Int. J. Res. Undergrad. Math. Ed. (2015) 1:234–267
DOI 10.1007/s40753-015-0005-y

Peer-Assisted Reflection: A Design-Based Intervention for Improving Success in Calculus

Daniel L. Reinholz[1]

Published online: 20 May 2015

Abstract Introductory college calculus students in the United States engaged in an activity called Peer-Assisted Reflection (PAR). The core PAR activities required students to: attempt a problem, reflect on their work, conference with a peer, and revise and submit a final solution. Research was conducted within the design research paradigm, with PAR developed in a pilot study, tried fully in a Phase I intervention, and refined for a Phase II intervention. The department's uniform grading policy highlighted dramatic improvements in student performance due to PAR. In Phase II, the department-wide percentage of students (except for the experimental section) who received As, Bs, and Cs in calculus 1, compared to Ds, Fs, and Ws (withdrawal with a W but no grade on a transcript), was 56 %. In the experimental section, 79 % of students received As, Bs, and Cs, a full 23 % increase. Such increased success has rarely been achieved (the Emerging Scholars Program is a notable program that has done so.)

Keywords Explanation · Formative assessment · Peer assessment · Reflection

Introduction

This paper documents the use of reflection tools to improve student success in calculus. Since the calculus reform movement (Ganter 2001), calculus learning has been a major research focus in the United States (US), with over 2/3 of departments reporting at least modest reform efforts (Schoenfeld 1995). Despite some successes, introductory college

This work is based on Daniel Reinholz's dissertation work, chaired by Alan Schoenfeld at the University of California, Berkeley.

✉ Daniel L. Reinholz
daniel.reinholz@colorado.edu

[1] Center for STEM Learning, University of Colorado, Boulder, 393 UCB, Boulder, CO 80309, USA

calculus remains an area of persistent difficulty. In the US, each fall semester, over 80,000 students (27 %) fail to successfully complete the course (Bressoud et al. 2013).

Calculus concepts, such as functions (Oehrtman et al. 2008) and limits (Tall 1992), are notoriously difficult for students. These conceptual difficulties are exacerbated by the challenges of the high school to college transition (e.g., developing greater independence, learning new study habits, forming new relationships; cf. Parker et al. 2004). Moreover, many students enter college unprepared; only 26 % of 12th grade students achieve a level of proficient or better on the National Assessment of Educational Progress (NAEP) exam (NCES 2010). Through K12 instruction, students often develop learning dispositions that do not align well with the requirements of collegiate mathematics (Schoenfeld 1988). All of these factors impede student success in calculus.

To help students succeed in calculus, I engaged in three semesters of study using the design-based research paradigm (Cobb et al. 2003). Design-based research aims to make practical and theoretical contributions in real classroom settings (Brown 1992; Burkhardt and Schoenfeld 2003; Gutiérrez and Penuel 2014). By specifying the theoretical underpinnings of my design in detail, I developed a practical instructional tool and refined principles of why it works (Barab and Squire 2004). Over three semesters of design, I developed a collaborative activity called Peer-Assisted Reflection (PAR). PAR was developed in a pilot study, tried fully in a Phase I intervention, and refined for a Phase II intervention.

The core PAR activities required students to: (1) work on meaningful problems, (2) reflect on their own work, (3) analyze a peer's work and exchange feedback, and finally (4) revise their work based on insights gained throughout this cycle. PAR was based on theoretical principles of explanation (Lombrozo 2006) and assessment for learning (Black et al. 2003). In particular, PAR leverages the connection between peer analysis and self-reflection (Sadler 1989) to help students develop deeper mathematical understandings (Reinholz 2015).

During Phase I and Phase II of the study, I used quasi-experimental methods to study the impact of PAR on student outcomes in calculus. The study took place in a mathematics department with many parallel sections of the same course, all of which used common exams and grading procedures. This paper focuses primarily on student understanding as measured by exam performance. Two companion pieces (Reinholz forthcominga, forthcomingb) provide in-depth analyses of student explanations and the evolution of the PAR design.

This paper is organized into three major components. The first component describes the PAR intervention, including its theoretical basis, core activities, and a brief history of its evolution. The next component focuses on the impact of PAR on student performance during Phase I and Phase II of the study. Finally, I further elaborate the intervention by discussing the impact of students' revisions and the mechanisms that appeared to make PAR such an effective intervention.

Background

Efforts to Improve Calculus Learning

To date, two of the most notable efforts to improve calculus learning in the US were the calculus reform movement (Ganter 2001) and Emerging Scholars Program (ESP;

Fullilove and Treisman 1990). Internationally, calculus continues to be an area of interest, as evidenced by the recent ZDM special issue focused on calculus research (Rasmussen et al. 2014). As these researchers note, a number of advances have been made (e.g., in understanding how students learn specific concepts), but "these advances have not had a widespread impact in the actual teaching of and learning of calculus" (Rasmussen et al. 2014, p. 512). Accordingly, I focus primarily on the ESP and calculus reform movement, both of which have had notable impacts on the teaching and learning of calculus.

The ESP is based on Treisman's observational study of minority learners (Fullilove and Treisman 1990); the ESP seeks to reproduce the learning conditions of successful students from the original study. Students in the ESP attend special 2-hour problem sessions, twice a week, in addition to their traditional calculus section. In the sessions, groups of 5–7 students work collaboratively on exceptionally difficult sets of problems; both the quality of the problems and the collaborative environment are essential (Treisman 1992). The ESP is open to students of all races, but enrolls primarily minority students; African American students have increased their success rates by 36 % through their participation (Fullilove and Treisman 1990). Versions of the ESP at other institutions (e.g., the University of Texas at Austin, the City College of New York) have also improved outcomes for minority students (Treisman 1992). Although ESP-style learning has been difficult to implement in regular calculus sections, the ESP provides evidence of the impact of meaningful problems in a supportive, collaborative learning environment.

Calculus reform interventions often introduced technology and/or collaborative group work to help students solve real-world problems. These studies generally reported positive improvements in engagement and deeper understanding, with mixed performance on traditional exams (Ganter 2001); however, it is difficult to generalize from these studies, due to lack of common measures (e.g., many of them did not compare student passage rates directly). To contextualize the present study, I report on some notable efforts. The Calculus Consortium at Harvard impacted a number of universities, with one of the most notable outcomes the gain of 12 % improvement in passage rates documented at the University of Illinois at Chicago (Baxter et al. 1998). Nevertheless, this finding is limited, because students were not compared using the same exams. Smaller positive gains were noted in the Calculus, Concepts, Computers, and Cooperative Learning (C4L) program, which showed a 4 % improvement in course GPA scores (Schwingendorf et al. 2000). Other notable efforts, such as Calculus and Mathematica (Roddick 2001) and Project CALC (Bookman and Friedman 1999) did not directly compare student outcomes in Calculus I, but comparisons of the GPAs of traditional and reform students in subsequent courses showed mixed results. As a whole these studies show promise, but in many cases the outcomes were difficult to interpret due to the methodological difficulties of conducting such studies. I attempt to account for some of these issues in the present study by comparing students using a common set of exams.

Explanation and Understanding

Although mathematical understanding has been defined in a number of ways, there is general consistency between recent attempts to create useful definitions (NGAC 2010;

NCTM 2000; Niss 2003; NRC 2001). These standards and policy documents tend to focus on learning as both a process of acquiring knowledge and as the ability to engage competently in social, disciplinary practices (Sfard 1998). The standards focus on holistic learning, and as a result, focus on a large number of skills and practices. This is an important shift for improving mathematical teaching, learning, and research, but it also highlights the difficulty of measuring learning in a meaningful way.

Explanation is a highly-valued mathematical practice, and is considered a "hallmark" of deep understanding in the common core state standards (NGAC 2010). This aligns well with the five NCTM process standards, of which explanation is fundamental to three (reasoning and proof, communication, and connections) and important to the other two (problem solving and representation; NCTM 2000). Explanation is also prevalent in the Danish KOM standards (e.g., in reasoning and communication; Niss 2003).

Explanation also supports learning, because it helps individuals uncover gaps in their existing knowledge and connect new and prior knowledge (Chi et al. 1994). In this way, explanation provides students with opportunities to grapple with difficult mathematical concepts in a supportive environment, so that they can learn to overcome conceptual difficulties rather than avoid them (Tall 1992).

Focusing on student practices, explanation supports productive disciplinary engagement (Engle and Conant 2002). This framework provides a lens for understanding the types of activities students engaged in through PAR. Engle and Conant (2002) describe four principles for productive disciplinary engagement: (1) problematizing, (2) authority, (3) accountability, and (4) resources. As a whole, these principles require that students work on authentic problems and are given space to address the problems as individuals, but are held accountable to their peers and the norms of the discipline. PAR was designed to support such engagements, because it provides students with opportunities to explain and justify their ideas on rich mathematical tasks, and receive and incorporate feedback from their peers.

Using Assessment for Learning

The PAR intervention was designed using principles of assessment for learning to support student understanding. Recognizing students as partners in assessment (Andrade 2010), such activities focus on how students can *evoke* information about learning and use it to *modify* the activities in which they are engaged (Black et al. 2003). Generally speaking, assessment for learning improves understanding (e.g., Black et al. 2003; Black and Wiliam 1998). In the present study, students analyzed their peers' work to develop analytic skills that they could later use to reflect on their own work (Black et al. 2003).

Through reflection, an individual processes their experiences to better inform and guide future actions (Boud et al. 1996; Kolb 1984). In the context of PAR, these experiences focused on problem-solving processes, such as metacognitive control, explanation, and justification. I use the terms analysis and reflection, rather than assessment, to distinguish PAR from other activities focused on assigning grades, which contain little information to support such learning processes (Hattie and Timperley 2007).

To self-reflect, a learner must: (a) possess a concept of the goal to be achieved (in this case, a high-quality explanation), (b) be able to compare actual performance to this goal, and (c) act to close the gap between (a) and (b) (Sadler 1989). Through actual

practice analyzing examples of various quality, students can develop a sense of the desired standard and a lens to view their own work. This allows students to reflect on and improve their mathematical work (Reinholz 2015).

Many researchers have studied students analyzing the work of their peers (Falchikov and Goldfinch 2000), but they focus primarily on peer writing (e.g., Min 2006). Most of these studies have focused on calibration between peer and instructor grades, rather than peer analysis as a tool for learning (Stefani 1998). Even studies focused on learning rarely measured quantitative changes in student outcomes (Sadler and Good 2006). The present study is unique because it focuses on the impact of peer analysis on student outcomes in a domain where such activities are rare.

In this article I define PAR as a specific activity structure, but in theory, PAR could be implemented in other ways. PAR involves students analyzing one another's work and conferencing about their analyses. Through peer-conferencing, students explain both their own work and the work of their peers, which promotes understanding.

Research Questions

In alignment with prior work, this paper addresses three research questions:

- Did PAR improve student exam scores and passage rates in introductory calculus?
- How did PAR impact student performance on problems that required explanation compared to those that did not?
- In what ways did PAR appear to support student learning?

The first research question focuses on whether or not PAR can help address the persistent problem of low student success in calculus. Although PAR targets student explanations specifically, I was interested in whether or not PAR could improve student performance more broadly (research question two). I address these two questions through quantitative analyses of student performance during Phase I and Phase II of the study. To understand the ways that PAR appeared to support learning, I analyzed how students revised their work and also conducted interviews with students.

Core PAR Activities

Students were assigned one additional problem (the "PAR problem") as a part of their weekly homework (for a total of 14 problems throughout the semester). The core PAR activities required students to: (1) complete the PAR problem outside of class, (2) self-reflect, (3) trade their initial work with a peer and exchange peer feedback during class, and (4) revise their work outside of class to create a final solution. Students turned in written work for (1)–(4), but only final solutions were graded for correctness. During their Tuesday class session, students were exposed to each other's work for the first time (unless they worked together outside of class). Each student analyzed their partner's work silently for 5 min before discussing the problem together for five more minutes, to ensure that students focused on one another's reasoning and not just the problems themselves. This meant that students spent a total of approximately 10 min of class time each week dedicated to PAR; this was a relatively small amount of the 200 min of class time that students met each week. Most of the time students spent working on PAR took place outside of class.

Through PAR, students practiced explanation; gave, received, and utilized feedback; and practiced analyzing others' work. PAR feedback was timely (before an assignment was due; cf. Shute 2008) and the activity structure (submission of both initial and final solutions) supported the closure of the feedback cycle (Sadler 1989). Through repeated practice analyzing others' work, students were intended to transition from external feedback to self-monitoring (see appendices A and B for the self-reflection and feedback forms). To support students to meaningfully engage in PAR conferences, students practiced analyzing hypothetical work during class sessions and discussed it as a class.

PAR problems were inspired and modified from: the Shell Centre, the Mathematics Workshop Problem Database (a database of problems used in the ESP), *Calculus Problems for a New Century*, and existing homework problems from the course. I further narrowed the problem sets by drawing on Complex Instruction, a set of equity-oriented practices for K-12 group work (Featherstone et al. 2011), and Schoenfeld's (1991) problem aesthetic. Ideal problems: were accessible, had multiple solution paths to promote making connections, and provided opportunities for further exploration. Most problems required explanation and/or the generation of examples; as a result, each pair of students was likely to have different solutions. Thus, these tasks could be considered real mathematical problems, not just exercises (related to problematizing; cf. Engle and Conant 2002).

I illustrate PAR by discussing a student interaction around PAR10 (the 10th assigned problem). PAR10 required students to trace their hand, use simple shapes to estimate the enclosed area, and estimate an error bound (see Fig. 1). This interaction was chosen because it illustrates how peer discussions were able to support meaningful revisions.

To begin, Peter and Lance completed PAR10 as homework. Figure 2 shows Peter's work to estimate the error of his method. Peter illustrated his hand and labeled 28 unit squares that were all entirely inside of the hand. Peter's method to calculate error was to "make rough estimates" of how much area he left out, which he reasoned "should be within 5 % of the actual value." However, Peter had no bound on the accuracy of his "rough estimates," so he may have actually provided an underestimate, rather than an upper bound of error. A portion of Peter's self-reflection is given in Fig. 3. Targeted checkboxes helped students focus on specific areas of communication.

In class, Peter traded his work with Lance and they spent 5 min silently reading and providing feedback. Lance's feedback (see Fig. 4) told Peter to calibrate his units of measurement to a known unit, but Peter ignored this advice (see Fig. 5).

After exchanging written feedback, the students discussed the problem. Peter had focused on estimating the error from the inside, while Lance focused on estimating from the outside, but both solutions were incomplete. These different perspectives connected productively in the PAR conversation.

[12] Lance: Were you trying to…get an under?

[13] Peter: Yeah, initially.

[14] Lance: What I was thinking was you could make an over-approximation. Take this right here and create an over-approximation and then subtract what you got from here with your under-approximation and it should get you this space.

PAR10: Hand Area

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 that your approximation is very close to the actual area.

1. 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).
2. 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.)
3. Use the method you described above to approximate the area of the outline of your hand. (Show your work.)
4. 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.)
5. Calculate an estimate for the error for your method.
6. 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.)

Fig. 1 *PAR10: Hand Area*

[15] Peter: So it's showing it has to be between those two values. That's the error.

[16] Lance: Right, that actual value is going to be between your low approximation and your high approximation.

…

[23] Peter: It says you want to think about bounding your error with some larger value. So that would make sense then.

Lance asked Peter if he was trying to get an under approximation (line 12), and then stated that he was trying to get an over approximation (line 14). Peter and Lance realized that combining these ideas together, they could get bounds on the actual value

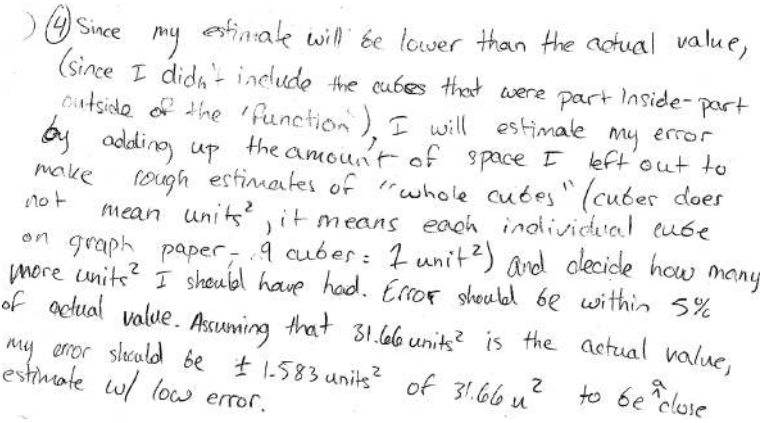

Fig. 2 Peter's initial solution to PAR10, Part 4

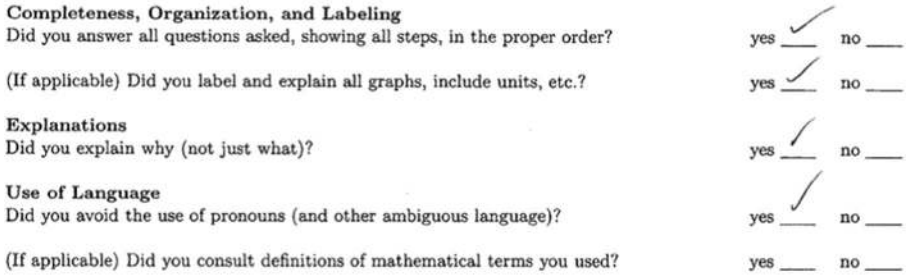

Fig. 3 Peter's self-reflection

from above and below (lines 15 and 16). Peter used this idea to come up with a correct method for bounding the error (see Fig. 6).

In Peter's final solution, he calculated an over approximation and an under approximation, and reasoned that the actual value must be between the two. Peter used the boxes from his initial solution that were entirely inside the hand as an underestimate. He then added additional boxes that surrounded the outside of his hand and reasoned that "since this will be an over approximation, I know that the true area under the curve will be less than the area I calculate by error." While not all PAR conversations resulted in productive revisions, this example illustrates the PAR process. PAR was designed to provide students with the authority to grapple with rich problems while holding students accountable to their peers through conferencing, key components of productive disciplinary engagement (Engle and Conant 2002).

Development of the Intervention

Although literature supported using peer analysis to promote self-reflection, it did not specify instruction in detail. Thus, PAR was developed over the three semesters of study. An in-depth analysis of the evolution of the PAR design is reported on in a companion piece (Reinholz forthcomingb). For the present paper, I highlight three crucial areas of development: (1) the use of real student work, (2) randomization of partners, and (3) student training.

Use of Real Student Work

The basic PAR activity structure was developed in a pilot study in a community college algebra classroom (during spring 2012). The class consisted of 50 % females and 79 % traditionally underrepresented minorities (of 14 students after dropouts) who were simultaneously enrolled in a remedial English class. Experienced educational designers and community college instructors advised me to have students analyze hypothetical work to mitigate possible issues from students having their work analyzed by peers.

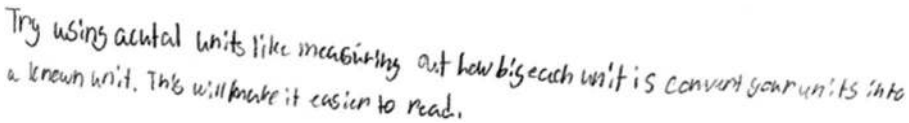

Fig. 4 Feedback received by Peter

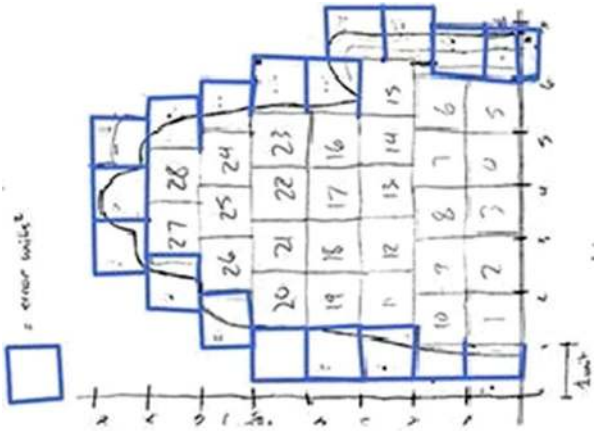

Fig. 5 Peter's final illustration of his hand area

However, even by mid-semester, students struggled to analyze hypothetical work. For instance, on the 6th homework assignment students were asked to rank order and analyze four sample explanations. Not a single student provided a clear rationale for their ordering. Some students also remarked that they did not understand the purpose of these activities.

To address these difficulties, I instead had students analyze each other's work and provide verbal and written feedback. Contrary to initial concerns, students seemed comfortable and engaged with the activity. Students now received immediate feedback from their peers as to whether or not their explanations were understood. As one student, Teresa, described:

> I didn't get it before, why you were always asking us to explain, but now it makes sense.When you don't explain things people can't tell what you're doing.

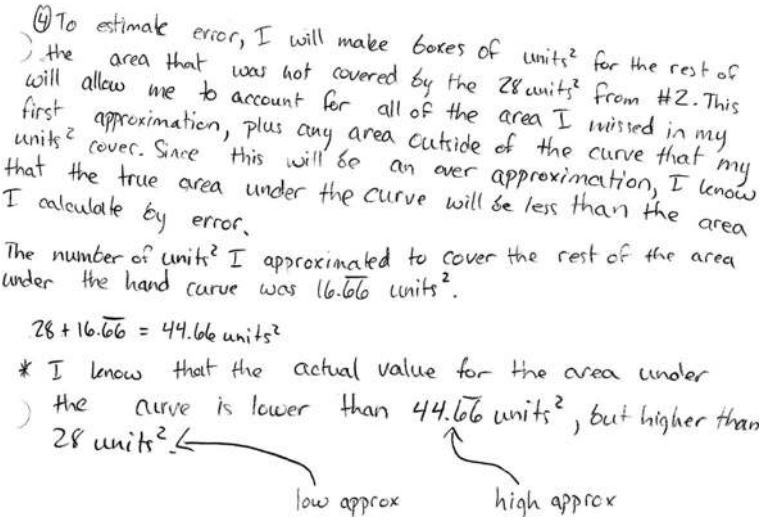

Fig. 6 Peter's final solution to PAR10, part 4

Students were now able to discuss their analyses, and the activity was more meaningful, because students could see themselves as helping a peer. Because students had to present to a peer, they were held accountable for the quality of their work by peers in addition to the instructor (Engle and Conant 2002). Finally, students received feedback that they could use to revise their own work (providing additional opportunities or "resources" for improvement). For all of these reasons, the revised activity structure was adopted as the basis for PAR. The PAR procedures for how and when students would engage in this process were solidified at the beginning of Phase I (as described in the Core PAR Activities section).

Randomization of Partners

During Phase I, a small subset of students had short, superficial peer conferences.

Consider Nicki and Alex's discussion of PAR10. Nicki had a mostly correct solution while Alex had only solved half of the problem. In the transcript below, both students spoke sarcastically, as though they were not using PAR as a serious learning opportunity:

[2] Nicki: I think you did it right, except for the last 3 parts. *(in a sing-song voice)*

[3] Alex: Yeah, totally! *(sarcastically)*

[4] Nicki: Do you know how to do it, just using triangles?

[5] Alex: Yeah, I got that.

[6] Nicki: You gotta add the ones underneath, and subtract the other ones.

[7] Alex: Yep.

[8] Nicki: It looks pretty good, and then for more accuracy, you could do some more triangles.

[9] Alex: Even more triangles. *(sarcastically)*

[10] Nicki: And more triangles. *(sarcastically)*

[11] Alex: I said yours is awesome, and, yeah.

In contrast to most student conversations (e.g., Peter and Lance's conference above), Nicki did not discuss the concepts at all. She simply told Alex what to revise (lines 6, 8, and 10).

Alex provided no feedback, which was atypical. Alex revised his solution, but apparently did so without understanding, because he still answered the question incorrectly.

These superficial conversations took place between a small subset of students who worked with the same partners repeatedly. These students appeared to be

working with their friends in the class, and usually did not provide useful feedback, or simply provided answers to one another. This issue was addressed in Phase II by having students sit in a random seat on PAR day; I did not observe such conversations during Phase II. Having students sit in a random seat as they entered the classroom also meant that little to no time was required for students to find partners.

Training of Students

Students practiced analyzing sample work through a weekly training activity. Each activity was written to correspond to that week's PAR problem. Students were typically given 2–3 min to think about three samples of student work, and then spent about 3–5 min discussing the work as a class. This is time that other sections would typically spend on lecture. I provide an example in which students began the class session by silently analyzing the work given in Fig. 7.

After students analyzed the work silently, they had a whole class discussion about their analyses. In what follows, students describe some of their observations about what is problematic about the second sample solution given above (see lines 4, 10, and 14).

[3] Instructor: What about number 2?

(Three students shake their heads no, Patrick, Colton, and Barry)

[4] Patrick: In the lab we just did, we created that graph to show that midpoints aren't always more accurate.

[5] Instructor: So midpoints aren't always the best. What else?

[6] Sue: Is this just general, or about the PAR?

[7] Instructor: These are always about the PAR

Instructions: Classify the "bullseye" (correct explanation), "on the board" (a mostly correct idea that is communicated poorly or has a minor error), and "off the mark" (incorrect) solutions.

Prompt: Explain (in principle) how y ou could improve your method to make your estimate as accurate as one could want (i.e., minimize the error).

Sample 1: If I took a limit as the width of the rectangles approaches 0 (making the number of rectangles approach ∞), the difference in the area under the curve and the rectangles would approach 0.

Sample 2: You could use midpoints rather than endpoints and it will be more accurate because there will be less overlap.

Sample 3: If I had more rectangles there would be less overlap and the approximation would be better.

Fig. 7 Sample solutions to the PAR10 training activity

[8] Sue: Wouldn't midpoints be better?

[9] Instructor: What do you think, would midpoints be better?

[10] Barry: Would it even matter, because it says "as accurate as you would want," and you can only get so accurate with midpoints?

[11] Instructor: What do you guys think about that? Did you not hear him, or do you disagree?

[12] Jim: I couldn't hear him.

[13] Instructor: Could you shout it from a mountain Barry?

[14] Barry: Yeah, so I just said that the prompt asks how you could improve the method to make your estimate as accurate as you would want, but using midpoints you can only make it so accurate, which is a problem.

After the instructor asked students what they thought about the explanation (line 3), a number of students shook their heads, indicating they thought it had problems. In line 4, Patrick connected the calculator lab that the class had been working on to the existing prompt, noting that midpoints don't necessarily create the most accurate estimate. Sue was unsure about this, so she asked to clarify (line 8). Rather than answering himself, the instructor allowed the class to respond (giving them authority in the discussion). Barry gave an explanation for why the midpoint method is insufficient to produce arbitrary accuracy (lines 10 and 14). As this brief transcript highlights, students had opportunities to analyze various explanations and explain their reasoning (developing authority; cf. Engle and Conant 2002). This gave students opportunities to calibrate their own observations to the perspectives of their peers and instructor. I now analyze the impact of the intervention.

Phase I (Fall 2012)

Materials and Methods

Phase I took place in a university-level introductory calculus course in the US targeted at students majoring in engineering and the physical sciences.[1] The course met 4 days a week for 50 min at a time. Ten parallel sections of the course were taught using a common syllabus, curriculum, textbook, exams, grading procedures, calculator labs, and a common pool of homework problems (instructors chose which problems to

[1] All research reported in this manuscript was conducted in accordance with the ethical standards in the Helsinki Declaration of 1975, as revised in 2000, as well as national law, with approval of the appropriate Institutional Review Board.

assign). Many of the PAR problems were drawn from this pool, but some were used only in the experimental section.

The calculus course was carefully coordinated, with all instructors meeting on a weekly basis to ensure alignment in how the curriculum was taught. Historically, the course had been taught using primarily a lecture-based format, which I confirmed through observations of three of the comparison sections. Instructors generally dedicated the same number of days to the same sections within the book and covered similar examples. The experimental section also used a lecture format, with some opportunities for student presentations and group work. The primary difference between the experimental and comparison sections was the use of PAR, as described in the Core PAR Activities section. Students had some opportunities to analyze hypothetical work to develop analytic skills, but during Phase I the systematic training procedure had not yet been implemented; students only engaged in three training activities during the entire semester.

Participants

Most sections of the course had 30–40 students, with a few large sections of 50–90 students (see Table 1). Students enrolled in the course as they normally would, with no knowledge that there was an experimental section. On the first day of class, students in the experimental section were given an opportunity to switch sections, but none did. The study classifies as quasi-experimental, because students were not randomly assigned to sections. While there may be systematic differences in the students who enrolled in different sections, I have no data to suggest that any particular section was atypical. As I describe later, the analysis of Exam 1 scores (as a proxy for a pre-test) indicates that the sections were indeed comparable. Moreover, demographics were collected during Phase II, and there were no significant differences between sections.

Michelle, who had a PhD in mathematics education and nearly 10 years of teaching experience, taught the experimental section and one of the comparison sections. Michelle taught two sections to help control for the impact of teacher effects. Of the two sections Michelle taught, I had her use PAR in the larger section, to garner evidence that PAR could be used in a variety of instructional contexts (not just small classes). Michelle used identical homework assignments and classroom activities in

Table 1 Phase I data collection table

	Experimental	Comp. 1	Comp. 2	Comp. 3	Comp. (Other)	Total
Instructor	Michelle	Michelle	Heather	Logan	–	–
Students	56	18	38	67	230	409
Participants	53	17	29	54	–	163
Video Obs.	45	6	6	6	–	63
Interviews	14	–	–	–	–	14
PAR Conv.	54	–	–	–	–	54

both sections, except for PAR (students in both sections completed the PAR problems, but the comparison students did not conference about their work).

Comparison instructors were chosen who had considerable prior teaching experience. Heather, a full-time instructor with over a decade of teaching experience, taught another observed comparison section. Logan, an advanced PhD student, taught the final observed comparison section. All observed instructors had taught the course a number of times before. Teachers in the comparison sections taught the course as they normally would.

Data Collection

To document changes in student understanding, I collected exam scores and final course grades for students in all sections of the course; all students took common exams, which allowed me to compare the experimental section to the department average scores. To study student interactions, I video recorded class sessions of the experimental section and three comparison sections. I performed all video observations with two stationary video cameras: one for the teacher, one for the class. As a researcher, I attended all class sessions, taking field notes of student behaviors and class discussions. In the experimental section, I also scanned students' PAR assignments and made audio records of students' conversations during peer-conferences. After the second midterm, I conducted semi-structured interviews with students in the experimental section about their experiences with PAR. A summary of the data collected is given in Table 1 (enrollment numbers are for students who remained in the course after the W-drop date, which was approximately halfway through the semester). Any students who enrolled but did not take the first exam were removed from all analyses; taking the first exam was used as an indicator of a serious attempt at the course.

Exam Design and Logistics

A five-member team wrote all exams. After three rounds of revisions, the course coordinator compiled the final version of the exam. Exams were based on elaborated study guides (3–4 pages); students were given the study guides 2–3 weeks in advance and only problems that fell under the scope of the study guides were included on exams. This ensured that exams were unbiased towards particular sections.

Exams were designed to follow a standard template: one page of procedural computations, one page of true/false questions, and the rest of the exam was conceptual problem solving. Depending on the specific topics covered on the exam, other idiosyncratic problem types were included, such as curve sketching. This typology of problem types is described in Table 2.

Midterms had nine, mostly multi-part questions on average, and the final exam was slightly longer. Table 3 shows the breakdown of problems by type for the Phase I exams, which shows that the exams had a similar breakdown of problems.

Exams were administered in the evenings, each covering 3–4 weeks of material, except for the comprehensive final exam. Grading was blind, with each problem delegated to a single team of 2–3 graders, to ensure objectivity. Each team of graders designed their own grading rubrics, with approval from the course coordinator. These

Table 2 Exam problem types

Problem type	Description
Problem solving	Non-rote mathematical problems. Over 80 % had multiple parts and required written explanations.
True/false	Students must explain why it is true, or provide a counterexample and explain why it is false.
Pure computation	Procedural practice of limits, derivatives, and integrals.
Miscellaneous	Multiple choice, fill-in-the-blank, and curve sketching.

rubrics followed standard department procedures for many types of problems, such as true/false or procedural computations. Students needed to show their work and explain their reasoning to receive full credit on any problem other than pure computations of limits, derivatives, and integrals, and multiple-choice questions (in contrast, true/false questions did require explanations). Partial credit was offered on all problems.

Exam Content

According to a US national study of calculus programs, "the vast majority of exam items (85.21 %) could be solved by simply retrieving rote knowledge from memory, or recalling and applying a procedure, requiring no understanding of an idea or why a procedure is valid" (Tallman et al. forthcoming). In contrast to typical exams, the exams used in the present study emphasized explanation and deeper problem solving. Two typical problems are given in Figs. 8 and 9.

Figure 8 is a typical problem-solving problem. Prompts (a)–(c) were a nontrivial calculus problem (maximization with the use of a parameter), and the final two prompts (d) and (e) required students to explain and justify their work. Figure 9 provides two sample prompts from a true/false problem. On average, each exam included four such prompts.

For true/false questions, students were required to provide an explanation justifying their answer. Even if they gave a counterexample to show the statement was false, they needed to provide a written justification of their counterexample to receive full credit. Problem solving and true/false problems (shown above) all required written explanations, and comprised about two-thirds of each exam (see Table 3). The only problems from exams in the present study that would accurately be classified as procedural recall

Table 3 Percentage of points by problem type (phase I)

Type	Exam 1	Exam 2	Exam 3	Final exam
Problem solving	50 %	50 %	47 %	53 %
True/false	20 %	15 %	15 %	20 %
Pure computation	25 %	19 %	19 %	27 %
Miscellaneous	15 %	16 %	19 %	–
Total points possible	99	104	104	150

A rectangle is inscribed with its base on the x-axis and its upper corners on the parabola $f(x) = 3P - x^2$ (where $P > 3$). Complete the following parts of the question to find the dimensions of such a rectangle with the greatest possible area.
a) Draw a diagram of the situation, labeling the variable dimensions of the rectangle.
b) Write an equation for a function A that expresses the area of the rectangle in terms of the variable(s) defined in (a). What is the domain of this function?
c) Find the dimensions of the inscribed rectangle that will maximize its area.
d) Explain how you know that the value you found in (c) is the maximum area.
e) Describe what happens to the shape of the inscribed rectangle as P increases. (Relate your answer to how both the length and the width of the rectangle depend on P.)

Fig. 8 Sample problem solving problem

are the "Pure Computation" problems, which comprised less than 30 % of any given exam. Although miscellaneous questions did not require explanations, none of them could be solved through simple recall.

Results

Success in the course was defined as receiving an A, B, or C (the grade requirement for math-intensive majors like engineering), compared to receiving a D, F, or W (withdrawal with a W on the transcript, which is not calculated into one's GPA). I computed success rates using student course grades, 70 % of which was based on exam scores and 30 % on homework and calculator lab scores. The course coordinator scaled homework and lab scores to ensure consistency across sections. The experimental section had an 82 % success rate, which was 13 % higher than the 69 % success rate in the comparison sections. This effect was marginally significant, $\chi^2(1, N=409)=3.4247$, $p=0.064$. This improvement is comparable to other active learning interventions in STEM, which result in a 12 % improvement in passage rates on average (Freeman et al. 2014). Moreover, students in the experimental section were more likely to persist in the course; the experimental drop rate was only 1.75 %, while the drop rate for non-experimental sections was 5.87 %.

To account for the nesting of students within classes, I created a two-level random effects (HLM) model using the lme4 package in R (see Table 4). The null model included class section and exam number as second-level variables. In the alternative model, the use of PAR was added as a fixed effect. Only the final three exams were included, because, as I describe below, Exam 1 was used as a proxy for a pretest score. I used the anova package in R to compare the two models, $\chi^2(1)=3.9635$, $p=0.0465$*, which were significantly different. This indicates that PAR had a significant impact on student exam scores. Michelle's comparison section ($M=66.84$ %, $SD=23.47$) performed numerically similar to the rest of the comparison sections ($M=67.32$ %,

Indicate whether each of the following statements is True or False. If the statement is true, explain how you know it is true. If it is false, give a counterexample. (A counterexample is an example that shows the statement is false.)
a) If $\int_0^1 f(x)dx = 4$, then $\int_0^{\frac{1}{2}} f(x)dx = 2$
c) If $f'(x) > 0$ for $x < -3$ and $f'(x) < 0$ for $x > -3$, there must be a local maximum at $x=-3$.

Fig. 9 Sample true/false problem

Table 4 Comparison of nested models for phase I exam scores

Parameter	Model 1	Model 2
Fixed effects [estimate (SE), *t*-value]		
Intercept	67.583 (3.689), 18.2	66.866 (3.60), 18.57
PAR intervention		6.972 (3.06), 2.282
Random effects [Variance (SD)]		
Section	11.84 (3.44)	6.02 (2.45)
Exam number	36.19 (6.02)	35.65 (5.97)
Residual	361.48 (19.01)	361.60 (19.02)
Overall model tests		
AIC	10061.1	10059.1
BIC	10081.3	10084.4
Deviance	10053.1	10049.1

SD=19.37), while Michelle's experimental section performed considerably better (*M*= 73.03 %, *SD*=18.37). Given the small sample size of Michelle's comparison section, I combined all of the comparison sections for the remaining analyses.

Table 5 shows the results for each individual exam. To account for the nesting of students within classes, I used the intracluster correlation (ICC) to compute a variance inflation factor (Biswas et al. 2007, p. 79). For each exam, the ICC within sections was small, near 0.01, so *t* values were divided by about 1.179. To test for the equivalence of groups, I used students' Exam 1 scores as a proxy for a pre-test, because Exam 1 occurred early in the semester. Nevertheless, it is likely that instruction in the PAR section still had an impact on students' Exam 1 scores, so Exam 1 should not be thought of as a true pre-test. Table 5 indicates that there were no significant differences in Exam 1 scores, and that all other differences were at least marginally significant (after adjustments). This indicates that the student populations in parallel sections were comparable, and that the intervention had a significant effect. The numerical (but not statistically significant) difference in Exam 1 scores is consistent with the interpretation that PAR instruction had some impact on student scores, but less impact because it was still early in the semester. Overall, the effect sizes were small to medium (Cohen 1988). To contextualize these results, I note that active learning interventions in STEM classrooms result in a 6 % improvement in exam scores, on average (Freeman et al. 2014).

To address understanding of different problem types (research question two), I used the typology of problem types given in Table 2. Student performance by problem type

Table 5 Phase I exam (percentage) scores (SD in parentheses; $p<0.01$**, $p<0.05$*, $p=0.06$†)

	Exam 1	Exam 2	Exam 3	Final exam
Experimental (*N*=56)	70.2 (17.4)	79.8 (14.8)	74.4 (17.6)	67.3 (21.3)
Comparison (*N*=353)	67.2 (17.8)	75.1 (15.6)	66.4 (19.2)	60.2 (22.7)
Difference	**3.0**	**4.8**†	**8.0****	**7.1**†
ES (Cohen's *d*)		0.27	0.38	0.27

(aggregated over all exams) is given in Table 6. As before, all *t*- and *p*-values are adjusted using the ICC correction factor.

Student performance by problem type was calculated as a percentage of the total possible points for each problem type, to account for the different number of points assigned to different problem types. Students in the experimental section scored numerically higher on all aspects of the exams, but differences in problem solving were not significant. This contrasts with prior studies on calculus reform that showed that students often fell behind on traditional procedural skills (Ganter 2001).

Discussion

Analyses of student exams addressed the first two research questions: (1) students in the experimental section had 13 % higher success rates (marginally significant) than the other sections, and (2) these improvements were evident throughout the exams, not just on explanation-focused problems (see Tables 4, 5 and 6). Although exams were analyzed by problem type, this analysis did not account for differences in item difficulty between items. Finally, using Exam 1 as a proxy for a pre-test score, I established that there were no significant differences between groups in baseline calculus understanding, so the effects found can likely be attributed to PAR.

Phase II (Spring 2013)

Materials and Methods

Phase II took place in a subsequent semester of the same calculus course. The same coordinator ran the course, and the curriculum and lecture-based teaching styles were the same as Phase I. In the experimental section, students engaged in PAR, just as in Phase I. There were three revisions to the Phase I design: (1) minor updates to the reflection and feedback forms, (2) the assignment of random partners (see the Randomization of Partners section), and (3) weekly training in analyzing work (see the Training of Students section).

Participants

Phase II once again had a single experimental section with 3 observed comparison sections. I taught the experimental section, to ensure full implementation of the design.

Table 6 Phase I Mean (percentage) scores (SD in parentheses) by problem type

	Experimental *(N= 56)*	Comparison *(N= 353)*	Difference	*t*	*p*
Problem solving	63.94 (14.4)	59.81 (15.2)	4.13	1.50	0.14
True/false	67.98 (16.4)	58.44 (17.2)	9.54**	3.05	0.003
Pure computation	70.05 (15.7)	61.07 (19.3)	8.98**	2.86	0.005
Miscellaneous	79.85 (12.5)	74.61 (17.0)	5.24*	2.027	0.045

I was a graduate student with approximately 3 years of teaching experience. I had not taught introductory calculus in the last 4 years. Comparison instructors were chosen who had prior experiences with teaching and with this particular course to provide a fair comparison; some of the other instructors had little teaching experience or had not taught this course before. Sam, a post-doctoral researcher working on mathematics education projects, taught one of the observed comparison sections. Graduate student instructors from Phase I, Tom and Bashir, taught the other two observed sections. These instructors and I had comparable experience with this course, but their teaching experiences were more recent. Grading procedures were the same as in Phase I.

Data Collection

The data collected are summarized in Table 7.

Data collection procedures were the same, except for a few minor changes: (1) one camera was used rather than two to reduce logistical difficulties, (2) a research assistant conducted interviews and performed video observations (to maintain objectivity), and (3) students were offered one extra credit homework assignment as an incentive to give an interview, which greatly increased the number of respondents. Also, I had one of the comparison instructors (Tom) assign PAR problems as regular homework, which I collected. Finally, I collected background demographic data for students in the four observed sections.

Results

There were no significant differences in academic background data (ACT scores and high school GPA) between the four observed sections. Numerically, the lowest averages were in the experimental section (mean GPA: 3.43 vs. 3.56, and mean ACT scores: 25.45 vs. 26.3). There were no significant differences in gender or race. The population of students who answered the survey consisted of 19 % females and 17 % traditionally underrepresented minorities. While demographics for all students were not collected, this sampling seemed to be relatively representative of the typical student population of calculus at this institution. Although students were not surveyed specifically, based on an analysis of student PAR conversations, none of the students appeared to have limited English proficiency or were English Language Learners. Although demographics were not collected for Phase I, these Phase II results suggest that the natural distribution of

Table 7 Phase II data collection summary

	Experimental	Comp. 1	Comp. 2	Comp. 3	Comp. (Other)	Total
Instructor	Dan	Sam	Tom	Bashir	–	–
Students	34	37	31	28	206	336
Participants	34	34	27	24	–	119
Video Obs.	54	6	5	5	–	75
Interviews	22	–	–	–	–	22
PAR Conv.	86	–	–	–	–	86

students across sections was relatively balanced (i.e. various sections are indeed comparable).

As in Phase I, student course grades were used to compute student success rates. Course grades consisted 70 % of exam scores, with the other 30 % assigned to labs and homework. Once again, the course coordinator scaled student homework and lab scores to achieve consistency between sections. During Phase II, the experimental success rate was 79 %, while the comparison success rate was only 56 %. This 23 % difference in success rates was even larger than the 13 % difference during Phase I. This result was a statistically significant, $\chi^2(1, N=336)=6.3529$, $p=0.0117$*. To contextualize these results, I note that active learning interventions in STEM result in a 12 % improvement in passage rates on average (Freeman et al. 2014). Moreover, students in the experimental section were more likely to persist in this course; the experimental drop rate was 10.5 %, while the drop rate for non-experimental sections was 15.25 %. These drop rates were much higher than during Phase I, likely due to differences in the students who enroll in the fall and spring versions of this course.

Differences between experimental and comparison sections were also evident in exam scores (see Table 8). Once again I used a variance inflation factor (mean ICC= 0.03) to adjust the t and p values.

As before, Exam 1 provides a baseline to further establish the equivalence of the experimental and comparison groups. Because there were no significant differences for Exam 1 (a proxy for a pre-test), but the differences were significant for the other three exams, the differences can likely be attributed to PAR. The improvements in exam performance (row 3) were even larger than in Phase I (row 4). The effect sizes were medium (Cohen 1988). To contextualize these results, I note that active learning interventions in STEM classrooms result in a 6 % improvement in exam scores, on average (Freeman et al. 2014). Once again, the smaller, non-significant differences in Exam 1 scores are likely an indicator of early benefits of PAR instruction.

As in Phase I, I created nested two-level random effects models to account for the nesting of students within classes (see Table 9). Using anova to compare the two models, I found that the PAR intervention had a significant effect, $\chi^2(1)=8.6565$, $p=0.00325$**. The average exam scores in the experimental section ($M=75.2$ %, $SD=18.6$) were much higher than in the comparison sections ($M=64.1$ %, $SD=21.1$)

Research question two was addressed by analyzing exams by problem type. Phase I results were generally replicated; students in the experimental section did better on all problem types, even purely computational (see Table 10). Differences for problem solving and miscellaneous problems were statistically significant, pure computation

Table 8 Phase II exam (percentage) scores (SD in parentheses; $p<0.05$*, $p<0.01$**)

	Exam 1	Exam 2	Exam 3	Final exam
Experimental (*N= 34)*	68.4 (21.3)	81.7 (16.1)	75.7 (16.8)	75.7 (17.5)
Comparison (*N*=302)	62.2 (19.1)	66.5 (19.4)	61.5 (22.4)	65.9 (23.1)
Difference (Phase II)	**6.2**	**15.2****	**14.2****	**9.8***
Effect Sizes (Cohen's *d*)		0.62	0.55	0.37
Difference (Phase I)	**3.0**	**4.8***	**8.1****	**7.1***

Table 9 Comparison of nested models for phase II exam scores

Parameter	Model 1	Model 2
Fixed effects [estimate (SE), *t*-value]		
Intercept	66.98 (2.09), 31.99	65.48 (1.66), 39.55
PAR intervention		12.28 (3.27), 3.75
Random effects [variance (SD)]		
Section	19.09 (4.37)	5.11 (2.26)
Exam number	5.10 (2.26)	4.49 (2.12)
Residual	429.67 (20.97)	439.21 (20.96)
Overall model tests		
AIC	8254.0	8247.3
BIC	8273.3	8271.5
Deviance	8246.0	8237.3

was marginally significant, and true/false was not significant. I report adjusted t and p values, accounting for intracluster correlation.

It is likely that the differences between the experimental and comparison sections are overstated for miscellaneous problems. There were a total of 450 points possible across all exams, with only 35 associated with miscellaneous problems. Because these problems made up such a small percentage of the exams, they are likely to be less reliable than categories such as problem solving, which made up approximately 50 % of the exams.

Discussion

Phase II provided a replication of Phase I's results; (1) students in the experimental section had 23 % higher success rates than other sections, and (2) they performed numerically better on all aspects of the common exams (gains for problem solving and miscellaneous were significant). Once again, item difficulty was not taken into consideration. During Phase I the impact of PAR was measured while controlling for teacher effects. Thus, it is unlikely that improvements during Phase II can be attributed entirely to teacher effects. Moreover, Phase II featured an improved version of the Phase I design, which likely accounts for at least some of the additional improvement. Phase II demonstrates that multiple teachers could use PAR successfully.

Table 10 Phase II mean (percentage) scores (SD in parentheses) by problem type

	Experimental *(N= 34)*	Comparison *(N= 302)*	Difference	*t*	*p*
Problem solving	64.51 (12.0)	52.62 (15.6)	11.89**	3.00	0.0042
True/false	59.15 (16.8)	52.25 (18.1)	6.90	1.28	0.21
Pure computation	71.72 (12.9)	64.02 (17.6)	7.70†	1.81	0.076
Miscellaneous	73.29 (14.3)	60.85 (20.7)	12.44*	2.61	0.012

Comparison of Phases I and II

Students in the experimental sections numerically outperformed the students in the comparison sections for all problem types. Nevertheless, there were notable differences between phases. During Phase I, the differences for true/false and pure computation problems were significant, while they were not during Phase II. Also, during Phase II the differences for problem solving problems were significant, while they were not significant during Phase I. These differences may be attributable to differences in teaching style across phases; Michelle was much more likely to use IRE-style questioning in her classroom, emphasizing procedural computations, while Dan was more likely to require open-ended explanations from the students. Moreover, analyses did not account for the difficulty of items on exams, which may also account for some of these differences.

The average success rate for comparison sections during Phase I was considerably higher than during Phase II (69 % vs. 56 %). This difference was reflected in average exam scores, which were 5–10 % higher on exams 1–3 comparing Phase I to Phase II students; notably, Phase II students scored higher on the final exam compared to the Phase I students, by 5.7 %. In the design of the Phase II final exam, the course coordinator noted that the previous semesters' exam was too difficult, and made efforts to decrease the length and difficulty level of the Phase II exam. The course coordinator also noted that students during spring semesters historically tend to have lower success rates than those in the fall, because they are generally students who were not on the "standard" track, meaning that they may have had to take additional remedial mathematics courses before they could take calculus.

Because Tom and Bashir both taught during Phase I and Phase II, the average scores from their sections also provide a point of comparison. Bashir's scores increased between phases (63.7 to 69.4 %), while Tom's remained mostly the same (65.5 to 65.8 %). Given the differences in student populations, this seems to indicate that both instructors improved in their teaching across semesters, but without knowing more about their specific classes no more definitive conclusions can be drawn. The next major section describes student revisions, and the section following that describes the PAR mechanisms that supported learning.

Improvement in Student Explanations

PAR was designed to improve student understanding generally, and student explanations specifically. While in-depth analyses of student explanations are beyond the scope of this paper, they are discussed in a forthcoming paper (Reinholz forthcominga). To contextualize student improvements on exams, I provide a brief summary of those results.

Student explanations were analyzed on three of the PAR problems (PAR 5, 10, and 14) to see the progression of student explanations over the course of the semester. Student work was analyzed from the Phase I experimental section, the Phase II experimental section, and a comparison section from Phase II. Student explanations were scored using a rubric consisting of four categories: accuracy, mathematical language, clarity, and use of diagrams. Solutions were double coded, with 94.1 % agreement.

Aggregating explanation scores across the semester, the Phase II section scored more than 4.5 times higher than the comparison section, and more than 1.5 times higher

than the Phase I section. The results held across individual dimensions as well, with Phase I scoring higher than the comparison section on all aspects of their explanations and Phase II scoring higher than Phase I on all aspects of their explanations. As this brief summary of some of the results highlights, students improved their explanations considerably as a result of PAR.

Students Revisions in PAR

The quantitative results from Phases I and II provided evidence of the positive impact of PAR on student performance in calculus. To better understand how students learned from PAR, I analyzed PAR assignments in the Phase II experimental section to look for changes in PAR scores as a result of revision.

Materials and Methods

To understand the impact of PAR for different students in the course, I broke the class into thirds (High, Middle, and Low), according to students' final scores on the PAR assignments. I used a random number generator to select three students from each of these groups. Of these nine students who were randomly sampled, there were four cases in which I had recorded a score for their final solution, but did not have a scan of the student's PAR packet. These were students who turned in their assignment separately from the rest of the class, and as a result some assignments did not get scanned. I dropped these four solutions from the analysis. I had a total of 122 PAR packets to analyze, each with an initial and final solution.

To measure the impact of PAR on student solutions, I blindly re-scored each student's initial and final solutions. Although I did not conduct double scoring to establish inter-rater reliability, the purpose of this analysis was to investigate changes in scores, so any systematic biases in scoring should be present in both the scoring of initial and final solutions.

Results

The sampled students turned in all of their PAR homework assignments, except for two students in the Low group didn't turn in PAR14 (the final problem). This was a 98 % completion rate for PAR homework assignments. In contrast, the comparison section had only a 70.2 % completion rate for the same problems.

The distribution of initial scores is given in Fig. 10. Students in the Low and Middle groups had relatively similar distributions of scores; neither group achieved any scores of 9–10, and most of the scores were distributed from 0 to 5 points. In contrast, students in the High group had a very different distribution of scores: some students achieved scores of 9–10, and the majority of initial scores were distributed between 4 and 8.

Figure 7 provides the distribution of final scores on the PAR problems. Of the three groups, only the High group achieved a considerable amount of 10 scores; there were no scores of 10 in the low group. In the Middle group there were a number of 8 and 9 scores, with some 8–9 scores in the Low group as well. A comparison of Figs. 10 and 11 shows a considerable shift in the distribution of PAR scores after revision.

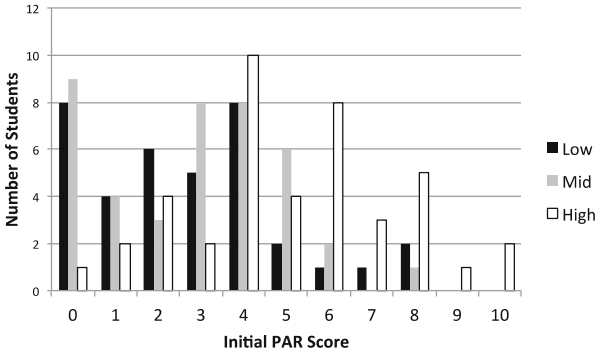

Fig. 10 Distribution of initial PAR scores

Of the solutions in the experimental section that were turned in, there were only three instances in which the students did not revise their work from initial to final solution. It was a single student in the Middle group who did not revise twice, and a student in the Low group who did not revise once. Students in all groups significantly improved their solutions after revision, as evidenced by Table 11.

The data in Table 11 suggest that students in the Middle group benefitted most from PAR. To better understand these results, I looked at the distribution of revision scores. I grouped the amount of change for each group into four categories: no change, 1–3 points change, 4–7 points change, and 8–10 points change (see Fig. 12).

Figure 12 shows that students in the High group were most likely to make small or medium improvements to their solutions. Students in the Middle group were the only group of students that consistently made large improvements to their solutions. Students in the Low group were least likely to benefit, making small gains more often than the other students. To better understand the 12 students who improved their scores by 8–10 points, I analyzed their PAR conversations and consulted my daily field notes. I identified two potential explanations for such drastic improvements: PAR conversations and office hours. The results are summarized in Table 12. The PAR conversation was considered as a potential source of improvement: (1) if a student's partner correctly solved the aspect(s) of the problem that the student was struggling with, or (2) the major errors were discussed in the conversation or written feedback. I considered office hours

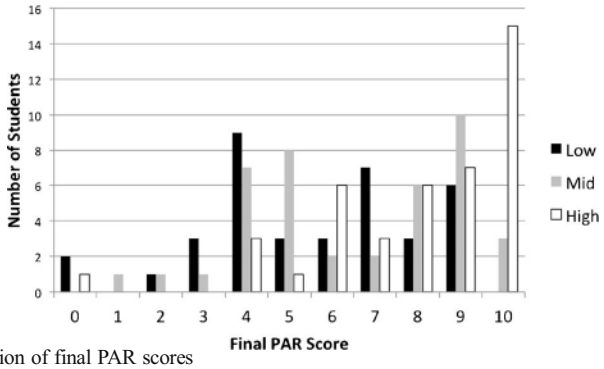

Fig. 11 Distribution of final PAR scores

Table 11 Average PAR scores, by group

	Initial score	Final score	Change	*t*	*p*
High	5.09	7.98	2.89**	8.95	$3.4 \cdot 10^{-11}$
Middle	2.82	6.55	3.73**	7.73	$1.8 \cdot 10^{-9}$
Low	2.78	5.60	2.82**	6.31	$2.6 \cdot 10^{-7}$

as the source if I had documented working with the particular student on that problem during office hours. Without observations of students' revision processes I do not have sufficient data to claim that these processes caused improvement, but the results suggest that PAR may support meaningful revisions in these ways.

Discussion

All students made significant improvements to their homework solutions as they revised from initial to final solutions. Students in the Low and Middle groups had similar distributions of initial PAR scores. However, students in the Middle group were much more likely to considerably improve their solutions after revision. These data suggest that one of the key differences between students who scored in the Low and Middle groups may be how they benefited from PAR and their revisions. Table 12 indicates that when students made considerable improvements in their revisions it was mostly due to their PAR conversations and additional time spent working on the problem after their conferences.

The help-seeking literature suggests that students with moderate need are the most likely to seek help (Karabenick and Knapp 1988). This is consistent with the finding that students in the middle group were the most likely to improve as a result of seeking external help. However, the students in the low and middle groups had relatively similar initial scores, so it is unclear what factors may have caused some of them to seek help while others did not. In general, low-performing students may be less likely to seek help due to low self-efficacy or negative emotions related to failure (Karabenick and Knapp 1988), which may have been factors at work here. This is an area for future research.

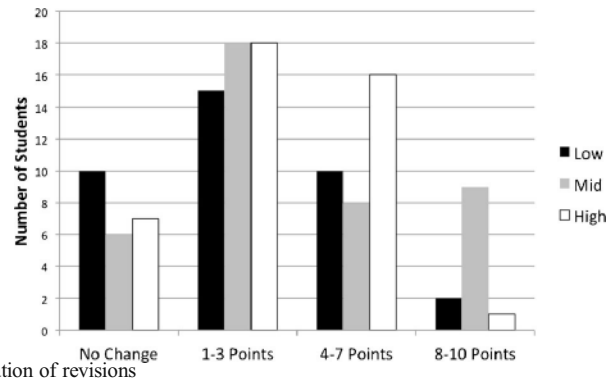

Fig. 12 Distribution of revisions

Table 12 Potential sources of considerable improvement on PAR problems

Source	Office hours	PAR conversations	Undetermined
Number of students	5	5	2

PAR Mechanisms

Materials and Methods

To understand student experiences with PAR, I analyzed interviews from Phase II. I focused on Phase II data because there was a much higher response rate than Phase I, which meant the interviews were more likely to represent a range of opinions. The following analyses focus on the first interview question that was asked: "Let's discuss the PAR; what's working well and not so well for you?" I focus on this question, because it was likely to elicit a balance of positive and negative aspects of PAR.

After transcribing student responses, I read through all of the transcripts multiple times to identify themes. After a set of themes was identified, I developed codes, both for positive and negative reactions. Using this set of codes I re-analyzed each student response and marked whether or not each code was present.

Results

Of the codes that were developed, six appeared most frequently in the data. Codes were only included if they described responses from at least two students. These codes and their frequency in the data are summarized in Table 13. Given the frequency of positive and negative reactions, students were generally quite positive towards PAR.

The positive student reactions described four mechanisms of PAR that appeared to support learning. PAR required students to work in iterative cycles: students made a preliminary attempt at a problem, received feedback and thought about the problem more deeply, revised, and turned in their final solution. Within these iterative cycles, students encountered new ideas to support their learning: by discussing with peers, by explaining and hearing explanations, and by seeing the work of others. These four sources can be consolidated into the acronym IDEA, meaning **Iteration**, **Discussion**, **Explanation**, exposure to **Alternatives**. I now discuss these mechanisms. The student quotes given below are intended to exemplify each category of student responses.

Table 13 Frequency of student reactions to PAR (N=22 interviews)

Positive reactions				Negative reactions	
Iteration	Discussion	Explanation	Alternatives	Workload	Both confused
6	13	9	11	2	8

Iteration

Rather than viewing homework as something that is attempted once, turned in, and forgotten about, PAR forced students to revisit their work. As a result, students seemed to view their first draft of the problem as a work in progress, and didn't expect it to be correct. Six students noted this in their interviews. As Barry said,

> I like the PAR. It's like we get to come to class and be wrong, and that's okay. Then later we get to revise our work and be right.

Mike made a similar remark,

> PAR is good. I like how we can put our initial solution down, and even if it's wrong it doesn't really matter, because we can just talk about it with a group member the next day, and figure it out together. And generally you don't get stuck on a wrong solution, you figure it out.

This activity structure seemed to increase students' perseverance. Rather than giving up when they could not solve a problem on their own, students seemed to realize that getting input from peers, the instructor, or other resources was often sufficient to help them solve challenging problems. This perseverance was evident in students' homework assignments; the number of students who fully completed the challenging PAR problems in the experimental section was much higher than in the comparison section (98 % vs. 70.2 %).

Discussing the Problem Together

Most students appreciated the opportunity to collaborate with their peers. More than half of the students (13 of 22) noted this. As Tom said,

> I like the PAR because it got you to interact and communicate with the other students…no one likes to just watch someone talk at a board all day. The self teaching and student interaction helped…PAR helps us be more social, so you can talk to other students, set up study groups, and get to know your classmates.

The value of student discussions was exemplified by the first example provided in this paper, in which Peter and Lance revised their methods for approximating error.

Explanation

Many students came to appreciate the importance of explaining their ideas. Nine students made mention of this. As Barry noted in his interview,

> I like the PAR. I'd take it over the other homework. You do about the same amount of work, and I think you learn more from it. You have do explain what you did, rather than just say here, I got this magical number. You actually understand the process and I think that helps more in learning than just getting the magical number.

As Maria said, PAR helped her learn:

> how to make it easier to read from another person's perspective. 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 it helps me figure out how to communicate better. It helps me to explain things in a way that is readable to others and not just myself.

Exposure to Alternatives

Students sometimes recognized errors in their own solutions simply by looking at one another's work. Half of the students (11 of 22) noted this. As Harry said,

> I really like looking at other people's initial models. I can see what they are thinking, it puts me in their head, and I can see that. A lot of times I'm really wrong and I can see different ways to do the same thing.

This was also evident in students' conversations. Consider the following excerpt from PAR7. As soon as the students finished silently reading one another's work, Revati exclaimed that she saw her errors,

> [1] Revati: I know I did it all wrong. I was reading yours and was like, "oh my goodness. How did I miss this?" Okay, so. You did a really good job explaining, so you have all that right. And your math is all correct so… good job! You could have turned this in as your final and gotten 100 %
>
> [2] Federico: Okay, thank you. Em, well I think now you know the errors?

PAR provided students with opportunities to analyze, explain, and discuss the work of their classmates. These opportunities seemed to help students make mathematical connections and develop deeper understandings of the problems. Students in comparison sections rarely had opportunities during class sessions to explore their peers' reasoning in depth.

Negative Reactions to PAR

Although students were generally positive towards PAR, there were areas that students felt could be improved. The most common criticism of PAR was that sometimes both partners would get stuck and that made it hard to make progress on the problem (8 of 22 students noted this). As Michelle said,

> If it's a confusing problem we just get together and talk about how neither of us know what is going on or we have no idea how to do it.

The other main issue that students had (only two students noted) was the amount of work required by the course. Given the large number of assignments they had, PAR felt

like it was too much on top of an already overloaded course. This criticism was not of PAR specifically, but of the organization of the course.

Conclusion

This paper focused on how reflection tools could promote improved understanding of calculus. Through cycles of problem solving, reflection, feedback and analysis, and revision, students had opportunities to exercise disciplinary authority and were held accountable to their peers (Engle and Conant 2002). PAR was supported by training exercises that helped students learn to analyze work and provide feedback. The PAR activities were conducted using a rich problem set, which provided opportunities for students to explain their ideas and compare multiple solutions with one another. Although these problems seemed to be an important part of the intervention, by themselves they were insufficient; these problems were also assigned in Tom's experimental section during Phase II with little impact.

Students in the PAR sections were given some opportunities to explain their ideas during class and engage in group work to support PAR. Although in-depth analyses of classroom activities are beyond the scope of this paper, similar activities were observed in some of the comparison classrooms as well (e.g., Heather's and Sam's sections). Accordingly, it seems reasonable to assert that the standard classroom activities in the PAR sections were not considerably different from the comparison sections; PAR was implemented in a primarily lecture-based environment, which was typical of calculus instruction at this institution.

Success rates in the experimental sections were higher than the comparison sections, 13 % in Phase I (marginally significant), and 23 % in Phase II (statistically significant). This demonstrated the impact of PAR on student success (research question one). These are impressive gains, showing that the impact of PAR compares favorably with other active learning interventions in STEM (Freeman et al. 2014). Moreover, these gains were replicated over two semesters. These gains are important, because student success in calculus remains an area of concern. The persistence rates were also higher in the experimental sections during both phases of study; it is possible that the community-building aspects of PAR may have made students less likely to drop the course.

Improvements were also apparent on exam scores during Phase I (experimental vs. comparison, same instructor: 6.19 %; and experimental vs. comparison, other instructors: 5.71 %) and Phase II (experimental vs. comparison: 11 %). Students improved numerically on all aspects of their exams, both explanations and procedures (research question two). These are considerable differences, especially given that the experimental sections included more students who would traditionally drop out of the course. No significant differences were apparent on Exam 1, which provides a proxy for a baseline pre-test score to establish the comparability of students in different sections. A companion paper (Reinholz forthcominga) focuses more directly on student explanations, and provides results consistent with the present findings.

During Phase I, Michelle taught two sections to control for teacher effects. Michelle's comparison section performed similarly to the other comparison sections, which suggests that improvements can be attributed to PAR, not the teacher. During Phase II, teacher effects were not controlled for specifically. As a result, it is possible

than some improvements may be attributed to the particular instructor, but given the impact of PAR during Phase I, and the improvements to the design for Phase II, it is unlikely that improvements can be attributed entirely to teacher effects. A goal of future studies would be to further replicate these results through a randomized experimental design.

The present study also makes an important contribution to literature on assessment for learning. Despite a large body of work on peer assessment, most of it has focused on calibration between instructor and peer grades, not how assessment can be used to promote learning. PAR demonstrates the effectiveness of such techniques, particularly in a content area where such practices are uncommon. Moreover, the iterative nature of PAR seemed to help students develop the perseverance required to solve challenging problems. The impact of PAR on student dispositions is an area for further study.

Design-based revisions provided greater affordances to support student learning (research question three). In particular, students worked with random partners, had regular training opportunities, and used streamlined self-reflection and peer-feedback forms. The underlying principles of having students analyze each other's work and provide feedback to each other appear to be productive activities that may work in a variety of contexts.

PAR was developed with two very different student populations in different contexts: primarily traditionally underrepresented minorities in a remedial algebra class and mostly White students in introductory college calculus. Since this initial study, PAR has been used in differential equations, introductory mechanics (physics), engineering statics, and thermodynamics. Given that PAR has been implemented in a variety of contexts, it appears to be a general method that could be effective across a broad variety of STEM learning contexts. To implement PAR and the corresponding training activities requires no more than 20 min of in-class time each week, which means that it is possible to include PAR as a part of a variety of different classrooms. As the impact of PAR is studied in new contexts, it will provide further insight into the activity structure and how students learn through peer analysis.

Acknowledgments The author thanks Elissa Sato, Hee-jeong Kim, and Bona Kang for their helpful feedback on an earlier version of this manuscript. The research reported here was supported in part by the Institute of Education Sciences pre-doctoral 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 of Education Sciences or the U.S. Department of Education.

Appendix A: Self-Reflection Form (Phase II of Study)

On a scale from 0% to 100%, how confident do you feel in your solution? _____

Completeness, Organization, and Labeling
Did you answer all questions asked, showing all steps, in the proper order? Yes____ No____

(If applicable) Did you label and explain all graphs, include units, etc.? Yes____ No____

Explanations
Did you explain why (not just what)? Yes____ No____

Use of Language
Did you avoid the use of pronouns (and other ambiguous language)? Yes____ No____

(If applicable) Did you consult definitions of mathematical terms you used? Yes____ No____

Justification
Did you justify your solution (in at least 1 of the following ways): Yes____ No____

- By checking if answers to different parts of the question are consistent?
- By explaining (in writing) how you know your solution is correct?
- In some other way? If so, how? _____

Note: Show explicitly on your solution *how* you justified your solution.

(Optional:) Is there anything in particular you'd like to discuss with your partner?

Appendix B: Peer-Feedback Form (Phase II of Study)

Communication: Give at least one suggestion to improve the communication of the solution. (Focus on explanations, imprecise use of language, organization, labeling, etc. Be specific: don't say "it was hard to follow" or "part 2 was unclear;" say *why* it was hard to follow, *what* was unclear, and *how* to improve it.)

Correctness: Note any errors you found. (Focus on misunderstanding of concepts, misuse of mathematical language, calculational errors, incomplete answers, etc. Be specific: don't just say "part 2 was wrong;" say exactly *what* is wrong, *why* it is wrong, and *how* to improve it.)

(Optional:) What other feedback to you have? How else could the solution be improved?

References

Andrade, H. L. (2010). Students as the definitive source of formative assessment: Academic self-assessment and the self-regulation of learning. In *NERA Conference Proceedings 2010*. Rocky Hill, Connecticut: Paper 25.

Barab, S., & Squire, K. (2004). Design-based research: putting a stake in the ground. *The Journal of the Learning Sciences, 13*(1), 1–14. doi:10.1207/s15327809jls1301_1.

Baxter, J. L., Majumdar, D., & Smith, S. D. (1998). Subsequent-grades assessment of traditional and reform calculus. *PRIMUS, 8*(4), 317–330.

Biswas, A., Datta, S., Fine, J. P., & Segal, M. R. (2007). *Statistical advances in the biomedical sciences: Clinical trials, epidemiology, survival analysis, and bioinformatics*. New York: Wiley.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7–74.

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

Bookman, J., & Friedman, C. P. (1999). The Evaluation of Project Calc at Duke University, 1989–1994. *MAA NOTES*, 253–256.

Boud, D., Keogh, R., & Walker, D. (1996). Promoting reflection in learning: A model. In *Boundaries of Adult Learning* (Vol. 1, pp. 32–56). Routledge.

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*(4), 685–698. doi:10.1080/0020739X.2013.798874.

Brown, A. (1992). Design experiments: theoretical and methodological challenges in creating complex interventions in classroom settings. *The Journal of the Learning Sciences, 2*(2), 141–178.

Burkhardt, H., & Schoenfeld, A. H. (2003). Improving educational research: toward a more useful, more influential, and better-funded enterprise. *Educational Researcher, 32*(9), 3–14. doi:10.3102/0013189X032009003.

Chi, M. T. H., De Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*(3), 439–477. doi:10.1016/0364-0213(94)90016-7.

Cobb, P., Confrey, J., Disessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher, 32*(1), 9–13. doi:10.3102/0013189X032001009.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. NY: Routledge.

Engle, R. A., & Conant, F. R. (2002). Guiding principles for fostering productive disciplinary engagement: explaining an emergent argument in a community of learners classroom. *Cognition and Instruction, 20*(4), 399–483.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: a meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287–322.

Featherstone, H., Crespo, S., Jilk, L. M., Oslund, J. A., Parks, A. N., & Wood, M. B. (2011). *Smarter together! Collaboration and equity in the elementary math classroom*. Reston, VA: National Council of Teachers of Mathematics.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 201319030. http://doi.org/10.1073/pnas.1319030111.

Fullilove, R. E., & Treisman, P. U. (1990). Mathematics achievement among African American undergraduates at the University of California, Berkeley: an evaluation of the mathematics workshop program. *The Journal of Negro Education, 59*(3), 463–478.

Ganter, S. L. (2001). Changing calculus: a report on evaluation efforts and national impact from 1988–1998. *AMC, 10*, 12.

Gutiérrez, K. D., & Penuel, W. R. (2014). Relevance to practice as a criterion for rigor. *Educational Researcher, 43*(1), 19–23.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. doi:10.3102/003465430298487.

Karabenick, S. A., & Knapp, J. R. (1988). Help seeking and the need for academic assistance. *Journal of Educational Psychology, 80*(3), 406–408. doi:10.1037/0022-0663.80.3.406.

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development* (Vol. 1). Upper Saddle River: Prentice-Hall.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences, 10*(10), 464–470. doi:10.1016/j.tics.2006.08.004.

Min, H. (2006). The effects of trained peer review on EFL students' revision types and writing quality. *Journal of Second Language Writing, 15*(2), 118–141.

National Center for Education and Statistics. (2010). *The nation's report card: Grade 12 reading and mathematics 2009 National and pilot state results (No. NCES 2011–455)*. Institute of Education Sciences, U.S. Department of Education: Washington, DC.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common core state standards for mathematics*. Washington, DC: Authors.

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

Niss, M. (2003). Mathematical competencies and the learning of mathematics: The Danish KOM project. In *3rd Mediterranean conference on mathematical education* (pp. 115–124).

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. 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.

Parker, J. D. A., Summerfeldt, L. J., Hogan, M. J., & Majeski, S. A. (2004). Emotional intelligence and academic success: examining the transition from high school to university. *Personality and Individual Differences, 36*(1), 163–172.

Rasmussen, C., Marrongelle, K., & Borba, M. C. (2014). Research on calculus: what do we know and where do we need to go? *ZDM – The International Journal on Mathematics Education, 46*(4), 507–515.

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

Reinholz, D. L. (forthcoming a). Using peer-review to improve undergraduate calculus explanations: A design-based approach.

Reinholz, D. L. (forthcoming b). Design bridges: Supporting inferences across multiple levels of design-based research.

Roddick, C. D. (2001). Differences in learning outcomes: calculus & mathematica vs. traditional calculus. *PRIMUS, 11*(2), 161–184.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144. doi:10.1007/BF00117714.

Sadler, P. M., & Good, E. (2006). The impact of self-and peer-grading on student learning. *Educational Assessment, 11*(1), 1–31.

Schoenfeld, A. H. (1988). When good teaching leads to bad results: the disasters of "well-taught" mathematics courses. *Educational Psychologist, 23*(2), 145–166.

Schoenfeld, A. H. (1991). What's all the fuss about problem solving. *ZDM – The International Journal on Mathematics Education, 91*(1), 4–8.

Schoenfeld, A. H. (1995). A brief biography of calculus reform. *UME Trends: News and Reports on Undergraduate Mathematics Education, 6*(6), 3–5.

Schwingendorf, K. E., McCabe, G. P., & Kuhn, J. (2000). A longitudinal study of the C4L calculus reform program: comparisons of C4L and traditional students. *CBMS Issues in Mathematics Education, 8*, 63–76.

Sfard, A. (1998). On two metaphors for learning and the dangers of choosing just one. *Educational Researcher, 27*(2), 4–13.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. doi:10.3102/0034654307313795.

Stefani, L. A. J. (1998). Assessment in partnership with learners. *Assessment & Evaluation in Higher Education, 23*(4), 339–350.

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

Tallman, M., Carlson, M., Bressoud, D. M., & Pearson, M. (forthcoming). A characterization of Calculus I final exams in U.S. colleges and universities.

Treisman, U. (1992). Studying students studying calculus: a look at the lives of minority mathematics students in college. *The College Mathematics Journal, 23*(5), 362–372.