems was made to maximize the utility of the corresponding proof comprehension tests: developing these tests is time-consuming and, thus, it makes sense to develop tests that have a better chance of being implemented. First, we chose proofs that are presented (or could be presented) in courses taken by a large population of undergraduate students. The three theorems above commonly appear in what is known as transition-to-proof courses in the United States: mathematics courses designed to introduce undergraduate students to the notion of mathematical proof, preparing them to make the transition from computational mathematics courses such as calculus, to more theoretical courses such as real analysis and abstract algebra. This course is commonly required in mathematics undergraduate programs, and in proof-based mathematics courses at the undergraduate level.

Second, we wanted to use proofs that would be interesting to instructors and researchers. Theorem 1 is a central theorem in transition-to-proof courses, being one of the first indirect proofs that students encounter. The proof is also often misunderstood by students, as there is

frequently confusion as to whether or not the constant generated in this proof is itself a prime number (e.g. Hazzan & Zazkis, 2003). The proof of Theorem 2 proceeds by induction. Inductive proofs are both a central concept in transition-to-proof courses and a concept that is notoriously difficult for students (e.g. Dubinsky, 1986, 1989; Harel, 2001). Theorem 3 is a more advanced theorem, the proof of which employs one of the more sophisticated methods covered in transition-to-proof courses.

In this section, we describe the method by which we developed these three tests, and illustrate this method by following the development of items for the comprehension test of the following proof of Theorem 1:¹

Proof: Suppose the set of primes is finite. Let $p_1, p_2, p_3, \dots, p_k$ be all those primes with $p_1 < p_2 < \dots < p_k$. Let n be one more than the product of all of them. That is, $n = (p_1 p_2 p_3 \cdots p_k) + 1$. Then n is a natural number greater than 1, so n has a prime divisor q . Since q is prime, $q > 1$. Since q is prime and $p_1, p_2, p_3, \dots, p_k$ are *all* the primes, q is one of the p_i in the list. Thus, q divides the product $p_1 p_2 p_3 \cdots p_k$. Since q divides n , q divides the difference $n - (p_1 p_2 p_3 \cdots p_k)$. But this difference is 1, so $q = 1$. From the contradiction $q > 1$ and $q = 1$, we conclude that the assumption that the set of primes is finite is false. Therefore, the set of primes is infinite.

The method consisted of six main phases aimed at ultimately developing and validating short, multiple-choice tests that could be easily administered and graded. Our rationale was simple: instructors and researchers will more likely use these tests (and gain the benefits of doing so) if they have confidence that the tests measure what they intend to measure, and if they can be administered and graded in a short period of time. The purpose of the first two phases was to generate the questions and choices of items of a long multiple-choice test. In the third and fourth phases we identified and improved items from the long multiple-choice test that were mathematically inaccurate, ambiguous, or items prompting student responses that were not indicative of their understanding of the proof. The main purpose of the fifth phase was to generate a shorter test (one that could be administered in less than 20 minutes) with items that

¹ To obtain access to the complete tests, please contact the first author of the paper.

² In this section, OE# stands for Open Ended, A# for Answers, and MC# for Multiple Choice.

³ A total of 201 students took the proof comprehension test for Theorem 1, 192 students took the test for

spanned the same range of difficulties and correlate highly with students' performance on the long-version of the tests. Finally, the purpose of the sixth and last phase was to validate that student responses on the short, multiple-choice test accurately reflected their understanding of the proof. Table 1 summarizes the purpose and actions taken at each one of these phases.

Phase	Action	Purpose
1. Generating open-ended items	For each facet of our proof comprehension assessment model, we generated up to four open-ended questions.	To generate the questions for the items of our long multiple-choice tests.
2. Conducting pre-test interviews	We interviewed 12 students, asking them to answer each item generated in stage 1.	To generate the choices (correct answers and foils) for our long multiple-choice tests.
3. Reviewing items	We asked other math educators and mathematicians to review the long multiple-choice tests generated after stage 2.	To improve items, especially items that were mathematically inaccurate or ambiguous.
4. Conducting pilot interviews	We interviewed 12 students, asking them to think aloud as they answered the long multiple-choice tests refined in stage 3.	To identify and improve items in which students' responses were not indicative of their understanding of the proof.
5. Administering the test to a large population	We gave the long multiple-choice tests to approximately 200 students.	To verify that these tests had high internal reliability, to identify problematic items with poor

		discriminatory power, and to identify items that can be removed to generate shorter multiple-choice tests.
6. Conducting validating interviews	We interviewed 12 students, asking them to think aloud as they answered the shorter multiple-choice tests generated after stage 5.	To verify that the final, shorter multiple-choice tests accurately measured students' understanding.

Table 1. Actions taken and purposes of each phase.

3.1 Generating open-ended items

Our process to develop each proof comprehension test began by generating open-ended versions of *each* type of assessment item in our model. The main objective of this first step was to collect data that would help us generate multiple-choice questions with foils that represented authentic ways in which students interpret and answer such items. In general, we found that the literature was helpful in providing models for the ways in which students understand certain mathematical concepts, models that are relevant for the generation of authentic foils for items concerning the meaning of terms and statements (e.g., research on student understanding of what is a prime number or an infinite set is useful in the design of this type of item for the proof comprehension test of Theorem 1). However, the literature was less helpful when it came to the generation of foils for other types of items, such as those concerning the justification of specific claims in a given proof, or the transfer of the general methods of that proof to another context. Student answers to open-ended versions of these types of items provided us with a good source of foils for the corresponding multiple-choice questions.

An important step in the proof of Theorem 1 is the statement that n is a natural number greater than 1, and thus n has a prime divisor q . Aiming at generating multiple-choice questions assessing students' understanding of this step, we generated the following three open-ended items at this stage:

OE1:² Why is it valid to conclude that n is a natural number?

OE2: Why is it valid to conclude that n is greater than 1?

OE3: Why does n have to have a prime divisor?

In our model, these items involve the justification of three specific claims in this important step of the proof of Theorem 1.

3.2 Conducting pre-test interviews and generating items for the long multiple-choice test

Once we generated open-ended versions of each type of assessment item in our model, we recruited 12 mathematics undergraduate students who had recently completed a transition-to-proof course to participate in pre-test interviews. These participants were interviewed individually and asked to read a proof and think-aloud as they answered the corresponding open-ended items. This process was repeated until the participant had read all three proofs and answered all open-ended questions.

Interviews were audio recorded and analyzed to identify (1) different ways in which participants had interpreted the questions; (2) common answers to each item; and (3) the different ways in which participants justified those answers. The result of this analysis was the generation of possible answers that captured the main ways in which participants had responded to each item. At this point, depending on those possible answers, we created two types of multiple-choice items: items asking test takers to select *all* answers that apply from the given list of options, and items asking them to select *the best option* from a list of possible

² In this section, OE# stands for Open Ended, A# for Answers, and MC# for Multiple Choice.

answers. Finally, we wrote the choices for any given item in ways that resulted in options of similar length and structure.

For example, the following are four answers that capture what we thought were the main ways in which participants in the pre-test interviews responded to the open-ended item OE1 above:

A1: Since the natural numbers are closed under addition and multiplication, and all prime numbers and 1 are natural numbers, n is a natural number.

A2: Since n is a number greater than 0 (also expressed as n being a positive number), n is a natural number.

A3: Since all prime numbers and 1 are integers (or whole numbers), the product of primes plus 1 is an integer (or a whole number), and thus n is a natural number.

A4: Since it is a given in the proof that n is a natural number (or defined to be a natural number), we know that n is a natural number (the claim of the statement is that n is greater than 1).

Based on these four possible answers, we created the following multiple-choice item:

MC1. In the proof, why is it valid to conclude that n is a natural number? Please select **the best** option.

(a) Because the product and sum of natural numbers is a natural number.

(b) Because n is greater than 0.

(c) Because $1, p_1, p_2, \dots, p_k$ are all integers.

(d) Because it is a given in the proof that n is a natural number.

While (b) and (c) are true statements that could be used as parts of a valid justification to conclude that n is a natural number, it is clear that (a) provides *the best* explanation out of the four available options. We did not always obtain four different responses from the pre-test

interview data. In these cases, when possible, we generated foils based on existing research about the ways in which students understood some of the concepts or inferences in these tasks. If the existing research was not sufficiently robust to do this, we generated multiple-choice items with only three options. Finally, as a heuristic, we aimed at creating foils that we thought at least 5% of students in a transition-to-proof course would choose, even though at that point we did not have data on how popular each choice would be.

3.3 Reviewing items for the long multiple-choice test

Once we had a first version of all multiple-choice items, we sent them to mathematicians and mathematics educators (specifically, members of the advisory board for the grant that supported this project) asking them to check the items for accuracy and appropriateness. The feedback these colleagues provided helped us improve the items in several ways. First, colleagues were able to point out mathematical inaccuracies in our items, sometimes highlighting interpretations in which our intended foils would be correct answers. For instance, advisory board members pointed out that we needed to be careful with items involving whether or not zero is an element of the set of natural numbers. While in the context of transition-to-proof courses in the US it is common to exclude zero from the set of natural numbers, this is not the case in other contexts (e.g. set theory, logic, computer science). Thus, a foil that rests on an assumed incorrect characterization of zero as a natural number could be problematic. For instance, while A1 above would probably still be the best possible response to OE1, the ‘whole numbers’ version of A3 is correct if we assume that the set of whole numbers is equal to the set of natural numbers (i.e. if we assume that zero is a natural number).

Second, these colleagues made suggestions that would improve the choices in an item so that they would more accurately target the assessment of a particular type of student

understanding. For instance, in order to assess the comprehension of the concept of finite sets, we created the following multiple-choice item:

MC2. Which of the following are examples of finite sets? Please select **all** that apply.

- (a) The set with the following elements: 1, 2, 3.
- (b) The set of real numbers between -2 and 2.
- (c) The set of all fractions $\frac{1}{r}$ where r is a natural number.
- (d) The set of integers greater than 4 and smaller than 10.

The four choices in this item came from pre-test interview data; these answers are taken almost verbatim from the participants' responses to the question: "please give an example of a finite set and explain why it is finite." One advisory board member suggested that the sets in choices (a) and (d) were very similar, in the sense that students could (almost) instantly inspect all of their elements. The suggestion was to change option (d) so that it presented a larger finite set that did not allow students to immediately list all of its elements and thus required a different strategy to determine whether or not the set was finite. Accordingly, we changed option (d) for: "The set of integers greater than 4.5 and less than 9999." At the end of this process we ended up with 20 to 24 multiple-choice items, depending on the specific test.

3.4 Conducting pilot interviews and finalizing the long multiple-choice test

Next, we piloted these long multiple-choice tests with another 12 mathematics undergraduate students who had recently completed a transition-to-proof course. The purpose of the pilot interviews was to make sure that participants' responses to our multiple-choice items were indicative of their understanding of the different aspects of the proof. We wanted to verify that participants who answered our items correctly did so because they understood that aspect of the proof, not for other reasons such as simply being able to eliminate all other item

choices. We also wanted to make sure that students who answered incorrectly did so because they did not understand the aspect of the proof being assessed, and not for other reasons such as inaccurately interpreting the intent of the question or the meaning of one of their choices.

Pilot interviews followed a protocol similar to that of pre-test interviews: participants were interviewed individually, and they were asked to read a proof and think-aloud as they answered the corresponding multiple-choice items. This process was repeated until the participant had read all three proofs and answered all multiple-choice questions. Interviews were audio recorded and analyzed to identify (1) instances in which a participant selected the correct answer, but provided a mathematically incorrect justification for such selection, and (2) instances in which a participant who discussed what a correct answer would be, ended up selecting an incorrect answer. Having flagged all of these instances, two of the authors met to discuss the extent to which the item itself could have been at fault and sought ways to improve it.

We present an example to illustrate how we improved the test based on participants' behavior with respect to the following item:

MC3. In the proof, we define $n = p_1 p_2 \dots p_k + 1$. Would the proof still work if we instead defined $n = p_1 p_2 \dots p_k - 1$? Why or why not? Please select **the best** option.

- (a) Yes, because $p_1 p_2 \dots p_k + 1$ and $p_1 p_2 \dots p_k - 1$ are either both even or both odd natural numbers, so this will not affect the rest of the proof.
- (b) Yes, because $n = p_1 p_2 \dots p_k - 1$ is still a natural number greater than 1 and in this case $p_1 p_2 \dots p_k - n = 1$, which means q must still be 1.
- (c) No, because if $n = p_1 p_2 \dots p_k - 1$ one cannot be sure that n is a new prime number, which is the idea of the original construction.

- (d) No, because the difference between n and the product of all primes will be -1 , which will not work because we can't have a negative prime number.

This was a difficult item requiring test takers to determine if the general method of the proof would work under slightly different conditions. In this case, the general method does work and (b) correctly specifies why. However, when considering the alternative definition of n as $p_1p_2\ldots p_k - 1$, one must additionally assume that the set of prime numbers contains at least two elements in order to be able to reach the corresponding conclusion that $p_1p_2\ldots p_k - 1$ is necessarily greater than 1. Yet this additional assumption was not stated explicitly in the item above. When piloting this item we observed participants struggling with the possibility that if 2 were the only prime number, one would not be able to conclude that $p_1p_2\ldots p_k - 1$ was greater than 1, which led them to believe that maybe this new definition of n would not work. To address this issue, we simply modified the item statement to explicitly include this assumption. The item then read:

In the proof, we define $n = p_1p_2\ldots p_k + 1$. Would the proof still work if we instead defined $n = p_1p_2\ldots p_k - 1$, knowing that there is at least two prime numbers? Why or why not? Please select **the best** option.

3.5 Administering the long multiple-choice test and generating the short multiple-choice test

After finalizing all items of the multiple-choice tests, we administered these tests to approximately 200 undergraduates taking a transition-to-proof course. The instructors of these courses allocated 40 minutes of class time to distribute each of the three tests. On the day each test was distributed, students in the course received a packet that contained the theorem and its proof, instructions on the different types of items in the test, and all the multiple-choice questions (the order of the items was randomized). The test for the proof of Theorem 1 was distributed after instructors had introduced proofs by contradiction in class

(approximately a third of the way into the term), the test for Theorem 2 was distributed once instructors had discussed the principle of mathematical induction (usually by the middle of the term), and the test for Theorem 3 was distributed by the end of term, after instructors had discussed the notion of the cardinality of sets.³

Each test that the student completed had 20 or 21 items and students had 40 minutes to complete the test. We strove to create shorter tests (12 items) that had the following features: (1) an acceptable level of internal consistency (Cronbach's $\alpha > .7$); (2) included items that spanned the same range of difficulties of those of the long test and correlated highly with students' performance on the long-version of the tests; and (3) contained a wide range of types of items in our model. The initial Cronbach's alpha coefficient for each one of the three long tests was greater than 0.7. For each test, we calculated the corrected item total biserial correlations (i.e. the correlation between the item score and the overall test score after removing that item) and started removing the least discriminating items (initially all those with a correlation smaller than 0.2), aiming to end up with a set of 12 items that contained roughly half local and half holistic types of items according to our model (Table 2 illustrates this process for the test for Theorem 1). The short, 12-item tests all maintained an acceptable level of internal consistency (Cronbach's $\alpha > .7$).

3.6 Conducting validating interviews

Finally, we recruited another 12 mathematics undergraduate students who had recently completed a transition-to-proof course, and we interviewed them individually to validate that their responses on the short, 12-item tests accurately reflected their understanding of the proofs. In these interviews, participants were first asked to read one of the proofs and to

³ A total of 201 students took the proof comprehension test for Theorem 1, 192 students took the test for Theorem 2, and 152 students took the test for Theorem 3. The decreasing number of participating students was not only due to the regular reduction of class size as the term progressed: one of the participating instructors did not reach the topic of cardinality in class, which meant that we could not distribute the test for Theorem 3 in the two sections led by this instructor.

summarize in their own words the steps and main ideas of that proof. We then asked participants to take the short test for that proof and to then justify their answers to each one of the 12 items. Finally, participants were asked if there was anything about the proof that they had not understood. This process was then repeated with the other two proofs.

Item type	Item number	Test length		
		20	15	12
Local	1	0.27	0.28	
	2	0.30	0.33	0.32
	3	0.33	0.34	0.35
	4	0.26	0.28	0.28
	5	0.19		
	6	0.09		
	7	0.23	0.26	
	8	0.32	0.32	0.32
	9	0.40	0.39	0.38
	10	0.35	0.36	0.33
	11	0.35	0.38	0.38
Holistic	12	0.26	0.24	
	13	-0.07		
	14	0.37	0.35	0.35
	15	0.47	0.44	0.45
	16	0.24		
	17	0.30	0.31	0.29
	18	0.39	0.37	0.37
	19	0.33	0.38	0.37
	20	0.19		
Reliability		0.71	0.74	0.72

Table 2. Corrected item-total biserial correlation and reliability for test for Theorem 1 with different lengths (the final version of items MC1, MC2, and MC3 that we presented earlier are numbered 8, 3, and 14 respectively).

Validating interviews were audio recorded and analyzed to characterize participants' understanding of the steps and main ideas of the proof, as evidenced by their summaries, their justifications to their answers to each one of the 12 items in the short test, and what they said they had not understood about the proof. This was then compared to their actual choices for the items of the test and their overall score. Participants' performance on the short tests was reflective of their understanding of the proofs (as evidenced by their answers to interview questions) and did not warrant any further modifications to the tests items.

4. Discussion

There are several reasons why the development of the type of proof comprehension assessments described in this article is important. First, some instructors have claimed that they do not have a good sense for how well students understand the proofs that they present (e.g. Cowen, 1991; Weber, 2012). For mathematics instructors, these tests can provide formative feedback on their presentations. An instructor could have their students take the short test after he or she has presented the proof of that test, or simply have students answer specific test items using clickers or mobile devices in the classroom. Students' performance on these tests could alert instructors to areas of the proof that their students did not understand, as well as provide more general feedback to the overall quality of their presentation.

Second, for students, these assessment items can direct their attention toward important aspects of the proof that they may not have considered. Previous research has found that mathematics majors' judgments of how well they understood a particular proof are often inaccurate (Selden & Selden, 2003) and that students tend to believe they understand a proof completely if they can specify how each new statement in a proof follows from previous assertions (Weber & Mejía-Ramos, 2014). If these mathematics majors knew they would be

asked to apply the ideas of the proof to specific example objects or to solve a problem in another context, they would be encouraged to understand the proof at a deeper level.

Third, we noted that Yang and Lin's (2008) model for comprehending geometry proofs opened up interesting areas of research and we believe that assessments for proofs in undergraduate mathematics have the potential to do the same. Such an assessment can be used, for instance, to obtain baseline data about what aspects of proof comprehension mathematics majors struggle with or to evaluate the effectiveness of instructional interventions. In particular, several mathematics educators have proposed alternative methods for presenting proofs to students that they believe will increase understanding (Alcock & Wilkinson, 2011; Hersh, 1993; Leron, 1983; Rowland, 2001), with Leron's structured proofs and Rowland's generic proofs widely cited as research-based pedagogical suggestions. However we are aware of no studies that demonstrate the efficacy of these approaches⁴. We believe that the type of tests we described in this article can play an important role in the systematic assessment of any such novel intervention designed to improve proof comprehension.⁵

We also believe these tests can be used to further investigate the existence of more general proof competencies, such as the comprehension of particular proof techniques, or proofs in a broader topic. In another article, we report and discuss the strong correlations between students' performance on any two of the three proof comprehension tests we developed. Taken together with the high internal consistency of these tests, these results suggest that the comprehension of the simple proofs that are presented in a transition-to-proof

⁴ Malek and Movshovitz-Hadar (2011) found interesting suggestive evidence that Rowland's (2001) generic proofs might improve comprehension, but their study was a small-scale qualitative study with only three or four students reading the generic proofs that were provided.

⁵ Indeed, Fuller et al. (2014) and Weber et al. (2012) reported being unable to document any comprehension benefits of using structured and generic proofs (respectively), as measured by tests developed using Mejia et al.'s framework (2012).

course could be a meaningful single-dimensional construct (Mejia-Ramos et al., under review).

We argue that regardless of whether proof comprehension tests are used in the classroom or in the research lab, it would be greatly beneficial to instructors and researchers to have access to validated, short, multiple-choice tests. Indeed, instructors and researchers will be more likely to use these tests (and gain the benefits of doing so) if they have confidence that the tests measure what they intend to measure, and if they can be administered and graded in a short period of time. It takes undergraduate mathematics students under 20 minutes to complete the short version of any one of the tests we developed, and instructors can grade each test in just a couple of minutes. Furthermore, grading could be done automatically if the tests were administered through a computer.

In this article, we discussed the multi-stage process by which we developed and validated three of these types of tests. While these methods have been employed widely in the field of educational measurement and have started to be used for the development of instruments in undergraduate mathematics education (e.g. Carlson, Oehrtman, & Engelke, 2010; Epstein, 2013) we have demonstrated that our implementation of this methodology is capable of producing short, reliable tests that validly measure students' comprehension of the proofs that they read, something that until now had not been assessed in this manner.

Our hope is that other mathematics educators will employ this method to develop and validate their own proof comprehension tests, so that instructors and researchers will eventually have access to a large battery of these tests to use in their classrooms and their own research studies.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. DUE-1245626.

References

- Alcock, L.J., Hodds, M., Roy, S., & Inglis, M. (2015). [Investigating and improving undergraduate proof comprehension](#), *Notices of the American Mathematical Society*, 62, pp.742-752.
- Alcock, L. & Wilkinson, H. (2011). e-Proofs: Design of a resource to support proof comprehension in mathematics. *Educational Designer*, 1(4). Retrieved from: <http://www.educationaldesigner.org/ed/volume1/issue4/article14/>
- Alcock, L., & Weber, K. (2005). Proof validation in real analysis: Inferring and checking warrants. *The Journal of Mathematical Behavior*, 24, 125–134.
- Carlson, M., Oehrtman, M., & Engelke, N. (2010). The precalculus concept assessment: A tool for assessing students' reasoning abilities and understandings. *Cognition and Instruction*, 28(2), 113-145.
- Conradie, J., & Frith, J. (2000). Comprehension tests in mathematics. *Educational Studies in Mathematics*, 42, 225–235.
- Cowen, C. (1991). Teaching and testing mathematics reading. *The American Mathematical Monthly*, 98(1), 50-53.
- de Villiers, M. (1990). The role and function of proof in mathematics. *Pythagoras*, 24, 17–24.
- Deeds, D., and Callen, B. (Eds.). (2006). *Proceedings of the national STEM assessment conference*. Jointly sponsored by the National Science Foundation and Drury University. Springfield, MO: Drury University.
- Dreyfus, T. (1990). Advanced mathematical thinking. In P. Nesher & J. Kilpatrick (Eds.) *Mathematics and cognition: A research synthesis by the International Group for the Psychology of Mathematics Education*. Great Britain: Cambridge University Press.

- Dreyfus, T. (1991). Advanced mathematical thinking processes. In D. Tall (Ed.) *Advanced Mathematical Thinking* (pp. 215-230). Kluwer: The Netherlands.
- Dubinsky, E. (1986). On teaching mathematical induction I. *The Journal of Mathematical Behavior*, 5, 305-317.
- Dubinsky, E. (1989). On teaching mathematical induction II. *The Journal of Mathematical Behavior*, 8, 285-304.
- Epstein, J. (2013). The Calculus Concept Inventory— Measurement of the effect of teaching methodology in mathematics. *Notices of the AMS*, 60(8), 1018-1026.
- Fuller, E., Weber, K., Mejia-Ramos, J. P., Samkoff, A., & Rhoads, K. (2014). Comprehending structured proofs. *International Journal for Studies in Mathematics Education* 7(1), 1-32.
- Hanna, G. (1990). Some pedagogical aspects of proof. *Interchange*, 21(1), 6–13.
- Hanna, G. (1995). Challenges to the importance of proof. *For the Learning of Mathematics*, 15(3), 42-50.
- Harel, G. (2001). The development of mathematical induction as a proof scheme: A model for DNR-based instruction. In S. Campbell & R. Zazkis (Eds.), *Learning and teaching number theory* (pp.185–212). Norwood, New Jersey: Ablex.
- Hazzan., O. & Zazkis., R. (2003). Mimicry of proofs with computers: The case of linear algebra. *International Journal of Mathematical Education in Science and Technology* 34(3), 385-402.
- Healy, L., & Hoyles, C. (2000). A study of proof conceptions in algebra. *Journal for Research in Mathematics Education*, 31, 396–428.
- Hersh, R. (1993). Proving is convincing and explaining. *Educational Studies in Mathematics*, 24(4), 389-399.

- Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force Concept Inventory. *The Physics Teacher*, 30, 141-158.
- Hodds, M., Alcock, L., & Inglis, M. (2014) [Self-explanation training improves proof comprehension](#), *Journal for Research in Mathematics Education*, 45, pp.98-137
- Inglis, M., & Alcock, L. (2012). Expert and novice approaches to reading mathematical proofs. *Journal for Research in Mathematics Education*, 43, 358-390.
- Knuth, E. (2000). The rebirth of proof in school mathematics in the United States. *International Newsletter on the Teaching and Learning of Mathematical Proof*. (Spring 2000). Available for download at:
<http://www.lettredelapreuve.it/OldPreuve/Newsletter/000506Theme/000506ThemeUK.html>. Last downloaded: May 29, 2012.
- Knuth, E. (2002), Secondary school mathematics teachers conceptions of proof. *Journal for Research in Mathematics Education*, 33(5), 379-405.
- Konior, J. (1993). Research into the construction of mathematical texts. *Educational Studies in Mathematics*, 24(3), 251-256
- 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.
- Leron, U. (1983). Structuring mathematical proofs. *The American Mathematical Monthly*, 90(3), 174-184.
- Lin, F.-L., & Yang, K.-L. (2007). The reading comprehension of geometric proofs: The contribution of knowledge and reasoning. *International Journal of Science and Mathematics Education*, 5, 729-754
- Malek, A. & Movshovitz-Hadar, N. (2011). The effect of using Transparent Pseudo-Proofs in linear algebra. *Research in Mathematics Education*, 13, 33-58.

- Mamona-Downs, J., & Downs, M. (2005). The identity of problem solving. *The Journal of Mathematical Behavior*, 24, 385–401.
- Mantini, L, Mills, M., & Johnson, J. (2012). Reading Comprehension of Series Convergence Proofs in Calculus II. In S. Brown, S. Larsen, K. Marrongelle, and M. Oehrtman (Eds.) *Proceedings of the 15th Annual Conference for Research in Undergraduate Mathematics Education* (pp. 485-489). Available for download at:
http://sigmaa.maa.org/rume/crume2012/RUME_Home/RUME_Conference_Papers_files/RUME_XV_Conference_Papers.pdf
- Martin, W. G., & Harel, G. (1989). Proof frames of preservice elementary teachers. *Journal for Research in Mathematics Education*, 20(1), 41-51.
- 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(1), 3-18.
- Mejía-Ramos, J. P., & Inglis, M. (2009). Argumentative and proving activities in mathematics education research. In F.-L. Lin, F.-J. Hsieh, G. Hanna, & M. de Villiers (Eds.), *Proceedings of the ICMI Study 19 conference: Proof and Proving in Mathematics Education* (Vol. 2, pp. 88–93). Taipei, Taiwan: National Taiwan Normal University.
- Mejía-Ramos, J. P., Weber, K., de la Torre, J., & Lew, K. (under review). Student performance on proof comprehension tests in transition-to-proof courses.
- Mills, M. (2011). Mathematicians' pedagogical thoughts and practices in proof presentation. *Presentation at the 14th Conference for Research in Undergraduate Mathematics Education*. Available for download from:
http://sigmaa.maa.org/rume/crume2011/Preliminary_Reports.html

- Powers, R., Craviotto, C., & Grassl, R. (2010). Impact of proof validation on proof writing in abstract algebra. *International Journal of Mathematical Education in Science and Technology*, 41, 501–514.
- Raman, M. (2004). Epistemological messages conveyed by three high-school and college mathematics textbooks. *The Journal of Mathematical Behavior*, 23(4), 389-404.
- Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37–75). Boston: Kluwer.
- Rowland, T. (2001). Generic proofs in number theory. In S. Campbell & R. Zazkis (Eds.), *Learning and teaching number theory: Research in cognition and instruction* (pp. 157–184). Westport: Ablex.
- Schoenfeld, A. H. (1994). What do we know about mathematics curricula? *The Journal of Mathematical Behavior*, 13(1), 55-80.
- Segal, J. (1999). Learning about mathematical proof: Conviction and validity. *The Journal of Mathematical Behavior*, 18(2), 191-210.
- 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. (2001). Student difficulties in constructing proofs: The need for strategic knowledge. *Educational Studies in Mathematics*, 48(1), 101-119.
- Weber, K. (2004). Traditional instruction in advanced mathematics courses: A case study of one professor's lectures and proofs in an introductory real analysis course. *The Journal of Mathematical Behavior*, 23(2), 115-133.

- Weber, K. (2010a). Proofs that develop insight. *For the Learning of Mathematics*, 30(1), 32-36.
- Weber, K. (2010b). Mathematics majors' perceptions of conviction, validity, and proof. *Mathematical Thinking and Learning*, 12, 306-336.
- Weber, K. (2012). Mathematicians' perspectives on their pedagogical practice with respect to proof. *International Journal of Mathematical Education in Science and Technology*, 43(4), 463-482.
- Weber, K., & Mejía-Ramos, J. P. (2011). Why and how mathematicians read proofs. *Educational Studies in Mathematics*, 76, 329–344.
- Weber, K. & Mejía-Ramos, J. P. (2014). Mathematics majors' beliefs about proof reading. *International Journal of Mathematical Education in Science and Technology*, 45, 89-103.
- Weber, K., Fuller, E., Mejia-Ramos, J. P., Lew, K., Benjamin, P., & Samkoff, A. (2012, February). Do generic proofs improve proof comprehension? Presented at the *15th Annual Conference on Research in Undergraduate Mathematics Education*. Portland, Oregon.
- Yang, K.-L. (2012). Structures of cognitive and metacognitive reading strategy use for reading comprehension of geometry proof. *Educational Studies in Mathematics*, 80(3), 307-326.
- Yang, K.-L., & Lin, F.-L. (2008). A model of reading comprehension of geometry proof. *Educational Studies in Mathematics*, 67, 59–76.