Improving calculus explanations through peer review

Daniel Lee Reinholz

Department of Mathematics & Statistics, San Diego State University. 5500 Campanile Drive, San Diego, CA 92182

Abstract

This paper describes Peer-Assisted Reflection (PAR), a peer-review activity designed to help students explain mathematics. PAR was implemented in a single experimental section during two consecutive semesters (phases) of introductory college calculus. During the second semester (Phase II), students were explicitly taught how to provide better feedback to each other. As a result, the amount and quality of feedback provided by Phase II students was significantly improved from Phase I. During both phases, students who used PAR had significantly better explanations than a comparison group that did not include PAR, indicating that student explanations can be improved with relatively little intervention. To capture these improvements, I introduce a new analytic scheme that defines explanation as a cluster concept along four dimensions. This context-neutral scheme operationalizes explanations in a way that growth can be captured longitudinally.

Keywords: Explanation, Reflection, Peer Assessment, Calculus

1. Introduction

Mathematical understanding is multi-faceted. To engage proficiently with mathematics, students must: master concepts and procedures, engage strategically in problem solving, explain and reflect on their work, and develop productive dispositions towards knowing and doing mathematics (NRC, 2001). In particular, the sociocultural turn in mathematics describes learning as a process of being enculturated in appropriate social practices (Engle, 2012; Lave, 1996). Accordingly, the present study focuses on Peer-Assisted Reflection (PAR), an activity designed to support the development of a particular practice, explanation.

Each week, students completed a draft solution to a PAR homework problem and engaged in a structured peer review process. PAR created space for students to reflect on their explanations, exchange feedback with their peers, and incorporate their learning into improved explanations through revision. This paper focuses on the impact of PAR on student explanations, extending prior work that demonstrated PAR's impact on student success (Reinholz, 2015b).

PAR was developed over a multiple phase design experiment (Cobb et al., 2003; Collins et al., 2004). Two of these phases took place in calculus, a conceptually rich (Oehrtman et al., 2008; Tall, 1992) and challenging (Bressoud et al., 2013) area of mathematical inquiry. These phases are the focus of the present

Email address: daniel.reinholz@sdsu.edu (Daniel Lee Reinholz)

work. To capture changes in student explanations over the course of a semester, I introduce a new analytic scheme. This scheme provides a lens for quantifying changes in the quality of student explanations, designed to work across mathematical content domains. Additionally, this paper discusses the impact on the types of feedback students provided to each other as a result of being taught how to provide better feedback during Phase II. This paper aims to make two primary contributions to the literature: (1) it documents the impact of a new instructional technique on improving student explanations, and (2) it provides a content-neutral analytic scheme for capturing changes in the quality of explanations.

2. Theoretical Framing

2.1. Defining Explanation

Explanation plays a central role in a variety of disciplines, including: science (Braaten & Windschitl, 2011), mathematics (CCSSM, 2010), and psychology (Lombrozo, 2006). Explanations, as defined by philosophers of science, generally focus on describing *how* or *why* something happens (Braaten & Windschitl, 2011). In mathematics, explanation is considered an important aspect of proof (De Villiers, 2003; Harel & Sowder, 2007; Steiner, 1978), and valued in its own right. Psychological studies (Lombrozo, 2006) highlight the cognitive function of explanations in generalizing understandings, which helps promote learning (Chi et al., 1994; Wong et al., 2002).

Given its centrality as a mathematical practice and as a tool for learning, standards documents from across the world elevate explanation. In the US, the *Principles and Standards for School Mathematics* include communication as one of five mathematical process standards (NCTM, 2000). Similarly, the Danish KOM project dedicates the *communication competency* to explanation (Niss, 2011), and explanations are central to the reasoning strand in the Australian Curriculum (ACARA, 2009). The US *Common Core State Standards for Mathematics* even prize explanation as a "hallmark" of understanding (CCSSM, 2010).

Although there is consensus on the value of explanation, defining explanation is a challenge. For instance, science educators continue to debate the relationship between explanation and argumentation (Osborne & Patterson, 2011). Similarly, philosophers of science have yet to agree on a common definition of explanation (Wilson & Keil, 1998). Thus, while there is general consensus that explanation is valuable, there is less agreement on exactly what explanation is. I argue that proof offers a way forward for mathematics educators. Given the centrality of proof to mathematics, defining proof has received considerable attention (Weber, 2014). The progress the field has made in defining proof is instructive when it comes to explanation.

To define proof, educators have attempted to create a set of desirable criteria that all proofs should meet. However, there is no single shared understanding of what constitutes a proof in the mathematical community. While some proofs would unambiguously be classified as proofs by nearly all professional mathematicians (e.g., rigorous deductive arguments), there are many *proofs* where classification is up for debate (e.g., graphical or computation proofs). As such, it has been nearly impossible to distill an "essence of proof" that can be embodied into a set of criteria for classification purposes.

To overcome this difficulty, proof can be understood as a cluster concept (Lakoff, 1987; Weber, 2014). To define a cluster concept, one begins with a set of criteria related to proofs. When a particular proof meets all of the criteria, it is unambiguously classified as a proof. When a proof satisfies only some of the criteria, it is a borderline case that may be considered a proof by some and not by others. This accounts for the ambiguity with socially-defined concepts, but still allows for the community to move forward by identifying a set of desired characteristics for proofs. Not intended to be exhaustive, a subset of criteria one might consider includes: (1) proof as a "convincing argument, as judged by qualified judges" (Hersh, 1993), (2) proof as a social process, meaning proofs must be sanctioned by the mathematical community (Harel & Sowder, 2007; Mason et al., 2010), and (3) proof as an explanation of why something is true (Hanna, 1990).

In analogy, explanations can be defined as a cluster concept. The first feature of explanations is that they must explain something, which is called the 'explanandum' (Lombrozo, 2006). In a proof, a *key idea* (Raman, 2003) contains the essential elements required to construct the proof, but falls short of representing a proof per se. Key ideas for explanations are analogously defined as describing some essential component of the mathematics in a problem, but are incomplete in themselves. In this sense, a key idea provides the raw material from which a complete explanation can be constructed. In explaining mathematics, an individual must express their private understanding of this key idea so that it can be conveyed to others publicly. Because communication is social, this process is context-dependent.

Although explanations must be judged in context, they still must conform to the larger body of knowledge established in the field of mathematics, related to the *accuracy* of an explanation. Accuracy concerns whether or not an explanation is in some sense "right," regardless of how illuminating or easy to follow it is. In addition to being accurate, mathematical explanations must be *precise*, using language that is exact. This can be achieved by correctly using the standard lexicon of mathematics (Jones, 2000), the appropriate use of colloquial language, or even in inventing new definitions. Perhaps the most fundamental aspect of an explanation is that it should describe *how* or *why* something happens (Braaten & Windschitl, 2011). In mathematics, proofs are considered to play two distinct roles; they both *prove* and *explain* (De Villiers, 2003; Steiner, 1978). A proof that *proves* demonstrates *that* a statement is true (related to *accuracy*); a proof that *explains* also makes clear *why* a statement is true (Hanna, 1990). This idea of describing why is denoted as *justification*. Finally, visual representations are a useful way to develop and express mathematical thinking (Cuoco et al., 1996; NCTM, 2000). Visual representations, or *diagrams*, often embody understandings that may be difficult to express in words.

Taken together, the quality of an explanation can be measured along four dimensions: accuracy, precision, justification, and diagrams. As a cluster concept, would one generally expect that explanations of higher quality would be scored higher along each dimension, although there may be some ambiguous cases in which a high-quality explanation scores low on some of the dimensions. This scheme is not intended to be exhaustive, but to provide a first attempt at classifying explanations in a meaningful way. For instance, once could choose to extend the scheme by adding additional dimensions.

2.2. Explanation and Understanding

The sociocultural turn in education describes learning as a process of becoming a more central member of a community of practice (Lave, 1996). This contrasts with constructivist approaches to learning, which emphasize the acquisition of concepts, or development of understanding (Smith et al., 1993). These two metaphors for learning are both valuable; neither alone is sufficient (Sfard, 1998).

Historically, research has focused on explanation from a constructivist standpoint, rooted in cognitive psychology (Lombrozo, 2006). This literature emphasizes that explanation is not just a hallmark of understanding, but that the very act of explaining helps generate understanding. For instance, students who are prompted to self-explain while reading a text have a significantly greater understanding than those who instead study the text an additional time (Chi et al., 1994). Even though students do not always generate correct explanations, embodying their partial understandings into an explanation helps consolidate learning.

Recently, self-explanation has been used to improve students' proof comprehension (Hodds et al., 2014). Through three studies, undergraduate mathematics students in the UK were taught to self-explain while reading proofs. The researchers found that those who received instruction on self-explanation had a significantly greater level of proof comprehension; the results were replicated in laboratory and classroom environments.

Another well-known example of using explanation to promote learning is Peer Instruction (Crouch & Mazur, 2001). In Peer Instruction, classes are taught in a series of small segments consisting of: (1) a short presentation on a concept, (2) a conceptual question, (3) individual student responses to the questions, (4) peer discussions in which students explain the ideas and try to convince each other they are correct, (5) another poll of student responses, and (6) a short explanation of the correct response by the instructor. This technique is generally used to make large-lecture sections more interactive. A decade of study in introductory physics shows that this technique significantly improves student understanding (Crouch & Mazur, 2001).

As these studies highlight, explanation is a powerful tool for promoting learning. Simultaneously, explanation is considered a "hallmark" of deep understanding (CCSSM, 2010). Explanation requires students to synthesize, process, and describe mathematical ideas in their own words. As such, students explaining mathematics is characteristic of a "powerful" classroom, because students have the opportunities to develop mathematics themselves and act as authorities (Schoenfeld, 2014). Drawing on the metaphor of learning as participation, student explanations must be seriously considered as an outcome of interest in their own right, not merely an intermediary to developing "understanding." From this perspective, explanation and understanding are inextricably linked; the fact that one can explain a concept is not taken as evidence of understanding, but rather as the understanding itself.

2.3. Peer-Assisted Reflection

Defining explanation as a cluster concept highlights the social nature of explanations. As such, students need opportunities to actually explain mathematics to learn to better explain; direct instruction on explanations is likely insufficient (Weber, 2014). To test this conjecture, Peer-Assisted Reflection (PAR) was developed. PAR has now been used in over a dozen classrooms, across STEM disciplines such as mathematics, physics, and engineering. PAR aims to improve student explanations, but was not specifically designed to align with the four-dimensional scheme presented above.

In what follows, a particular implementation of PAR in introductory college calculus is described. PAR required students to: (1) complete a draft solution to a problem, (2) reflect on their work, (3) analyze peers' work and exchange feedback, and (4) revise their own work. Students completed steps (1), (2), and (4) of the PAR cycle outside of class, while students engaged in step (3), the peer conferences, during class. For their conferences, students traded their draft work with a peer, read over each other's solutions silently and provided written feedback for five minutes, and then had an additional five minutes to discuss their feedback.

Early work on PAR focused on developing a theoretical model for learning through peer assessment (Reinholz, 2015a). Empirical studies of PAR showed that students significantly improved their success in undergraduate calculus; during two semesters of a design-based research study, students improved their success (passing with an A, B, or C in the course) by 13% during Phase I and 23% during Phase II (Reinholz, 2015b). These studies also found a significant impact on the nature of peer discussions around calculus concepts (Reinholz, 2015c). Like prior studies around explanation in learning, this work focused primarily on the connections between explanation and understanding. The present study draws on students from the same populations described in these papers.

2.4. PAR Supports for Explanation

To support student explanations, tasks were carefully selected and modified from multiple sources: the Shell Centre, the Emerging Scholars Program, *Calculus Problems for a New Century*, and existing homework problems from the introductory calculus course at the institution where the study took place. Tasks were selected to align with Schoenfeld's problem aesthetic (Schoenfeld, 1991). Ideal tasks: were accessible, had multiple solution paths, and provided opportunities for further exploration. When possible, existing tasks were modified to require students to generate additional written explanations.

To support students to reflect and exchange feedback, students answered designated prompts each week. The reflection portion of PAR required students to answer six yes-no questions related to their drafts. Three of these concerned their written explanations: (1) did you explain why, not just what?; (2) did you avoid the use of pronouns and other ambiguous language?; and (3) did you consult definitions of mathematical terms you used? When students exchanged feedback, they were asked to fill out three prompts: (1) give at least one suggestion to improve the communication of the solution, (2) note any errors you found, and (3) provide any additional suggestions (optional). The purpose of these prompts was to focus students on communication, not just conceptual understanding, in their feedback.

Beyond this explicit focus on explanation in reflections and peer feedback, asking students to explain to their peers changed the demand characteristics of the explanation tasks. In addition to explaining the math so that their teacher judged it as correct, students needed to explain the math in a way that a peer could understand. When interviewed about their experiences with PAR, many students noted the importance of explaining to a peer. For instance, Maria stated (Reinholz, 2015b),

> It's one thing if I think it looks good, but other people look at it and say it doesn't make sense to me. So [PAR] helps me figure out how to communicate better. It helps me explain things in a way that is readable to others and not just myself.

As Maria's words highlight, PAR helped her figure out how to communicate to her peers, because she had opportunities to receive feedback from them about what was and was not understood.

Finally, to maximize the usefulness of feedback, students need opportunities to actually use the feedback they receive (Butler & Winne, 1995; Sadler, 1989). These opportunities were built into the PAR cycle; students were expected to revise their draft solutions after receiving feedback. On average, students scored approximately 30% higher on the PAR problems after revising their drafts (Reinholz, 2015b).

2.4.1. Teaching Students to Explain (Phase II Only)

During Phase II, specific efforts were made to improve the quality of student feedback. At a broad level, this involved framing PAR in a way such that students recognized the value of providing constructive criticism to one another. This was achieved by suggesting students be "critical friends" and by discussing how simply saying "everything looks good" does not actually help your peers learn. Moreover, the instructor emphasized on a regular basis how PAR would help students improve their explanations, and that practicing explanations would help their learning more generally. Finally, the instructor had occasional discussions with the class about the quality of feedback that they provided to each other. For instance, the instructor might highlight a particularly useful piece of written feedback provided by a certain student, or encourage the class as a whole to write more on their feedback forms.

In addition to general framing, Phase II students practiced analyzing sample student work on a weekly basis, for approximately 10 minutes of class time. This process involved students looking individually at three samples of written work and then discussing as a class how to provide feedback to help the hypothetical students improve their explanations. The sample work was always drawn from the PAR problems students had just completed, and the discussions took place immediately after students turned in their PAR packets. The samples were typically generated based on common solutions provided by students in previous semesters. The sample solutions varied in their quality. At times the samples were used to highlight deficiencies in explanations, and at other times they were used to demonstrate for students highly effective ways of communicating mathematically. The instructor helped draw students' attention to ways to improve the sample explanations both conceptually and communicatively.

To illustrate the discussions of sample work, I highlight an example from the week 10 PAR problem (Reinholz, 2015b), focused on computing the area of one's hand. Students were given three sample solutions to the following prompt: "Explain (in principle) how you could improve your method to make your estimate as accurate as one could want (i.e. minimize the error)." The sample responses presented to students were as follows:

1. If I took a limit as the width of rectangles approaches 0 (making the number of rectangles approach ∞), the difference in the area under the curve and the rectangles would approach 0.
2. You could use midpoints rather than endpoints and it will be more accurate because there will be less overlap.
3. If I had more rectangles there would be less overlap and the approximation would be better.

Related to this prompt, the instructor led a whole class discussion. A sample related to the second prompt is given:

1. Patrick: In the lab we just did, we created that graph to show that midpoints aren't always more accurate.
2. Instructor: So midpoints aren't always the best. What else?

...

3. Sue: Wouldn't midpoints be better?
4. Instructor: What do you think [looks at class], would midpoints be better?
5. Barry: Would it even matter, because it says "as accurate as you would want," and you can only get so accurate with midpoints?

As this brief excerpt shows, students discussed the quality of the sample explanations given to them. In line 1, Patrick connects the explanation provided to work that students had done in a calculator lab about Riemann sums. In line 5, Barry notes that while midpoints may improve the accuracy, in general, they will not improve the approximation to an arbitrary accuracy. In responding to sample solutions, students often focused on mathematical concepts (as shown above) in addition to specifics about language and communication.

3. Method

3.1. Course Context

The study took place over two consecutive semesters, Phase I and Phase II, of introductory calculus at a research university. The course had approximately 400 students enrolled each semester, distributed across over 10 sections with 30-100 students each. A mix of graduate teaching assistants, post-doctoral researchers, and full-time instructors taught the sections, which met for 50 minutes, four times weekly.

The course was organized by a course coordinator, a tenured faculty member with a strong commitment to mathematics education. The coordinator provided a common syllabus, pool of homework assignments, calculator lab assignments, textbook, study guides (with learning goals), and exams for all sections. After nearly a decade in this role, the course coordinator had a well-established canon of materials; the nature of the exams and study guides changed little over the years. Both students and instructors had access to many years of prior exams, which promoted a common understanding of course expectations. Moreover, there were weekly meetings to ensure that instructors taught the same syllabus in similar ways. In general, various sections of the class tended to cover the same sections in the book on the same day, and the coordinator meetings helped ensure that particular examples and ways of thinking were discussed in all sections.

3.2. Participants and Data Collection

Each phase included a single experimental section, with all other sections used as comparisons. Students were not randomly assigned to sections, so the study was quasi-experimental. Michelle taught the Phase I experimental section ($N = 56$). Michelle was a full-time instructor with a PhD in mathematics education and nearly a decade of teaching experience. The author taught the Phase II experimental section ($N = 35$). At the time I was a graduate student with three years of teaching experience. Each semester the experimental section and three comparison sections were observed (at least six times) and class sessions were video recorded. All of the observed instructors had taught the course before, and had considerable teaching experience (there was a mathematics education postdoctoral researcher, a full-time instructor with a decade of teaching experience, and advanced graduate students).

The primary data source for this paper is student explanations on the PAR problems. PAR problems were assigned in the experimental sections during Phase I and Phase II and a randomly chosen comparison section ($N = 30$) during Phase II. In the comparison section, students completed the PAR problems as homework, but they did not exchange feedback and revise their solutions. A mathematics graduate student with prior calculus teaching experience taught this section. This graduate student instructor had more experience with calculus than the researcher who taught during Phase II, but less experience than Michelle, the full-time instructor from Phase I. By collecting student work in these three sections, I compared student explanations across Phase I, Phase II, and the comparison section. Each of these groups consistent of a unique population of undergraduate students enrolled in first-semester college calculus.

3.3. Design

The primary instructional method in all observed sections was lecture, confirmed by classroom observations. Students in experimental sections also engaged with PAR each week. The PAR homework problems had multiple parts and most required written explanations (see Appendix A-Appendix C); students completed a total of 14 problems during the semester. PAR problems were related to concepts taught in class, but the mathematics on PAR problems was never formally taught to students. In fact, PAR problems were often introduced before related concepts were discussed in class. For instance, the week 10 problem, used above to exemplify the discussion of sample student work, focused on the approximation of area, and was assigned before students had studied Riemann sums, integration, or area. During Phase II, additional instruction was added to help students learn to analyze explanations and provide meaningful feedback to each other, as described above; Phase I did not include this instruction.

3.4. Choice of Problems

Homework problems were selected for analysis. Homework problems were chosen, rather than exam problems, because students had more time to write elaborated responses outside of the time pressure of an exam. Although all sections assigned homework problems from a common pool of problems, not all sections necessarily assigned the same problems from the pool, which limited the choice of problems that could be used to compare across phases. Moreover, problems used in Phase I were sometimes revised for Phase II. For instance, as instructors noticed there were aspects of the wording on the original problems that were confusing for students, they modified them before using them next semester.

Keeping the above considerations in mind, a sample of PAR problems was chosen for analyses. Due to changes in the problems between phases, the pool of 14 potential problems was reduced to 7. Of these 7 problems, 3 required little to no explanation, as they were mostly computational, so they were also not considered. Of the 4 remaining problems, 3 were chosen such that they roughly spanned and separated the semester into thirds. These were the PAR problems from weeks 5, 10, and 14, henceforth referred to as the early semester, mid-semester, and late semester, respectively. These problems were also used as homework problems in the comparison section, but without students engaging in the PAR process of feedback and revision. For these problems, the explanation prompts were analyzed; there was one prompt for the early semester problem, two prompts for the mid-semester problem, and two prompts for the late semester problem. Due to some rewording and reorganization of the explanation prompts for the early semester problem between phases, only one prompt could be analyzed.

3.5. Scoring Procedures

The first round of scoring involved choosing a sample of student explanations for double scoring. The goal was to find a sample that generally spanned what one could reasonably expect students in the course to produce as explanations. For each PAR problem, each class was broken into quintiles according to their scores on the 10-point rubrics used for course grading. The highest and lowest quintiles were removed from the analysis to eliminate outliers. The middle 60% of students remained; these quintiles represented high, medium, and low scores on the PAR problems, respectively. A random number generator was used to select three students from each of these groups in each section. The result was 27 samples of student work for each prompt, (3 sections of the course: Phase II, Phase I, Comparison) x (3 groups of students: high, middle, low) x (3 randomly sampled students per group). Since five prompts were analyzed, a total of 135 explanations were scored. All explanations were de-identified and grading marks and comments were removed.

All work was de-identified and double-coded by Michelle and the researcher (the two experimental section instructors). Before scoring, Michelle and the researcher discussed each of the problems, to identify *key ideas* (Raman, 2003) for each prompt. To identify key ideas, multiple sources of data were used: course study guides, discussions from weekly course coordinator meetings, related student work from old exams, and prior experience with the course. Using these sources of experience, the two scorers discussed what each of the prompts was about, in the context of this specific course.

Building on these discussions, the researcher looked through the corpus of data and verified that the types of explanations students provided generally aligned with the key ideas previously identified. After doing so, a rubric and codebook using samples of student work that were not sampled for double-coding was created. After Michelle reviewed the coding materials, double coding began. For each problem, two samples of work were scored independently and discussed. Afterwards, the remaining samples were scored and discrepancies were resolved. During the scoring process, the coders were also open to considering student explanations whose logic was not built on the previously identified key ideas as correct. Yet, in the corpus of sampled explanations, no such explanations were identified; when students correctly explained the mathematics, it was generally in alignment with the key ideas already identified.

There were 8 disagreements among the 135 explanations that were double scored, for 94.1% agreement; all disagreements were resolved after discussion. After this subset of the data were double coded, the remainder of the entire corpus of data ($N = 545$ explanations total) was then scored by the researcher.

3.6. Dimensions of Scoring

Explanations were scored along four dimensions:

- *Accuracy* is the extent to which the key idea is captured correctly. Explanations with no accuracy do not correctly express a key idea, and as a result, they tend to receive a 0 on all other dimensions.[1]
- *Precision* refers to the "exactness" of the written explanation, in terms of mathematical language and symbolic expressions used.
- *Justification* focuses on the extent to which the "how" or "why" underlying the key idea is described.
- *Diagrams* are scored based on how they augment the written explanations, for instance by making connections that were otherwise not apparent.

Scoring was on a 3-point scale: (0) incorrect, (1) partially correct, and (2) correct. When students introduced incorrect ideas, they could not score a 2 for accuracy. Only diagrams could be scored N/A, which occurred when no diagram was present. Incorrect (but present) diagrams received a 0.

The 3-point rubrics were calibrated to the level of sophistication of an introductory college calculus student. Given the difficulties that students have with this course, both at the local institution and across

[1] Only a single solution that scored 0 on accuracy received a score higher than 0 on any other dimension.

the USA, one can only expect a certain level of mathematical sophistication. Thus, what was scored as correct (a 2 on the rubrics) for these students falls short of what one would expect from a professional mathematician. Yet, even with these standards, students generally scored low on the rubrics, as shown below in the results. Rather than interpreting the rubrics in the absolute sense of correctness, the rubrics should be thought of as a way to show the spread of student responses; while there was often room for improvement upon student explanations scored as 2, these explanations still represent a deeper understanding of the problems than explanations scored as a 1.

4. Analysis of Student Explanations

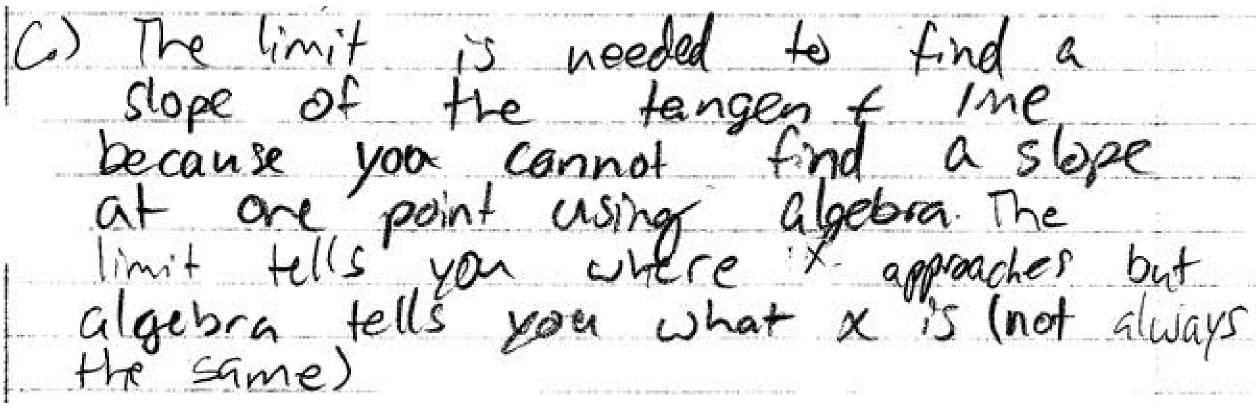
To illustrate how explanations were scored, a set of sample responses from each section is presented for the early and late semester problems (the mid-semester problem omitted due to space). In the middle quintile from each section, the highest scored response is provided, to provide a sense of an "average" explanation from each section. Thus, in addition to elaborating the coding scheme, these explanations provide a qualitative sense of how the explanations differed across sections. Following the description of scoring, quantitative results are provided.

4.1. Early Semester Problem

This problem required students to draw on their emergent understandings of limits to explain the definition of the derivative. Students were asked to illustrate the definition of the derivative graphically, explain the connection between secant lines and tangent lines, and explain why a limit is needed to find the slope of a tangent line. The task is given in its entirety in Appendix A, but only part (c) was scored, asking why the limit was needed to find the slope of the tangent line and why arithmetic and algebra were insufficient. The task was scored according to the *key idea* that: average rate of change (AROC) can only be calculated with two distinct points on the curve, so it cannot be used to find the slope of a tangent line, which (locally) intersects the function at a single point. The corresponding dimensions of scoring were:

- *Accuracy:* The explanation should state that AROC cannot be used to calculate the slope of a tangent line (i.e. the slope at a single point).
- *Precision:* The explanation should describe what AROC or IROC is, in words or an equation.
- *Justification:* The explanation should explain that AROC is insufficient *because* it requires two distinct points, or computing AROC a single point would result in 0/0. It should also describe that the limit solves this problem by allowing AROC to be computed over arbitrarily small intervals, which allows IROC to be approximated to any desired accuracy.
- *Diagrams:* No students used diagrams to support their explanations, so it was always scored as NA.

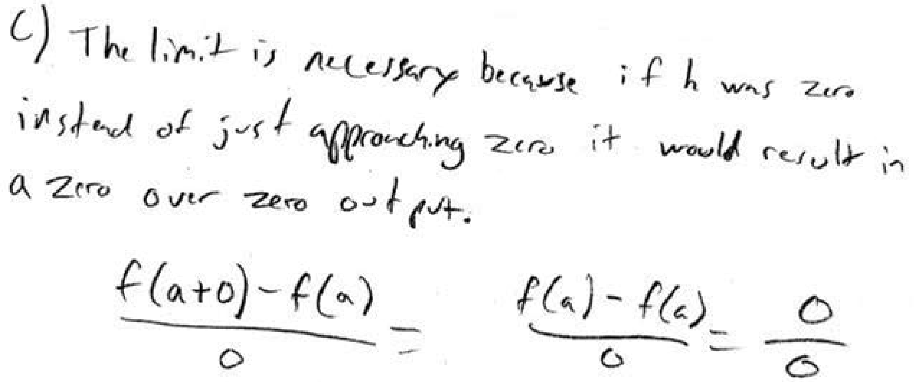
The average student responses from each section are given in Figures 1, 2, and 3. Adam (see Figure 1) received a 1 for *accuracy*. He correctly stated that "you cannot find the slope at one point using algebra," but also made the incorrect statement "the limit tells you where x approaches but algebra tells you what x is," so he received a 1 rather than 2. Adam received a 0 for *precision*, because he did not describe what AROC or IROC actually are. Adam did not address why AROC is insufficient or why the limit is needed, so he received a 0 for *justification*.



C.) The limit is needed to find a slope of the tangent line because you cannot find a slope at one point using algebra. The limit tells you where x approaches but algebra tells you what x is (not always the same)

Figure 1: Adam's solution to the early semester problem (comparison section)

Samantha's response (see Figure 2) received a 2 for *accuracy* because it described that attempting to compute AROC using a single point on the tangent line would result in 0/0, which would not work. Further, she illustrated the calculation of AROC at a single point, so she received a 2 for *precision*. Samantha received a 1 for *justification*, because she showed how this calculation would result in 0/0, but she did not address how the limit could be used to overcome this problem.

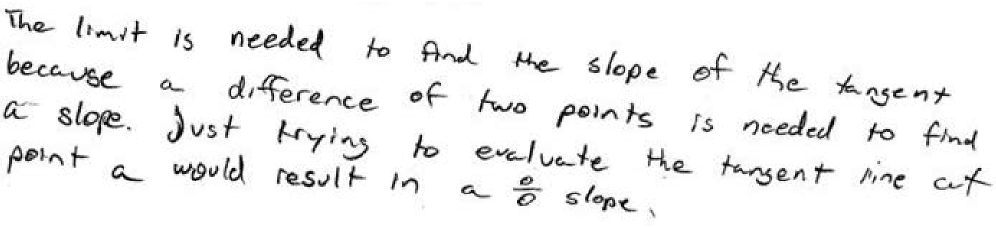


C) The limit is necessary because if h was zero instead of just approaching zero it would result in a zero over zero output.

$$\frac{f(a+0)-f(a)}{0} = \frac{f(a)-f(a)}{0} = \frac{0}{0}$$

Figure 2: Samantha's solution to the early semester problem (phase I experimental section)

Brian's response (see Figure 3) received a 2 for *accuracy* because it described how calculating AROC requires two points, but the tangent line only has one. It received a 1 for *precision*, because it described how "trying to evaluate the tangent line at point a would result in a 0/0 slope," but did not actually show this calculation. Moreover, Brian talked about finding “the difference between two points,” rather than the difference between values on the tangent line, which lacks precision. Brian received a 1 for *justification*, because he stated that the calculation would result in a 0/0 slope, but did not describe the underlying concepts (e.g., the role of the limit). Scores are summarized in Table 1.



The limit is needed to find the slope of the tangent because a difference of two points is needed to find a slope. Just trying to evaluate the tangent line at point a would result in a $\frac{0}{0}$ slope.

Figure 3: Brian’s solution to the early semester problem (phase II experimental section)

	Accuracy	Precision	Justification	Diagrams
Comparison (Adam)	1	0	0	-
Phase I (Samantha)	2	2	1	-
Phase II (Brian)	2	1	1	-

Table 1: Scores of the average student solutions to the early semester problem

4.2. Late Semester Problem

The late semester problem was the final PAR problem that students completed (see Appendix C). The problem focused on solids of revolution, requiring students to draw on their understandings of integration, volumes, and potentially limits. This problem, also known as the napkin ring problem, describes the creation of a bead by drilling a cylindrical hole through the center of a sphere. Students had to represent the bead as a volume of revolution and then rewrite their result to express the volume only in terms of the height, not the radius, of the original sphere. Students then explained two prompts related to the volume and the surface area of the bead. Parts (d) and (e) were scored for student explanations. Part (d) asked students how two beads with different starting sizes (e.g., a basketball and an orange) could have the same volume, and part (e) asked if they would also have the same surface area.

Part (d) was scored for the *key idea* that: as the outer radius of the sphere increases, the area of the cross-section that is rotated is smaller (i.e., the bead is thinner); this allows the volume of the beads to remain constant. When a cylinder is removed from the small sphere, the cross-sectional areas of the remaining portions are relatively large (in comparison to the total area), which results in a thick bead with a small outer circumference. In contrast, when the cylinder is removed from a large sphere, the cross-sectional areas of the remaining portions are relatively small, resulting in a thin bead with a large outer circumference. The dimensions of scoring were as follows:

- *Accuracy:* The explanation should state that the beads have different cross-sectional areas, which allows them to have the same volume.
- *Precision:* The explanation should describe the beads in terms of cross-sectional area or thickness, clearly identifying the relevant characteristics of the bead.
- *Justification:* The explanation should explain that different cross-sectional areas result in the same volume *because* they correspond to changes in the outer circumferences (i.e. the differences in thickness and circumference cancel each other out, to result in two beads with the same volume).
- *Diagrams:* Explanations could illustrate graphically the difference in cross-sections (e.g., see Figure 4).

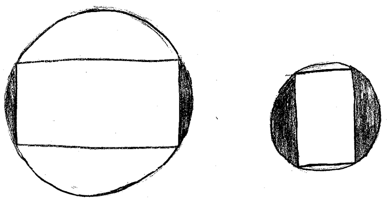
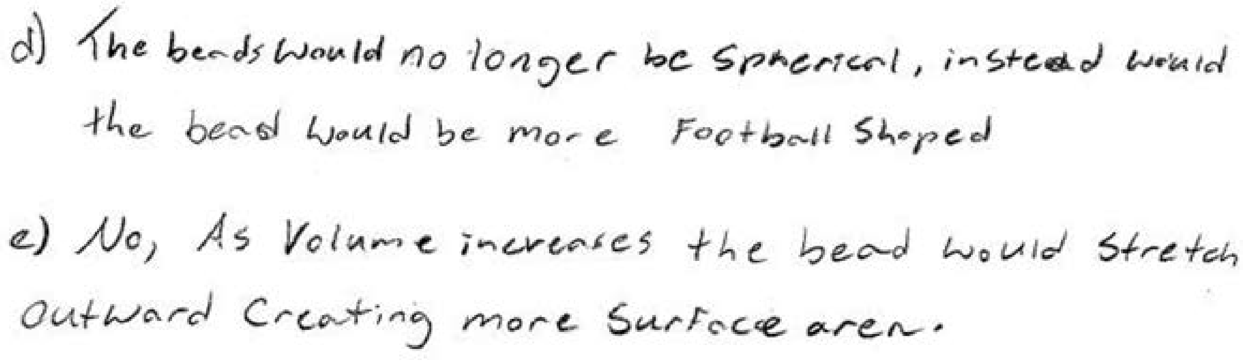


Figure 4: Katie's diagram (from Phase II) illustrating that the large sphere makes a thinner bead.

Part (e) was scored for the key idea that: beads with a larger outer radius have a larger outside surface area, due to the outside circumference. The bead made from the larger sphere has a larger surface area than the bead made from the smaller sphere, because the outside circumference is the most significant factor in determining surface area (the sphere's curvature matters but not as much). In the limit as $R \to \infty$, the outside of the bead approaches a cylinder, so the surface area approaches $2\pi Rh$, which diverges to ∞ (while the smaller spheres have finite surface area). The dimensions of scoring were as follows:

- *Accuracy:* The explanation should state that beads with a larger outer radius have a larger surface area.
- *Precision:* The explanation should describe the regions of the sphere that have an impact on surface area using terms like "outer radius" or "outside circumference."
- *Justification:* The explanation should describe that the surface area is larger *because* of how it is related to the outer radius. Not a single student mentioned the outside curvature of the bead, so this aspect of the solution was not included in analysis.
- *Diagrams:* Explanations could illustrate graphically the surface area of different beads.

Student solutions are provided in Figures 5, 6, and 7. Will (see Figure 5) stated: “the beads would no longer be spherical, instead would the bead would be more football shaped.” This statement indicates a misunderstanding of the problem situation, so Will received a 0 for all dimensions on part (d) except for diagrams, which was NA. In part (e), Will stated: “as volume increases the bead would stretch outward creating more surface area.” Yet, the problem defines all spheres as having equal volumes, so once again Will received a 0 for all dimensions other than diagrams, which was NA.



d) The beads Would no longer be Spherical, instead would the bead Would be more Football Shaped

e) No, As Volume increases the bead Would Stretch outWard Creating more Surface area.

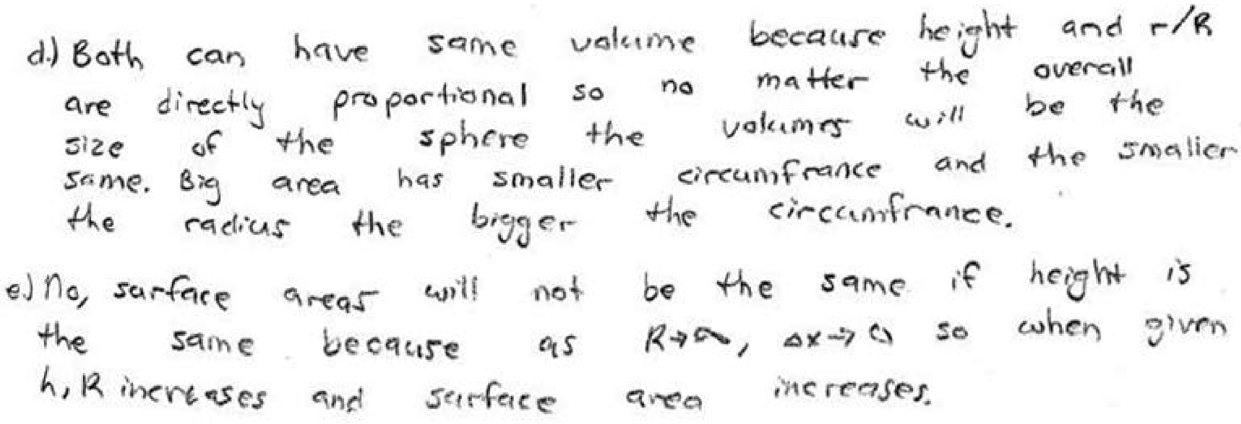
Figure 5: Will’s solution to the late semester problem (comparison section)

Silvia (see Figure 6) stated: “big area has smaller circumference and the smaller the radius the bigger the circumference.” Following Silvia’s parallel construction it is likely that she meant “the smaller the area the bigger the circumference” but accidentally wrote radius instead of area. This would indicate a correct description of the relationship between cross-sectional area and circumference. However, she only received a 1 for *accuracy*, because she said “height and r/R are directly proportional.” Yet, in this problem, h is fixed while r and R are changing, so h and r (or R) cannot be directly proportional. Because the language directly proportional does not accurately describe the relationship, and she made an error in her parallel construction (stating radius rather than area), she received a 0 for *precision.* She received a 0 for *justification* as well, because she used direct proportionality as a justification, which is incorrect.

For part (e), Silvia wrote “when given h, R increases and surface area increases,” so she received a 2 for *accuracy.* Her *precision* was a 1, because R while clearly denotes the outer radius, it is unclear what was meant by Δx. Her *justification* was a 0, because she did not relate R to surface area. *Diagrams* were NA.

Kevin (Figure 7) received a 2 for *accuracy*, because he stated: “beads of the same heights will have the same volume due to the differences in the thickness of remains portions.” He received a 2 for *precision* because his diagram (and labeling) illustrated the changing outer circumference and thickness of the beads. He received a 2 for *justification* because he explained the relationship between increasing outer circumference and decreasing thickness. Given the quality of his *diagram*, he also received a 2 for this dimension.

For part (e), Kevin received a 2 for *accuracy.* While his written explanation only stated that different size spheres will have different surface areas due to increased circumference, his diagram shows that the larger sphere would have larger surface area. Although Kevin illustrated the situation in his diagram, his written

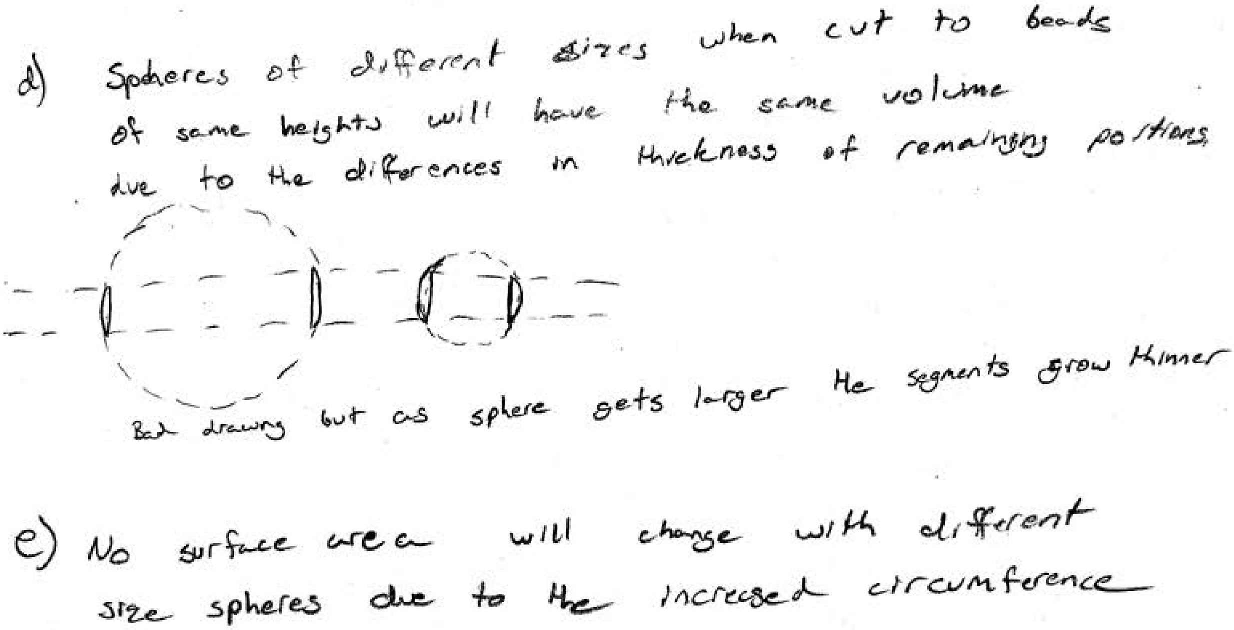


d.) Both can have same volume because height and r/R
are directly proportional so no matter the overall
size of the sphere the volumes will be the
same. Big area has smaller circumfrance and the smaller
the radius the bigger the circumfrance.

e.) No, surface areas will not be the same if height is
the same because as $R \to \infty$, $\Delta x \to 0$ so when given
h, R increases and surface area increases.

Figure 6: Silvia's solution to the late semester problem (phase I experimental section)

explanation did not clearly refer to outer circumference so he received a 1 for *precision*. Kevin received a 1 for *justification*. While his diagram shows that the larger circumference would have a larger surface area, he did not make this explicit in his writing. His *diagram* received a 2, because it illustrated how the surface area would increase with circumference.



d) Spheres of different sizes when cut to beads
of same heights will have the same volume
due to the differences in thickness of remaining portions.

Bad drawing but as sphere gets larger the segments grow thinner

e) No surface area will change with different
size spheres due to the increased circumference

Figure 7: Kevin's solution to the late semester problem (phase II experimental section)

The late semester scores are summarized in Table 2. Will's solution was representative of the other sampled solutions ($N = 9$) in the comparison section; all solutions sampled received 0 scores for all dimensions. In contrast, explanations in the experimental sections were of much higher quality.

	Accuracy	Precision	Justification	Diagrams
Comp. (Will)	0 / 0	0 / 0	0 / 0	- / -
Phase I (Silvia)	1 / 2	0 / 1	0 / 0	- / -
Phase II (Kevin)	2 / 2	2 / 1	2 / 1	2 / 2

Table 2: Scores of the average student solutions to the late semester problem (part d / part e)

5. Results

Initial data analyses involved exploring the distribution of student scores. In each section, the modal score for student explanations was 0 across all four dimensions. Yet, beyond these modal scores, the histograms were relatively flat, indicating that student explanations existed along a rich continuum, and that the coding scheme worked sufficiently well to capture this richness. To compare student outcomes across sections, the results are organized into three subsections: (1) aggregate scores, (2) scores disaggregated by dimension (3) scores disaggregated by time.

5.1. Aggregate Scores

To facilitate statistical comparison, scores were first analyzed in the aggregate (see Table 3). In each section, the dimensions of scoring (accuracy, precision, justification, and diagrams) were summed to provide a single score for each problem for each student. Then, for each section, these scores were averaged across all three problems to provide the results in Table 3. The first column of scores provides the mean score with all four dimensions included. Phase II scored 24.5% higher than the Comparison section (1.96 points on an 8-point scale), and 16% higher than Phase I. The second column gives the average scores without diagrams, because they were not explicitly prompted by the problems. Without diagrams (a 6-point scale), Phase II scored 26.7% higher than the Comparison section, and 16.3% than Phase I. Thus, the differences between sections were similar whether or not diagrams were considered. Table 3 also provides the average scores for what is defined as the **experimental** group, which represents the average of the Phase I and Phase II experimental sections.

Table 4 provides a statistical comparison of the mean scores across sections. The mean scores in the experimental sections, aggregated together, were significantly higher than the comparison section, with a medium effect size (using the Holm-Bonferroni correction for significance for multiple comparisons; Holm, 1979). The Phase II experimental section also scored significantly higher than the Phase I experimental section, with a medium effect size.

As these results show, PAR had a significant impact on the quality of student explanations. This impact was enhanced during Phase II, when students explicitly practiced analyzing sample student work.

	Total Score (Out of 8)	Diagram-Free Score (Out of 6)
Comparison	1.09 (1.98)	1.00 (1.80)
Phase I	1.77 (2.41)	1.63 (2.15)
Phase II	3.05 (3.05)	2.61 (2.47)
Experimental	**2.23 (2.72)**	**1.99 (2.32)**

Table 3: Mean scores (and standard deviations) by section

Mean Explanation Scores	Independent samples $t-$test	Sign.	Cohen's d
Comparison vs Experimental	$t = 4.9774, df = 257.45, p = 1.18 \cdot 10^{-6}$	**	0.62
Phase I vs Phase II	$t = 4.0695, df = 282.14, p = 6.12 \cdot 10^{-5}$	**	0.48

Table 4: Statistical comparisons of mean scores

5.2. Scores Disaggregated by Dimension

Table 5 provides the mean scores for each section, disaggregated by section; the mean scores for each section include the explanations that were scored for all three of the PAR problems that were analyzed. The table exhibits full monotonicity: Phase II students scored higher than Phase I students on all dimensions, and Phase I students scored higher than comparison students on all dimensions. The table exhibits monotonicity in the other direction as well: on average, students scored higher on accuracy than precision, higher on precision than justification, and higher on justification than diagrams.

	Average Scores by Dimension (Max = 2)			
	Accuracy	Precision	Justification	Diagrams
Comp.	0.50	0.30	0.22	0.08
Phase I	0.75	0.49	0.40	0.14
Phase II	1.07	0.84	0.69	0.44
Exp.	**0.87**	**0.62**	**0.51**	**0.25**

Table 5: Mean scores disaggregated by dimension (out of 2)

As these results show, students generally scored the highest on accuracy, and lower on precision and justification. This is consistent with how explanations were scored; if students received a 0 for accuracy, then they were likely to receive a 0 for precision and justification as well, because that indicates that they likely missed the key idea. These results also show that explaining *why* was the most difficult aspect of explaining for students, as indicated by the low scores on justification. Diagram scores were the lowest, which is not surprising, given that students were not explicitly asked to use diagrams.

Statistical comparisons by dimension are given in Table 6. The results are consistent with Table 5. The differences in dimensions were significant, with small to medium effect sizes.

Diagram-free Average Scores	Paired $t-$test	Sign.	Cohen's d
Accuracy vs. Precision	$t = 12.281, df = 545, p = 2.2 \cdot 10^{-16}$	**	0.53
Precision vs. Justification	$t = 5.5328, df = 545, p = 4.9 \cdot 10^{-8}$	**	0.24
Justification vs. Diagrams	$t = 9.1105, df = 545, p = 2.2 \cdot 10^{-16}$	**	0.39

Table 6: Statistical comparisons by dimension

5.3. Scores Disaggregated by Time

Table 7 provides the mean scores for each problem, disaggregated by time period. In this case, all four dimensions were summed together to provide the aggregate scores. The results in this table show a widening gap between student explanations in the three sections. At the beginning of the semester, students in the three sections performed similarly. Yet, as time went by, students in the experimental sections began to score higher than in the comparison section, and Phase II students scored notably higher than the Phase I students. The decrease in scores for comparison students is likely an artifact of the difficulty of the problems; student learning in the comparison section evidently did not keep pace with the growing challenges of the calculus course. In contrast, Phase I scores remained relatively constant, and Phase II scores actually increased over time.

	Early Semester	Mid-Semester	Late Semester
Comp.	2	1.08	0.64
Phase I	1.88	1.77	1.72
Phase II	1.92	3.21	3.52
Exp.	**1.92**	**2.30**	**2.34**

Table 7: Mean scores disaggregated by problem (out of 8)

Table 8 compares growth over time. Mean early semester scores were subtracted from the late semester scores and statistical comparisons were made on these change scores. The experimental sections improved significantly more than the comparison section during the semester (large effect size), and the Phase II section improved significantly more than the Phase I section (medium to large effect size).

Diagram-free Average Scores	Independent samples $t-$test	Sign.	Cohen's d
Comparison vs Experimental	$t = 4.8292, df = 131.66, p = 3.746 \cdot 10^{-6}$	**	0.84
Phase I vs Phase II	$t = 3.1713, df = 93.548, p = 0.002053$	**	0.66

Table 8: Statistical comparisons by time

6. Student Feedback

PAR students significantly improved their explanations, and this impact was more profound during Phase II when students discussed sample work as a class. To gain further insight into the impact of PAR, students' written feedback is now analyzed. The present analyses extend prior work showing that analyzing sample student work had a positive impact on peer conferences, supporting students to focus more on the process of solving problems, rather than the product (answer) or praise (Reinholz, 2015c).

6.1. Method

The entire corpus of student feedback from the Phase I and Phase II experimental sections was de-identified and analyzed for the early, mid, and late semester problems. Students from the comparison section are not included, because they did not engage with PAR. The first pass of coding involved reading all student feedback to identify themes. After open coding, themes were consolidated into categories for a second pass of coding (see Table 9). Feedback was coded for multiple categories, but each category was only coded once, given the difficulty of distinguishing multiple pieces of feedback of the same type.

The first five categories in Table 9 were for specific feedback about written explanations. The next two categories focused on graphs or diagrams. "Other feedback" focused on logistical issues of presentation, such as handwriting or spacing of text, not mathematics. "General clarifications" were only coded when no other feedback was present. On multiple occasions, students indicated in their peer conferences that they wrote down things like "you could explain a little more" just to put something on their paper.

6.2. Results

Table 10 shows the quantity of types of student feedback, with Phase I students giving an average of 0.87 types of feedback and Phase II students giving an average of 1.24 types. This indicates that Phase II students provided 42.5% more feedback (with respect to multiple categories) than Phase I students. These differences were significant, $\chi^2(2, N = 248) = 10.76, p = 0.0046$, Cramer's $V = 0.21$, for a small to medium effect size. During Phase I, 24% of responses had no written feedback, compared to 8.8% in Phase II.

Feedback Category	**Example**
Use of mathematical terms (e.g., suggested language, correction of terms)	You said it is an "exact approximation," which confused me.
Defining and referring to variables	In part (a), make sure you define the variables.
Avoid pronouns or clarify references	You need to say what "it" is.
Add information	You didn't explain why the limit is needed, you just said what the slope is.
Other feedback about the explanation (e.g., related to math concepts)	If you evaluate IROC without the limit you would wind up with 0/0 just using algebra. I would add that somewhere in your explanation.
Improve a graph or diagram	You need to draw a more accurate/cleaner graph; in part (a), h is the horizontal distance.
Improve a computation	Include more steps in your calculation
Other feedback not about explanation (e.g, handwriting)	Space out your answers, to make things easier to read.
General clarification	You could explain a little more.
None	Either left blank, or student said "everything looks good."

Table 9: Codes used to analyze student feedback

The percentage of feedback by category type (amount of feedback / N) is given in Table 11. Phase II students gave more feedback on the specifics of explanations and diagrams, focusing more on use of language, precision, and other specific aspects of how mathematical concepts were described. In contrast, Phase I students generally provide less of such feedback, but did provide more feedback about computations. Given the small sample sizes, Fisher's exact test was used. The differences in feedback types were significant, $p = 0.032$, Cramer's $V = 0.27$, indicating a medium effect size.

As a whole, the analysis of student feedback is consistent with differences in quality of student explanations. Students during Phase II provided significantly more feedback than their Phase I counterparts. Moreover, this feedback was more generally aligned with the dimensions of high-quality explanations that were scored. In fact, the only categories that Phase I students provided more feedback on were either

	Total Feedback		No Feedback	
	Phase I	Phase II	Phase I	Phase II
Early	58 (56)	58 (36)	9 (56)	1 (36)
Mid	38 (53)	32 (28)	15 (53)	2 (28)
Late	40 (48)	23 (27)	14 (40)	5 (27)
Total	**136 (157)**	**113 (91)**	**38 (157)**	**8 (91)**

Table 10: Quantity of types of feedback provided (*N* per section in parentheses)

computational or not related to mathematical explanations at all (e.g., focused on handwriting).

7. Summary and Discussion

Explanation is a highly-valued mathematical practice and a central part of mathematical understanding. Yet, most studies to date have focused on explanation as a tool for learning, rather than as an outcome of interest in its own right. Examining explanations as an outcome, this paper extends prior studies in two ways: (1) it documents the impact of PAR on student explanations, and (2) it provides an analytic scheme for capturing the quality of explanations.

PAR is a structured process through which students exchange feedback with one another and revise their mathematical work. While students only solved one PAR problem each week, engaging in the PAR process on a regular basis signaled that explaining one's mathematical thinking is a central part of mathematical practice. Thus, PAR helped create a classroom environment in which students knew they were expected to explain their thinking. Beyond the basic PAR cycle, Phase II students were taught explicitly how to provide better feedback to each other, by discussing sample student work. This further emphasized the importance of constructing clear explanations to students, and provided students with guidance to help their peers improve their explanations. In essence, PAR and the associated activities taught students "what counts" in this classroom environment, and helped them learn to attend to aspects of mathematical practice that are typically left implicit.

As a result, students who engaged with PAR had significantly improved explanations compared to those who did not. Moreover, students who did PAR improved the quality of their explanations significantly more over time than those who did not. These types of improvements are consistent with the nature of PAR as an intervention. The goal of PAR was to teach students to reflect on their work and to produce better explanations. Both of these processes support learning, so as students got better at PAR through practice, they also got better at learning more generally. These findings are also consistent with prior work that a

Feedback Category	**Phase I** ($N = 157$)	**Phase II** ($N = 91$)
Use of mathematical terms	7.0%	17.6%
Defining and referring to variables	4.6%	8.8%
Avoid pronouns or clarify references	0.6%	2.2%
Add information	5.1%	8.8%
Other feedback about the explanation	5.7%	17.6%
Improve a graph or diagram	22.9%	33.0 %
Improve a computation	10.2%	3.3%
Other feedback not about explanation	6.4%	5.5%
General clarification	24.2%	27.4%

Table 11: Percentage of students providing particular types of feedback

small amount of time spent teaching students to explain their ideas can have a large impact (Hodds et al., 2014).

The impact of PAR was enhanced during Phase II by teaching students how to provide better feedback; students constructed better explanations and provided better feedback in Phase II than in Phase I. Student explanations were likely enhanced for at least two reasons: (1) students received better feedback, which helped them improve their explanations more, and (2) students got better at giving feedback, which meant that were better at reflecting on their own explanations, because they knew what was important to attend to. Student feedback improved both in quantity and quality; students provided more explanation-focused feedback and moved away from commenting on procedures. In essence, students learned to emphasize explanations over procedures, which may have impacted their approach to learning more generally. This is a question for future study.

To capture changes in student explanations, I developed a new analytic scheme. The scheme was content-neutral, which allowed for changes to be measured over time. Because explanations are multi-faceted, they were considered as a cluster concept. Scoring focused on four dimensions of this cluster: accuracy, precision, justification, and diagrams. Students who used PAR improved on all dimensions, but these improvements were not all equal. Students received the highest scores on accuracy, then precision, justification, and finally diagrams. This indicates that future interventions might be designed to target specific aspects of mathematical explanations.

This approach was used to capture growth for students in the present study, and may be expanded to other contexts as well, such as physics or chemistry. For instance, once could consider the *precision* of language used in these disciplines, how ideas are *justified* based on physical principles (rather than mathematical logic), and consider the use of *diagrams* such as free-body diagrams in mechanics. Other discipline-specific dimensions may be added as well. Even in future studies of mathematics learning, the total number of dimensions could be expanded, or some of these dimensions may be replaced.

Despite its utility, the scheme has its limitations: (1) raters must agree on key ideas before scoring, which is context-dependent, (2) the scheme may fail to capture the same level of nuance as a deep qualitative analysis, (3) the scales were from 1-3, which may simplify some intermediate levels of understanding. Nevertheless, the high level of inter-rater agreement in this study suggests that reaching agreement on key ideas for a given context may not be that difficult. Moreover, one might use this scheme to complement to deeper qualitative analyses, using a mixed-methods approach. Finally, others could extend the scheme to provide additional levels of scoring, as desired.

In closing, explanation is not just a tool to promote other ends, but is a highly-valued practice in its own right. PAR provides a systematic method for helping students develop this practice with the use of relatively little class time. While PAR was structured to help students improve their explanations, the general idea of a regular routine built around peer feedback is likely applicable to other mathematical practices, such as developing conjectures, verifying proofs, constructing new proofs, and modeling real-world situations. These are areas for future study.

Acknowledgements

This research was supported by the Institute of Education Sciences, U.S. Department of Education predoctoral training grant R305B090026 to the University of California, Berkeley. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

ACARA (2009). *Shape of the Australian curriculum: Mathematics*. Technical Report National Curriculum Board Sydney, Australia.

Black, P., Harrison, C., & Lee, C. (2003). *Assessment for learning: Putting it into practice*. Berkshire, England: Open University Press.

Braaten, M., & Windschitl, M. (2011). Working toward a stronger conceptualization of scientific explanation for science education. *Science Education*, *95*, 639–669.

Bressoud, D. M., Carlson, M. P., Mesa, V., & Rasmussen, C. (2013). The calculus student: insights from the mathematical association of america national study. *International Journal of Mathematical Education in Science and Technology*, *44*, 685–698.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of educational research*, *65*, 245–281.

CCSSM (2010). *Common core state standards for mathematics*. Technical Report Authors Washington, DC.

Chi, M., De Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive science*, *18*, 439–477.

Cobb, P., Confrey, J., diSessa, A. A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, *32*, 9–13.

Collins, A., Joseph, D., & Bielaczyc, K. (2004). Design research: Theoretical and methodological issues. *The Journal of the learning sciences*, *13*, 15–42.

Crouch, C. H., & Mazur, E. (2001). Peer Instruction: Ten years of experience and results. *American Journal of Physics*, *69*, 970–977. URL: `http://scitation.aip.org.colorado.idm.oclc.org/content/aapt/journal/ajp/69/9/10.1119/1.1374249`. doi:`10.1119/1.1374249`.

Cuoco, A., Goldenberg, E. P., & Mark, J. (1996). Habits of mind: An organizing principle for mathematics curricula. *Journal of Mathematical Behavior*, *15*, 375–402.

De Villiers, M. (2003). *Rethinking proof with the geometer's sketchpad*. Emeryville, CA: Key Curriculum Press.

Engle, R. (2012). The productive disciplinary engagement framework: Origins, key concepts, and developments. In D. Yun Dai (Ed.), *Design research on learning and thinking in educational settings: Enhancing intellectual growth and functioning.*. New York, NY: Routledge.

Hanna, G. (1990). Some pedagogical aspects of proof. *Interchange*, *21*, 6–13.

Harel, G., & Sowder, L. (2007). Toward comprehensive perspectives on the learning and teaching of proof. In F. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 805–842). Charlotte, NC: National Council of Teachers of Mathematics.

Hersh, R. (1993). Proving is convincing and explaining. *Educational Studies in Mathematics*, *24*, 389–399.

Hodds, M., Alcock, L., & Inglis, M. (2014). Self-Explanation Training Improves Proof Comprehension. *Journal for Research in Mathematics Education*, *45*, 62–101. URL: `http://www.jstor.org.colorado.idm.oclc.org/stable/10.5951/jresematheduc.45.1.0062`. doi:`10.5951/jresematheduc.45.1.0062`.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, *6*, 65–70.

Jones, K. (2000). Providing a Foundation for Deductive Reasoning: Students' Interpretations when Using Dynamic Geometry Software and Their Evolving Mathematical Explanations. *Educational Studies in Mathematics*, *44*, 55–85. doi:10.1023/A:1012789201736.

Lakoff, G. (1987). *Women, fire, and dangerous things*. Chicago, IL: University of Chicago Press.

Lave, J. (1996). Teaching as learning, in practice. *Mind, Culture, & Activity*, *3*, 149–164.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in cognitive sciences*, *10*, 464–470.

Mason, J., Burton, L., & Stacey, K. (2010). *Thinking mathematically.* (2nd ed.). Harlow, England: Pearson Education Limited.

NCTM (2000). *Principles and standards for school mathematics*. Reston, VA: The National Council of Teachers of Mathematics.

Niss, M. (2011). The Danish KOM project and possible consequences for teacher education. *Cuadernos de Investigacin y Formacin en Educacin Matemtica*, *6*, 13–24.

NRC (2001). *Adding it up: Helping children learn mathematics*. Mathematics Learning Study Committee, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Oehrtman, M., Carlson, M., & Thompson, P. W. (2008). Foundational reasoning abilities that promote coherence in students' function understanding. In C. Carlson, & C. Rasmussen (Eds.), *Making the connection: research and teaching in undergraduate mathematics education* (pp. 27–42). Washington, DC: Mathematical Association of America.

Osborne, J., & Patterson, A. (2011). Scientific argument and explanation: A necessary distinction? *Science Education*, *95*, 627–638.

Raman, M. (2003). Key ideas: what are they and how can they help us understand how people view proof? *Educational Studies in Mathematics*, *52*, 319–325.

Reinholz, D. L. (2015a). The assessment cycle: A model for learning through peer assessment. *Assessment & Evaluation in Higher Education*, (pp. 1–15). doi:10.1080/02602938.2015.1008982.

Reinholz, D. L. (2015b). Peer-Assisted Reflection: A design-based intervention for improving success in calculus. *International Journal of Research in Undergraduate Mathematics Education*, *1*, 234–267. doi:10.1007/s40753-015-0005-y.

Reinholz, D. L. (2015c). Peer conferences in calculus: The impact of systematic training. *Assessment & Evaluation in Higher Education*, (pp. 1–17). doi:10.1080/02602938.2015.1077197.

Sadler, D. (1989). Formative assessment and the design of instructional systems. *Instructional science*, *18*, 119–144.

Schoenfeld, A. H. (1991). Whats all the fuss about problem solving. *ZDM*, *91*, 4–8.

Schoenfeld, A. H. (2014). What Makes for Powerful Classrooms, and How Can We Support Teachers in Creating Them? A Story of Research and Practice, Productively Intertwined. *Educational Researcher*, *43*, 404–412. doi:10.3102/0013189X14554450.

Sfard, A. (1998). On two metaphors for learning and the dangers of choosing just one. *Educational Researcher*, *27*, 4–13.

Smith, J., diSessa, A. A., & Roschelle, J. (1993). Misconceptions Reconceived: A Constructivist Analysis of Knowledge in Transition. *Journal of the Learning Sciences*, *3*, 115–163.

Steiner, M. (1978). Mathematical explanation. *Philosophical Studies*, *34*, 135–151.

Tall, D. (1992). Students difficulties in calculus. In *Proceedings of Working Group 3 on Students Difficulties in Calculus* (pp. 13–28). International Congress on Mathematics Education volume 7.

Weber, K. (2014). Proof as a cluster concept. In *Proceedings of the Joint Meeting of PME* (pp. 353–360). Citeseer volume 38.

Wilson, R., & Keil, F. (1998). The shadows and shallows of explanation. *Minds and Machines*, *8*, 137–159.

Wong, R. M., Lawson, M. J., & Keeves, J. (2002). The effects of self-explanation training on students' problem solving in high-school mathematics. *Learning and Instruction*, *12*, 233–262. doi:10.1016/S0959-4752(01)00027-5.

Appendix A. Early Semester Problem

Draw an accurate graph of your favorite nonlinear function $y = f(x)$ (no formula for $f(x)$ needed) and pick a point on the x-axis and label it "a". (Make the graph fairly large so you can clearly draw other things on it.) Recall that the derivative of a function f at a point $x = a$ is defined by

$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

(a) Illustrate and label on your graph each of the following quantities that appear in the definition. Then write a short statement explaining in terms of your graph what each quantity means (1-2 sentences for each quantity).

(i) $f(a)$ (ii) h (iii) $f(a+h)$ (iv) $\frac{f(a+h)-f(a)}{h}$ (v) $f'(a)$

(b) Explain in terms of the graph what the equation $f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$ means (be sure to talk about secant lines and the tangent line).

(c) Explain why the limit is needed to find the slope of the tangent line. (Why can't we just use arithmetic and algebra?)

Appendix B. Mid-Semester Problem

In this problem, you will trace the shape of your hand and approximate the area of the picture that you create. Your main tasks are to devise a method for approximating the area and to show your approximation is very close to the actual area.

(a) Put your hand flat on the grid provided (with fingers touching, no gaps) and trace the shape of the outline of your hand. Make sure that the shape you trace is a function (if not, erase the parts of the shape that would make it not a function).

(b) Devise a method to approximate the area of the region inside the curve you have traced. Explain your method in detail, and explain why it should work. (Don't perform any calculations yet.)

(c) Use the method you described above to approximate the area of the outline of your hand. (Show your work.)

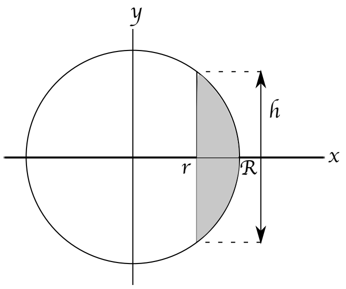
(d) Describe a method for estimating the error in your method of approximation. (Error is something you would like to make *small!* Thus an estimate for the error means being able to say the error is *less than* some value.)

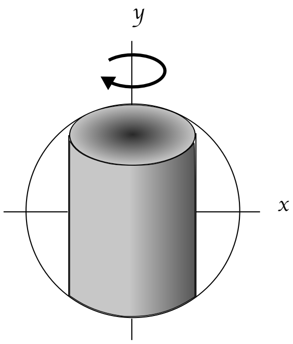
(e) Calculate an estimate of the error for your method.

(f) Explain (in principle) how you could improve your method to make your estimate as accurate as one could want (i.e., minimize the error). (You do not actually have to perform the calculations, just explain what you would do.)

Appendix C. Late Semester Problem

A bead can be formed by removing a cylinder of radius r from the center of a sphere of radius R (see the figure below).





(a) Use calculus to find the volume V of the bead with $r = 1$ and $R = 2$.

(b) Use calculus to find the volume V of a bead in terms of the variables r and R.

(c) The bead's height h is labeled in the figure. Rewrite your formula from (b) to show that $V = \frac{\pi}{6}h^3$.

(d) Since your answer in part (c) expresses the volume entirely in terms of h (and not r or R), it means that all beads of the same height have the same volume. In other words, if you started with a sphere the size of an orange and a sphere the size of a basketball and made them each into beads a height of 2 inches, the beads would have the same volume. Explain how this can be true. (Hint: think about the shape of the beads)

(e) Do all beads of the same height h also have the same outside surface area (not including the surface area of the cylindrical hole inside)? (Note: you do not need to use an integral to compute the surface area, just discuss it intuitively.)