JOURNAL OF RESEARCH ON EDUCATIONAL EFFECTIVENESS
https://doi.org/10.1080/19345747.2023.2287594

Routledge
Taylor & Francis Group

INTERVENTION, EVALUATION, AND POLICY STUDIES

OPEN ACCESS Check for updates

The Impacts of a Standards-Based Grading System Emphasizing Formative Assessment, Feedback, and Re-Assessment: A Mixed Methods, Cluster Randomized Control Trial in Ninth Grade Mathematics Classrooms

Steven L. Kramer[a], Michael A. Posner[b], Alexander S. Browman[c], Nancy R. Lawrence[d], Jennifer Roem[e] and Kathleen Krier[a]

[a]Research, The 21st Century Partnership for STEM Education, Wayne, PA, USA; [b]Department of Mathematics and Statistics, Villanova University, Villanova, PA, USA; [c]Department of Psychology, College of the Holy Cross, Worcester, MA, USA; [d]Human Development and Family Studies, Colorado State University, Fort Collins, CO, USA; [e]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

ABSTRACT

We investigated the impact in ninth-grade mathematics classrooms of Proficiency-based Assessment and Reassessment of Learning Outcomes (PARLO), a standards-based grading system. Key components of PARLO are basing student final grades on the number of learning outcomes on which the student is high-performance or proficient, providing students with formative feedback, and encouraging students to reassess for full credit after further study. Our mixed-methods study employed a cluster randomized control trial with 2,736 participating ninth graders at 14 Treatment and 15 Control schools. Data included student achievement tests, interviews with 35 teachers, and student surveys. The program improved student performance on end-of-course algebra and geometry tests by a statistically significant 0.33 SD but did not impact students' value of or expectancies for success in mathematics. However, treatment effects on mathematics performance were moderated by these psychological antecedents of motivation, such that students with higher math expectancies and value benefited more from the treatment. Furthermore, teacher interviews suggested that PARLO may have also had positive effects on growth mindsets, mastery goals, autonomy, and relatedness.

ARTICLE HISTORY
Received 22 February 2022
Revised 25 October 2023
Accepted 27 October 2023

KEYWORDS
Standards-based grading; high school mathematics; growth mindsets; expectancy value theory; self-determination theory

CONTACT Steven L. Kramer skramer@21pstem.org Research, The 21st Century Partnership for STEM Education, Wayne, PA, USA.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/19345747.2023.2287594.

Introduction

Among the most central and deeply rooted features of most American classrooms is a grading system that is designed to rank, sort, and evaluate students as *learners*, but is not optimally designed to *help students learn*. In the late 19th and early 20th centuries, many educators sought to build a "meritocratic" system that would give students access to a common elementary education, with students encouraged to persist according to their interests and abilities. Specifically, the system was designed so that more naturally talented and quicker learning students would receive higher grades than those with less natural talent or who learned more slowly, and only those with high grades would be encouraged or even permitted to progress to the next level of the education system (Farrington & Small, 2008; Schneider & Hutt, 2014).

Thus, the grading system that is still typical in U.S. classrooms today was adopted at a time when a primary goal of public education was to sort students based on their supposed inherent intellectual abilities. As a result, the grading system does not optimally support the theoretical goal of the modern education system: learning. Specifically, classroom grading is built around *summative assessments* that typically only provide students with a limited amount of time to learn a topic and a limited number of opportunities to demonstrate how much they have learned. After assessment, students are often not provided with additional opportunities to improve their skills or to demonstrate that they have increased their learning. Further, early assessments are typically averaged with later assessments. This can discourage persistence in the face of initial difficulty because no matter how well students eventually learn the material they will be evaluated partly based on their initial difficulties.

A Proposed Improvement: Adding Formative Assessment

To better support student learning, educators in the 1960s began recommending formative assessment as a supplement to summative assessment (Bloom, 1968). In contrast to the traditional system, where assessments are used solely to provide a final, summative judgment of student performance, formative assessment uses assessments intermittently to provide teachers and students with feedback about each student's progress. This enables teachers and students to adapt their teaching and learning strategies to help each student progress based on their individual needs. Numerous meta-analyses have shown that formative assessment has significant positive implications for student engagement, learning, and performance, especially for previously low-performing students (e.g., Black & Wiliam, 2010; Kingston & Nash, 2011).

To date, most formative assessment programs that have been implemented, especially in mathematics, have been teacher-facing. This approach to formative assessment provides teachers with feedback about their students' current level of proficiency, allowing the teachers to adapt their teaching strategies to meet the individual needs of their students (e.g., Supovitz et al., 2018). However, formative assessment can also be student-facing, wherein assessments are used to provide students with feedback about their own current level of proficiency and the nature of their mistakes so they can adapt their learning and become more proficient.

Numerous educators have called for more widespread implementation of student-facing formative assessment (e.g., Leahy et al., 2005; Wiliam, 2011). Mills and Silver (2018, p. 180) articulated the argument for mathematics: "Teachers on their own cannot make students learn mathematics; students must become partners in their own learning in order for effective teaching and learning to occur." Supporting student agency by incorporating student-facing formative assessment may be especially efficacious for middle and high school students, an age when students are exploring their independence and have a particular need for status and respect (Yeager, 2017; Yeager et al., 2017).

Nonetheless, prior research on student-facing formative assessment in math class has been limited, mostly focusing on improving the feedback that students receive about their work. For example, Murphy et al. (2020) reported results of a cluster randomized control trial (RCT) that evaluated "Assistments," an online program that provides students immediate feedback on the correctness of homework problems. "Assistments" increased the mathematics achievement of seventh graders by a statistically significant 0.2 standard deviations (SD). Programs like "Assistments," are philosophically compatible with and potentially complementary to, but differ from, the program we investigated, which focused on reengineering the grading system to better support student learning.

Standards-Based Grading: The PARLO Assessment System

The present work examines the effects of the *Proficiency-based Assessment and Re-assessment of Learning Outcomes (PARLO)* system. The system was developed by a partnership between Dylan Wiliam, a pioneer researcher in formative assessment, and the Math Science Partnership of Greater Philadelphia (MSP-GP), an NSF-funded project that brought together 13 institutions of higher education and 46 Pennsylvania and New Jersey school districts to improve grades 6-12 mathematics and science education. It was partly inspired by a program at the Young Women's Leadership Charter School (YWLCS), a public charter high school in Chicago that changed its grading system to allow students to reassess for full credit whenever they had mastered classroom content. The goal was to promote students' agency by fostering a sense of partnership with their teachers in ensuring that they mastered the big ideas from each course. As a result of these changes, YWLCS regularly achieved the highest graduation rate of any nonselective public school in Chicago, despite serving a similar student body as neighboring schools (mostly low-income Black and Latina students; Farrington & Small, 2008).

The PARLO assessment system is designed to be implementable in a single teacher's classroom, embedded in a school that may have a traditional American grading system. Consequently, PARLO teachers do assign final summative grades. However, PARLO reengineers the classroom assessment system by better integrating summative assessment and student-facing formative assessment. Specifically, while students receive summative scores on quizzes and assignments, such scores are not recorded indelibly to be used in a weighted average that determines final grades. Rather, the teacher uses such assessment evidence both to evaluate the extent to which a student is proficient in each of the course's learning outcomes *at that particular moment in time*, and then to provide students with personalized feedback designed to guide further learning. Students are then given opportunities to do further work, at home or in school, and to be reassessed for full credit. In other words, summative assessments become formative tools designed to promote further learning, instead of merely yardsticks to measure the student's efficiency as a learner.

The first step in implementing PARLO is clarifying and sharing learning intentions and success criteria (Wiliam, 2011). To do this, the teacher organizes their course instruction around 10-15 learning outcomes per semester, which together define the material to be mastered. These learning outcomes are then shared with students and their parents so that they can be partners in the student's educational progress. An important characteristic of PARLO that distinguishes it from similar programs such as Mastery Learning (Bloom, 1968) is the way it defines success criteria. In addition to establishing criteria for "proficient" performance, teachers also share criteria for "high performance." Achieving high performance on a learning outcome is intended to challenge talented students, while potentially being achievable through hard work by nearly all students. In practice, teachers in the current study used Webb's Depth of Knowledge (Webb, 2002) to define success criteria, with high-performance requiring students either to solve problems at a greater depth of knowledge (application or strategic thinking), or else to tutor a fellow student and successfully bring them from not-yet-proficient up to proficient.

The second step in implementing PARLO employs the central idea of student-facing formative assessment: eliciting evidence of learner's achievement and providing feedback that moves learning forward (Wiliam, 2011). PARLO teachers use short quizzes, end-of-lesson written "exit tickets," notes from observing students working in groups, and other formative assessment techniques to provide students feedback about how well they are progressing toward proficient or high performance on each learning outcome. Note that other student-facing formative assessment programs like Assistments (Murphy et al., 2020) implement this step. In future iterations of PARLO, teachers might use programs like Assistments to support PARLO implementation.

The remaining two steps in implementing PARLO focus on changing the way teachers assign summative grades. As will be explained in more detail below, the changes to summative grading were the key features differentiating the Treatment from the Control schools in the current study. These two steps are:

1. Reassessment for full credit after further learning. Student grades are not averaged over the semester; instead, students are rated on each learning outcome as not-yet-proficient, proficient, or high-performance based on the best work they can show by the end of the semester. While students can reassess for full credit, before reassessment they must first engage in further learning, completing activities such as:
   a. Error logs, wherein students explain what they did wrong on a problem, rework it, and describe what they would need to remember so as not to make the same mistake again.
   b. Remediation plans, or contracts with students about activities to be undertaken before reassessment.

   c. Flashback days, or in-class opportunities for students to work individually or together to revisit learning outcomes and learn material at proficient or high-performance level.
2. Final Grades Based on Learning Outcomes. With project help, each school developed its own algorithm to determine a letter grade based on the number of Learning Outcomes scored Proficient and the number scored High Performance. Other factors sometimes used to assign grades, such as attendance, attitude, and homework completion, are not averaged as part of the final grade. Instead, they are viewed as means to the end of understanding course content.

At the time that we designed the intervention, a grading system with components like the four we described above was commonly known as "proficiency-based"—hence the name "PARLO." However, terminology has since shifted, and PARLO-type grading systems have become known under the name *standards-based grading* (see, e.g., Marzano, 2010).

Theoretical Framework

We hypothesized that the PARLO system would positively impact student learning in two ways: by providing *additional time and opportunity for students to learn*; and by *encouraging students' motivation/desire to learn*.

Opportunity to Learn

As noted above, traditional summative assessment systems encourage teachers to provide a limited amount of time for students to learn a topic and a limited number of opportunities to demonstrate how much they have learned. In contrast, PARLO teachers provide students with more opportunities to learn from their mistakes, improve, and be reassessed to demonstrate their increased proficiency. Further, student persistence is encouraged by giving them full credit for material learned, regardless of initial difficulties they may have had.

Motivation

According to expectancy-value theory (Wigfield & Eccles, 2000), students' motivation and classroom engagement are largely determined by the degrees to which they both *value* the material being taught and *expect to succeed* in class. Yet, studies have found that students exposed to an assessment system designed to compare and rank students report weaker feelings of intrinsic value (enjoyment of learning for its own sake), utility value (the belief that learning is useful and will have purpose and relevance in one's life), and expectancies for success than those exposed to a system that provided students with opportunities to improve their skills or to demonstrate that they have increased their learning (Covington & Omelich, 1984; Haley, 2015; Sanchez et al., 2017), and that these effects may be especially pronounced among students from less advantaged backgrounds (Jury et al., 2015; Smeding et al., 2013).

Additional Motivational Antecedents Addressed

This study was designed using concepts from the expectancy-value theory of motivation (Wigfield & Eccles, 2000). Additionally, after we had collected our data and begun to analyze it, we found concepts from three additional theories of motivation to be helpful in understanding what the teachers told us in interviews: growth mindset theory (Dweck, 2007), self-determination theory (Ryan & Deci, 2020), and achievement goal theory (Senko, 2016). Growth mindset theory suggests that students show greater motivation, engagement, learning, and performance in school, especially in the face of difficulty, if they believe that they can grow and improve their intelligence and academic skills through hard work (a growth mindset), versus if they instead believe that their ability levels are innate and unchangeable (a fixed mindset). Self-determination theory suggests that students will be more internally motivated to engage with their schoolwork when three needs are met: *competence* (roughly analogous to expectancies for success, described above), *autonomy* (the feeling that they have opportunities to make meaningful choices about their learning), and *relatedness* (the feeling that they are valued and belong in class). Finally, achievement goal theory suggests that students experience greater academic outcomes when they adopt *mastery goals*, i.e., engaging in schoolwork either to learn as much as possible, or else to meet self-determined standards of success in learning. As will be discussed below, each of these themes emerged in interviews with teachers and open-ended survey responses from students who were exposed to PARLO—a program that encouraged the capacity of students to improve their mathematics abilities (i.e., growth mindsets and mastery goals), and provided students with flexibility in how they could go about acquiring and demonstrating this improvement (i.e., self-determination). Below, we refer to feelings, attitudes, beliefs, and goals that contribute to student motivation and engagement as *motivational antecedents*.

Conceptual Model

Figure 1 displays our current conceptual model regarding how the PARLO system may influence students' academic experiences and learning. Because it was developed using the still-to-be described results of the present work, we return to the complete model in the Discussion section. For now, we wish to note the most important difference between the PARLO Treatment and the Control group in the present work: the two white boxes on the bottom left of the figure. Unlike the Control teachers, PARLO teachers were trained and expected to reassess students for full credit after further learning, and to base a student's summative grade on the number of proficient and the number of high-performance learning outcomes the student had mastered. The two gray boxes on the top left—organizing instruction by learning outcomes and providing formative assessment and feedback—are also necessary parts of the PARLO system and were implemented by teachers in our study. However, these two PARLO components were popular innovations already in place in the school districts where we conducted the study. Given this environment, the project delivered professional development to teachers in the

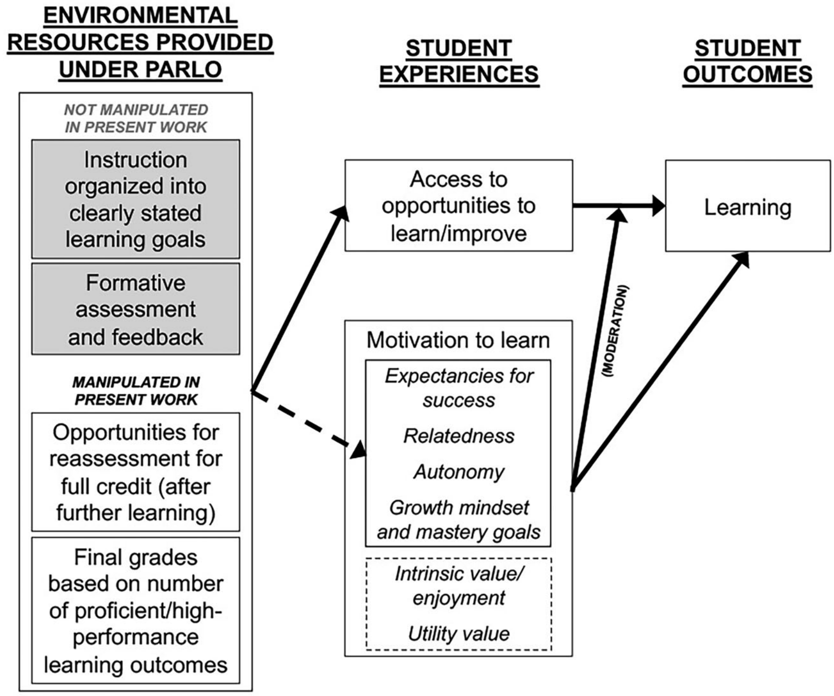

Figure 1. Conceptual model.

Control group to further assist them in implementing these two formative assessment-related practices. We anticipated that PARLO teachers would be more likely than Control teachers to implement the PARLO components displayed in the grayed-out boxes, but only because implementing the bottom two components encouraged them to do so. Thus, the current study tests the impact of experimentally adding the lower two components of the PARLO standards-based grading system—reassessment for full credit after further learning, and summative grades based on the number of proficient and the number of high-performance learning outcomes—to classrooms that were already endeavoring to implement the top two components.

We also note that the focus of the conceptual model, and of the current article, is on *changing teacher's assessment practices*. Specific inputs, such as software tools made available to teachers and specific professional development provided, evolved over the course of the project, and will likely be implemented differently in the future. Consequently these "program inputs" are not displayed in the model. Our intent is to generalize results to future implementations of standards-based grading that are similar to PARLO. See our Conclusion section for ideas about possible future implementations.

Prior Research on Standards-Based Grading and PARLO

While numerous researchers have written about standard-based grading, published work has focused on case studies of small-scale implementation in single classrooms or schools (e.g., Farrington & Small, 2008). Consequently, researchers have decried the lack of studies evaluating the practice (Marzano, 2010; Scarlett, 2018). The current study should help to fill that gap.

There have been two pilot studies of the PARLO system. First, Clymer and Wiliam (2007) implemented PARLO for a full school year in a single eighth grade science classroom. Average achievement on an end-of-class exam increased by 0.4 *SD* over the previous year's scores. In addition, student interviews indicated that these positive effects on performance may have emerged because students developed stronger growth mindsets. Second, Posner (2011) taught two sections of an introductory undergraduate statistics class for non-majors, nonrandomly employing PARLO in one section and traditional grading in the other. He found that the PARLO group scored significantly higher than the traditional grading group on measures of perceived intrinsic value, utility value, expectations for success, and motivation in statistics.

The Current Study

While these initial results were promising, the studies described were only small pilots that did not provide a randomized control-style investigation of the effects of the PARLO assessment system. The present work describes the results of such an investigation. Specifically, we employed a mixed-methods, cluster randomized control approach with school as the unit of analysis, to address the following two research questions: *What is the impact of the PARLO system on 9th grade students' expectancies for success, utility value, intrinsic value, long-term motivation, and academic performance in mathematics class? What are the mechanisms by which this impact occurs?*

In addition, while not part of our *a priori* hypotheses, our examinations of the qualitative data unexpectedly revealed that the relationship between the PARLO system and student motivation might be bi-directional: not only might PARLO improve student motivation, but students with higher initial motivation might also benefit more from the PARLO system. Consequently, we used our quantitative data to investigate an additional research question: *Is the impact of the PARLO system on student academic performance moderated by established antecedents of student motivation—specifically, students' value of and expectancy for success in mathematics?*

Method

This was a mixed-methods study, collecting qualitative and quantitative data throughout program implementation. Our data analysis followed a sequential explanatory design, performed in three phases. In the first phase, we conducted quantitative analyses to see whether the PARLO system had its predicted effects on motivation and on student achievement. In the second phase, we used qualitative data to look for explanations of the results we found in phase 1. Our qualitative analysis began inductively, thus allowing for the emergence of trends we did not hypothesize a priori. We summarized preliminary results as memos describing major themes observed and evidence for those themes. We then reviewed the memos in light of the quantitative findings, looking for explanations of those findings. The second phase yielded an unexpected result, indicating that student motivation might be moderating the PARLO system's effects on student achievement. Consequently, we conducted a third phase, investigating whether the quantitative data we had collected could confirm the moderating effect.

Participants

We recruited urban, suburban, and rural schools; public, charter, and religious schools; schools from high- and low-performing districts; and schools with a wide variety of racial make-ups. In order to ensure reasonably high fidelity of implementation, we only assigned a school to the project if both the administration and the ninth-grade mathematics teachers agreed in advance that they would participate as assigned either in the treatment or in the control condition, and would maintain participation regardless of subsequent random assignment. All ninth-grade algebra and geometry teachers at each school were asked to participate. Teachers received a stipend in exchange for their participation.

We ultimately recruited two cohorts of schools: a cohort of 20 schools (14 public schools, 3 charter schools, and 3 Catholic all-girl schools) that participated during the 2010–11 and 2011–12 school years; and a second cohort of 15 schools (14 public schools from one large urban school district and 1 additional public school) that participated during the 2011–12 and 2012–13 school years. Cohorts participated for two years (i.e., two separate ninth grade classes) because implementing PARLO required significant changes to teachers' instructional practices that we anticipated might take more than a single year to take root.

Several schools dropped out of the study after randomization but before data collection was completed. The final sample of schools in the quantitative study included 14 PARLO and 15 control schools across both cohorts. One additional PARLO school provided qualitative data but dropped out before quantitative data collection could be completed. In all cases, school-level attrition was caused by a change in administration, which is unlikely to be related to the treatment. The overall attrition rate was 17%, and the differential attrition between treatment and control schools was 10.4%. At the student level, the overall attrition rate was 16.9% and the differential attrition was 3.9%. These rates are within acceptable standards when the intervention is unlikely to affect attrition (USDOE, 2022). See Tables A1, A3, and A4 in the online appendix for details about attrition.

Our Institutional Review Board (IRB) required active consent (i.e., parental opt-in) before administering our quantitative measures during the 2010–11 school year—the first implementation year for Cohort 1. However, many students failed to return a consent form, creating a risk of self-selection and bias in the data. As a result, our external reviewer and advisory board for this project determined that the quantitative data collected in 2010–11 could not be used for valid analyses. With the approval of our IRB, this issue was resolved beginning in the 2011–12 school year through the use of passive consent (i.e., parental opt-out). Thus, our quantitative analyses focus on data collected during academic years 2011–12 and 2012–13.

Across two years of data collection, 3,273 students completed algebra or geometry pretests and had available demographic data for the analysis—1,936 in treatment schools and 1,337 in control schools. Of these, 2,736 students provided complete quantitative data, including baseline motivation and motivational antecedent scores, race, gender, and post-test scores, and thus were included in the analytic sample for our primary quantitative analysis. Of the total 2,736 students, 1,649 were from the classes of 38 teachers at the 14 PARLO treatment schools, and 1,087 were from the classes of 27 teachers at the 15 control schools. Tables A5 and A6 in the online appendix compare the demographic characteristics of treatment and control group students and teachers in the analytic sample.

Random Assignment and Conditions

In June 2010, all ninth-grade algebra and geometry teachers at all Cohort 1 schools who had agreed to participate attended three 6-hour days of professional development (PD) in the summer. After teachers completed this PD, project staff then randomly assigned participating schools either to the treatment or the control condition. Randomization was done in three blocks: charter vs. Catholic vs. public. The following year, we randomly assigned Cohort 2 schools to treatment or control conditions following a process that was identical but for two exceptions: the PD before random assignment lasted only two days, and blocking by school type was not necessary as all Cohort 2 schools were public. Participating teachers were aware of their condition assignment. See the online appendix for details about random assignment and for a depiction of the study timeline.

Control Condition

When this study was conducted, two formative assessment practices were coming into widespread use, i.e., becoming business-as-usual for many teachers in local school districts: sharing learning intentions and success criteria by organizing instruction around learning outcomes tied to state standards; and using formative assessment strategies such as providing frequent feedback. Both practices were necessary but not sufficient to implement PARLO. In order to further focus the study on the unique aspects of PARLO that were not already coming into popular use, before random assignment we provided all participating teachers professional development supporting these two widely used formative assessment practices. We had each teacher work with state standards for their course to translate them into 10 to 15 learning outcomes each semester that would be the focus of teaching in their classroom. We introduced Webb's Depth of Knowledge (Webb, 2002) to help them think about addressing each learning outcome at application and strategic-thinking levels. We used ideas from Wiliam (2011) to teach techniques for eliciting evidence from students and providing feedback to move their learning forward. The initial professional development lasted three days for Cohort 1 and two days for Cohort 2. Because the Control teachers were encouraged to implement the two aspects of PARLO reflected in the gray boxes of Figure 1, the project often described the Control schools as a "limited treatment condition," to be contrasted with the "full treatment condition" that implemented all four aspects of PARLO.

Control teachers did not participate in any project-related PD beyond the two or three days that they participated in before randomization, and they were not obligated to modify their instructional practices. In addition to administering the student content exam and Attitudes Toward Mathematics Inventory described below, control teachers completed questionnaire measures of the extent to which they used various practices in their teaching.

PARLO Treatment Condition

Teachers in treatment schools participated in three training opportunities not offered to control teachers. First, in August prior to first implementing the PARLO system in their classrooms, they attended three (for Cohort 1) or four (for Cohort 2) additional six-hour days of PD. Second, during the two years of project participation, they were given the opportunity to participate in monthly Professional Learning Community (PLC) meetings. The PLC meetings were held at five different locations and facilitated by teacher leaders who were mathematics content and PARLO experts. Third, they attended two six-hour days of follow-up PD during the summer between their two years of PARLO participation. Total participation time for treatment teachers in PD and PLC meetings amounted to about 88 hours over the two years. Seventy-nine percent of the treatment teachers attended 70% or more of the PD available to them.

The PARLO-specific PD focused on the two distinguishing characteristics of the PARLO standards-based grading system: 1) Reassessment for full credit after further learning; and 2) Basing a student's semester grade on their number of proficient and number of high-performance learning outcomes. Treatment teachers were provided with a project-developed software tool, PARLO *Tracker*, and trained in how to use it. Teachers could enter into Tracker the learning outcomes they had developed and enter evidence collected, including assessments and reassessments, about each student's progress on each learning outcome. Students and parents could access Tracker online to track student progress. (Tracker was developed with user input and became available at the beginning of the second year of the project, which is the first year that quantitative data used in the current study was collected. Use was optional, and when surveyed during Year 3 of the project, 16 of 25 PARLO teachers (64%) reported using Tracker regularly.) Additional topics covered in the professional development included: creating scoring rubrics for learning outcomes and sharing them with students; helping students track their proficiency; teaching students what to do when they are not yet proficient; requiring proof of new learning before reassessment, including use of tools like error logs and remediation plans; scoring proficiency in learning outcomes based on the best evidence to date instead of traditional methods like averaging scores; and converting learning outcome scores into semester grades. PARLO teachers were responsible for developing their own learning outcomes based on state standards, developing their own methods for informing students about the PARLO system and grading scheme, establishing their own classroom procedures, and deciding on their own means of collecting assessment and/or reassessment evidence. However, through PLCs they were able to share ideas and support each other in developing routines, assessments, and so on.

Quantitative Measures

Implementation Measures

In the spring of 2012, we administered a 7-item measure to Treatment and Control teachers, investigating whether they implemented key aspects of the PARLO system. In the spring of 2013 we administered an additional survey to PARLO teachers only, investigating in more detail to what extent they had implemented the PARLO system. See the online appendix section *Implementation: Additional Details* for a detailed description of the measures.

Algebra and Geometry Content Exams

The state where we conducted our research had recently adopted content standards for both geometry and algebra based on the Common Core State Standards for Mathematics (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010), but had not yet implemented an end-of-course test of the material. Consequently, we created a test to mimic as closely as possible a state-designed test of its published content standards. To do so, we administered (with permission) an adapted version of the Virginia Standards of Learning multiple choice Algebra and Geometry Tests. For each course (algebra or geometry), we selected items from the Virginia test that addressed our state's standards. Then, for any standards not covered, we added items adapted from textbook sources or from released items from older assessments in our participating state. In each participating classroom, we administered the appropriate content exam as both a pretest during the first five days of the school year (Algebra: $M = 34.2\%$, $SD = 11.0\%$, Cronbach's $\alpha = 0.72$; Geometry: $M = 36.0\%$, $SD = 11.3\%$, $\alpha = 0.75$) and, with items in a different order, during the last month of the school year (Algebra: $M = 46.9\%$, $SD = 17.5\%$, $\alpha = 0.90$; Geometry: $M = 49.7\%$, $SD = 14.9\%$, $\alpha = 0.94$). Following suggestions by May et al. (2009) for combining differing tests into a single measure, we used a linear transformation to standardize the pretests within each subject area (algebra or geometry) to have a mean of 0 and a standard deviation of 1, and then combined the two sets of scores to create a single measure. We did the same for post-test scores. See the online appendix for the complete tests used and technical details about computing student scores.

Motivation and Motivational Antecedents Measures

We assessed concepts from the expectancy-value theory of motivation using the Attitudes Toward Mathematics Inventory (ATMI; Tapia & Marsh, 2004). The four ATMI subscales are described below. In each participating classroom, we collected baseline scores on the four ATMI subscales during the first five days of the school year, and we collected outcome data on the same subscales during the last month of the school year. Students responded to all of the subscales using a 5-point scale (1 = "strongly disagree"; 5 = "strongly agree"). See the online appendix for the full set of items from the ATMI and technical details about computing student scores.

Perceived Intrinsic Value of Mathematics. The extent to which students saw mathematics as having intrinsic value was assessed using the ATMI's 10-item "enjoyment" subscale. Items included "I have usually enjoyed studying mathematics in school" (baseline: $M = 3.27$; $SD = 0.88$, $\alpha = 0.89$; outcome measure: $M = 3.08$; $SD = 0.86$, $\alpha = 0.89$).

Perceived Utility Value of Mathematics. The extent to which students saw mathematics as having utility value was assessed using the ATMI's 10-item "value" subscale. Items included "I can think of many ways that I use math outside of school" (baseline: $M = 3.73$; $SD = 0.67$, $\alpha = 0.86$; outcome measure: $M = 3.54$; $SD = 0.74$, $\alpha = 0.87$).

Expectancies for Success in Mathematics. The strength of students' expectancies for success in mathematics was assessed using the ATMI's 15-item "self-confidence" subscale. Items included "I believe I am good at solving math problems" (baseline: $M = 3.45$; $SD = 0.77$, $\alpha = 0.79$; outcome measure: $M = 3.31$; $SD = 0.83$, $\alpha = 0.81$).

Long-Term Motivation to Engage with Mathematics. The ATMI's 5-item "motivation" subscale assessed the degree to which students enjoyed the challenge of mathematics and planned to pursue the subject over the long term. Items included "I plan to take as much mathematics as I can during my education" (baseline: $M = 3.21$; $SD = 0.79$, $\alpha = 0.79$; outcome measure: $M = 3.10$; $SD = 0.85$, $\alpha = 0.81$).

Student-Level Covariates

We coded each student's race and gender based on information provided by the school or, when school-provided data were not available, based on a self-report survey administered simultaneously with the ATMI. A preliminary analysis using Akaike's Information Criterion (AIC; Akaike, 2011) indicated that the most informative way to operationalize "race" was to create a dichotomous variable, with "1" indicating the student was identified as White (51%) or Asian (4%), and 0 indicating that the student was identified as Black (30%), Hispanic (8%) or Multiracial/other (7%).

School-Level Covariates

We obtained from state databases the schoolwide percent low-income students enrolled in the academic year 2010–11, the first year of PARLO implementation for the study. We also aggregated student-level covariates (race, gender, and content pretest and baseline motivation-related scores) to the school level as potential additional covariates to control for in our analysis.

Qualitative Data

Qualitative data were collected throughout the three operational years of the project. The qualitative data sources used for this paper included 84 PARLO teacher interviews and an open-ended survey of 678 students from a nonrandom sample of 6 of the 14 PARLO treatment schools.

Teacher Interviews

During the first year of participation for Cohort 1 and Cohort 2, PARLO teachers were randomly selected to be interviewed; however, if a teacher was the only participant at their school, they were automatically selected to be interviewed. During the 2011–12 school year, the 13 Cohort 1 teachers who were interviewed were the same teachers who were interviewed the previous year. During the 2012–13 school year, all 25 remaining participating teachers were interviewed. Interviews were conducted in the fall and spring of each year, and some teachers were interviewed twice during the same year. Over the three years of the study, 11 teachers were interviewed once, 8 were interviewed twice, 7 were interviewed three times, and 9 were interviewed four times. Overall, 35 teachers at 15 schools participated in 84 interviews.

Interview data for the current article was drawn primarily from questions that addressed how PARLO affected students—for example, "Does student engagement 'look' different? By that, I mean are students interacting with math, with you, with each other differently under PARLO?" See the online appendix for detailed interview questions. Interviews were semi-structured, permitting teachers to expound upon questions that warranted further elaboration. With few exceptions, interviews were conducted following a lesson observation, either during teachers' preparatory or lunch periods, or, rarely, after school. Interviews lasted about 25 to 35 minutes and were audio recorded with the teacher's permission. Every interviewee agreed to be recorded. Each interview was transcribed, then entered into NVivo software for analysis.

Student Survey

During the 2011–12 school year, six of the treatment schools agreed to have their students complete an open-ended online 8-question survey (e.g., "How would you describe PARLO to next year's ninth grade students?") A total of 678 students completed the survey. See the online appendix for the full list of questions asked on the student survey.

Data Analysis

Quantitative Analyses

We conducted all statistical analyses using the *lmer* command from the *lmerTest* package of the R programming language (Kuznetsova et al., 2017). We used Restricted Maximum Likelihood (REML) to test all research questions, but as recommended by Zuur et al. (2009), we used maximum likelihood in preliminary analyses that employed AIC (Akaike, 2011) to choose variance components and covariates, or that used log likelihood to evaluate heteroscedasticity. We calculated degrees of freedom for statistical tests using Satterthwaite's approximation (Satterthwaite, 1946).

Our data set identified each student by course (geometry vs. algebra), teacher, school, and study year. Consequently, the data were structured as a five-level hierarchical linear model: students within course within study-year within teachers within schools. We used AIC to choose the most appropriate variance structure for our analysis, using students' mathematics content post-test as our dependent variable. This analysis indicated that after entering variance components for school and for course-within-year-within-teacher, adding additional variance components for year-within-teacher or teacher-within-school reduced the efficiency of the model. We therefore used the following three-level variance structure in our analyses: "student," within "course-within-study-year-within-teacher," within "school."

Our first quantitative analysis investigated the main effects of PARLO using mathematics content post-test scores as the dependent variable. We used AIC to select covariates for our model, choosing the following set: school level percent low-income students in 2011; a dichotomous indicator identifying students enrolled in geometry; a dichotomous variable indicating female sex; a dichotomous variable indicating White or Asian race (vs. Black, Hispanic, or multiracial/other); students' baseline expectancy score; linear and quadratic terms for the pretest; and interactions between the geometry indicator and the linear and quadratic pretest variables. We also included indicators to ensure that students were compared to other students in the same study-year and in the same blocking group used for random assignments. The reference group was Cohort 1 public school students. Other groups were: Catholic school students; Charter School students; Cohort 2 Public school students tested in 2010–11; and Cohort 2 public school students tested in 2011–12. All non-dichotomous student level variables were centered around the student mean, and percent low-income was centered around the mean of the school means.

Our second quantitative analysis investigated the main effects of PARLO on the four motivation-related subscales obtained by the ATMI. We used the same covariates that we used the first analysis, with the following exceptions: we used scores for all four motivation and antecedent subscales, whereas previously only expectancy scores were used. We also removed the quadratic term for the pretest and its interaction with the geometry indicator. See the online appendix for further details about how we selected covariates to use in each of our models.

Our two main-effects analyses used two approaches to account for missing data. First, we used Full Information Maximum Likelihood (FIML; Allison, 2012), an approach that, like multiple imputation, minimizes bias due to missingness. However, because it is difficult to analyze interactions between treatment and student-level covariates using FIML, and because these interactions were important in addressing *for whom* the PARLO system might be effective, we also ran the analyses using listwise deletion. The main effects estimated by the two analyses were very similar; consequently, we report detailed results of the listwise deletion analysis here. See Tables A10 and A12 of the online appendix for details of the FIML analysis.

Our third quantitative analysis, unlike the first two, was not pre-planned. We conducted the analysis to see whether we could corroborate an unanticipated finding from our qualitative data: According to teacher interviews, student motivation moderated PARLO's impact, such that more highly motivated students benefited more from the PARLO program (see Results for details). We therefore used interaction terms to investigate how our four motivation-related quantitative variables moderated PARLO's impact on mathematics content post-test scores.

Note that for each of the four measures of motivation, we had two observations per student: a score from the first week of the school year, and a score from the last month. Because motivation can change throughout the school year, our preferred approach to measuring motivation across the year would be to average the baseline and end-of-year scores together. We note that end-of-year motivation scores might be endogenous, which can make interpretations of some models confusing or bias results under certain circumstances. However, there is preliminary evidence (based on Monte Carlo results) that if x is an endogenous variable and w is an exogenous variable, the coefficient of the interaction term xw (i.e. the moderator effect) will be consistently estimated as long as w is binary and x is homoscedastic conditional on the other variables in the model (Bun & Harrison, 2018). PARLO treatment was a binary variable, and all four motivation variables were homoscedastic, so we could safely analyze the interactions using the averaged baseline-and-end-of-year scores. These moderation analyses used the same covariates we had used when testing the main effects of PARLO on mathematics post-test scores, except that instead of using the baseline confidence score as a covariate, we included the main effect for each moderator being tested. As a sensitivity check, we also re-ran the moderation analysis using baseline scores instead of year-average scores for all four motivation-related moderators. See the online appendix for details.

As recommended by the What Works Clearinghouse Procedures Handbook (USDOE, 2020), whenever we used multiple statistical tests to address a single issue we used the Benjamini-Hochberg (BH) correction to account for multiple comparisons. The BH procedure controls the false discovery rate for multiple comparisons, ensuring that the expected proportion of falsely identified statistically significant results equals the intended alpha, in this case 5%.

Qualitative Analyses

At the conclusion of teachers' participation in the study (June 2013), two qualitative researchers began coding three years' worth of interview transcripts. To enhance reproducibility, the researchers established a coding scheme using a procedure similar to the three-stage process outlined by Campbell et al. (2013). In the first stage, each researcher read the same three interview transcripts, jotting notes in the margins to identify possible broad themes. Once themes were identified, the researchers met to compare notes and look for agreement or disagreement. If there was a lack of consensus, they discussed why they had coded the text under a particular theme, until agreement was reached that either only one of their themes was accurate, or that both of their themes were accurate. In the latter instances, the text was double coded. At the end of this process, 17 broad themes were identified.

In the second stage, the remaining transcripts were divided, and each researcher coded half of the interview transcripts. In the third stage, each researcher assumed responsibility for the analyses of a defined set of themes. The researchers then prepared memos by theme and study year, with each memo summarizing major evidence for the existence of that theme in the interviews conducted that year. Initial analysis of the student survey followed a similar process, with an additional analysis that identified key words or phrases that appeared frequently in students' responses to survey prompts.

To produce the findings described in this article, a third researcher read all three years' worth of memos to both identify any claims that addressed how or why the PARLO system affected students' motivational antecedents, motivation, or academic performance, and to summarize evidence supporting those claims. Finally, for claims that were supported by numerous teachers' interviews, the research team made a final pass through all 84 transcribed interviews to quantify how many teachers supported that particular claim.

Results

PARLO Implementation

All PARLO teachers interviewed indicated they had implemented the new system, allowing reassessment for full credit and basing final grades on the number of proficient learning outcomes and the number of high-performance learning outcomes. In practice, most teachers rated individual assignments using a "traffic light" system: Red (not yet proficient), Yellow (approaching proficient), Green (proficient), or Blue (high performance; available only for high cognitive demand tasks). In interviews, teachers confirmed that learning outcome grades were developed by teacher judgments based on cumulative evidence, with more recent evidence weighted most heavily. Most PARLO teachers also required proof of learning before reassessment. In an online questionnaire administered to the 25 PARLO teachers participating during the 2012–13 school year, 88% reported using resubmission/correction of work, error logs, remediation plans, and/or flashback days at least "fairly often."

When surveyed in the spring of 2012, PARLO teachers were significantly more likely than Control teachers to agree that they planned to utilize "a 'high performance', 'proficient', or 'not yet proficient' assessment system" (76% vs. 22% agreement) and use "the traffic light system to monitor student progress" (50% vs. 15% agreement). Both these contrasts were statistically significant after using the BH adjustment to control for a false discovery rate of 5%.

Two additional differences between groups were significant after controlling for a false discovery rate of 10%, but not 5%. First, PARLO teachers were somewhat more likely to implement "holding students accountable for learning outcomes" (94% vs. 74%) and somewhat more likely to implement "collecting evidence of student learning" (88% vs. 67%). Both Treatment and Control groups indicated they were highly likely to implement "using formative assessment strategies in your classrooms": 91% for PARLO teachers and 93% for Control teachers. See the online appendix for additional details about the implementation measures and their results.

Overall, the above data indicate that the contrast between the PARLO teachers and Control teachers appears to have been as intended. A large majority of both groups implemented the PARLO characteristics depicted in the gray boxes of Figure 1, although there is evidence that the PARLO group may have implemented learning outcomes more extensively. The PARLO teachers were much more likely than Control teachers to base a student's grades on the number of proficient and number of advanced learning outcomes. While we do not have statistics on the percentage of Control teachers who gave students the opportunity to reassess for full credit after additional learning, all PARLO teachers interviewed indicated that they had adopted this practice.

PARLO's Impact on Quantitative Measures of Mathematics Performance

Main Effects: Did PARLO Increase Mathematics Performance?

Table 1 summarizes the PARLO treatment's main effects on student mathematics learning. Recall that students' mathematics content test scores were standardized, so that within each subject area (algebra or geometry), the post-tests had a mean of 0 and a

Table 1. PARLO treatment main effect on mathematics achievement.

	Estimate	Standard Error	Degrees of Freedom	p-value
School Level Fixed Effects				
Intercept	−0.66	0.12	20	<.0001
Catholic Girls' School	0.30	0.22	21	.18
Charter School	−0.05	0.29	33	.86
Cohort 2 Public School, 2011-2012 school year	0.23	0.15	14	.14
Cohort 2 Public School, 2012-2013 school year	0.34	0.15	14	.035
Proportion Disadvantaged	−0.53	0.23	12	.039
PARLO Treatment	**0.33**	**0.12**	**14**	**.014**
Course Level Fixed Effects				
Geometry Student	0.20	0.16	50	.23
Student Level Fixed Effects				
Baseline Expectancy of Success	0.20	0.02	2,687	<.0001
White or Asian	0.15	0.04	2,654	<.0001
Female	0.10	0.03	2,672	.0005
Pretest	0.26	0.02	2,710	<.0001
Pretest-squared	0.04	0.01	2,673	.0028
Geometry X Pretest	0.20	0.05	2,720	.0002
Geometry X Pretest-squared	−0.02	0.03	2,700	0.54
Random Effects	*N*	*Variance*		
School	29	0.03		
Course x Teacher x Year	85	0.14		
Residual	2,736	0.53		

Notes: This analysis utilized data from 2,736 students, 65 teachers, and 29 schools. "Mathematics Achievement" was defined as the Algebra post-test z-score for algebra students and the Geometry post-test z-score for geometry students. All non-dichotomous variables are grand mean centered. The reference group for the blocking variables was Cohort 1 Public School students who participated during the 2011-12 school year. The intercept represents the post-test score for students at average values for pretest, expectancy of success, and school-level proportion disadvantaged and in the reference group for all dichotomous variables (i.e. in the Cohort 1 Public School randomization block; in a control school, taking algebra, not White/Asian, and not Female).

standard deviation of 1. Consequently, the results reported in Table 1 can be interpreted as effect sizes in standard deviation units. After controlling for random-assignment grouping, students' pretest scores, schoolwide percent low-income, geometry vs. algebra class, race, sex, and baseline expectancy for success, students in the PARLO system scored 0.33 *SD* higher than students in the control group on the project-administered end-of-course tests ($p = 0.014$).

Using post-test standard deviation as the denominator and combining Treatment with Control groups, Geometry students gained approximately 0.91 standard deviations and Algebra students gained approximately 0.73 standard deviations from our pretest to our post-test. Thus, an impact of 0.33 *SD* might be thought of as roughly equivalent to about 36% (for Geometry students) to 45% (for Algebra students) of a year's learning.

Interaction Effects: For Whom Did Academic Performance Increase?

We investigated interactions between the PARLO program and each of the student level covariates in our model to determine whether baseline student characteristics moderated the effects of the PARLO system on mathematics content learning. There were no statistically significant interactions. In other words, students benefited from the PARLO system regardless of their course enrollment (geometry or algebra), their content pretest scores, their baseline expectancies for success in mathematics, their race, or their gender. See the online appendix for complete details regarding these analyses.

PARLO's Impact on Quantitative Measures of Students' Mathematics Motivation

Main Effects: Did PARLO Impact Motivation?

Table 2 summarizes the PARLO treatment's main effects on students' self-reported intrinsic value, utility value, expectancies for success, and long-term motivation to engage with mathematics. (See Table A11 of the online appendix for more detailed information including covariate coefficients and variance components.) As can be seen in the table, the PARLO system had substantively small and statistically non-significant effects on post-test measures of the four constructs, with effects ranging from −0.08 to −0.02 on a 5-point scale.

Interaction Effects: Were There Any Subgroups for Whom Average Motivation Changed?

While there were no main effects of treatment on any of the motivation related measures, it seemed possible that one or more subgroups of students might have been impacted. We investigated this possibility by testing whether any of our four measures were impacted by interactions between treatment condition and each of our student level covariates: gender, race, geometry vs. algebra class, and baseline expectancies, intrinsic value, utility value, long-term motivation, and mathematics content test scores. After using the Benjamini-Hochberg (BH) procedure to control for multiple comparisons, none of the interactions were significant. (See the online appendix for additional details.) Thus, we found no quantitative evidence that the PARLO system affected the intrinsic value, utility value, expectancies for success, or long-term motivation of any student subgroup.

Mechanisms Leading to PARLO's Positive Impact on Mathematics Performance

Opportunities to Learn

We had hypothesized that PARLO would improve student achievement partly by improving opportunities to learn. Teacher interviews provided detail about how increased opportunities may have operated.

Table 2. Effects of PARLO treatment on students' motivational antecedents and long-term motivation in math class.

Dependent Variable (Subscale)	PARLO effect Estimate	Standard Error	df	95% conf. interval	p-value	Effect size in SD units
Intrinsic Value	−0.06	0.06	15	(−0.18, +0.06)	.31	−0.07
Utility Value	−0.02	0.05	14	(−0.12, +0.07)	.60	−0.03
Expectancy	−0.06	0.04	51	(−0.15, +0.02)	.14	−0.08
Long-term Motivation	−0.08	0.06	19	(−0.21, +0.05)	.23	−0.09

Notes: This analysis utilized data from 2,698 students, 65 teachers, and 29 schools. Dependent variables are measured on a 1-5 Likert scale. *df* = degrees of freedom. Standard deviations for the dependent variables were as follows: Intrinsic Value: 0.86; Utility value: 0.74; Expectancies for success: 0.83; Long-term motivation: 0.85. Reported results controlled for the following covariates: Assignment block for randomization; School-level proportion disadvantaged; course assignment (geometry or algebra), student-level gender, race, pretest, and baseline score on all four subscales of the ATMI.

Reassessment Opportunities. The central feature of PARLO is encouraging students to re-study and reassess to achieve proficiency or high performance on each learning outcome, and a recurring theme—emerging among 89% of teachers interviewed (31/35)—was that teachers felt this central feature of PARLO led directly to better student learning of mathematics content. Specifically, teachers reported that reassessment opportunities were useful to students of all ability levels. Regarding struggling students, teachers made comments like the following:

> … some of the students that have typically struggled in math are able to see how their hard work pays off … I think it's something that they have not had before, because if they struggled on a test, even if they came in and got help, that test grade still remained, and they weren't able to demonstrate their understanding. (Teacher #31)

Teachers noted that high achievers also benefited, specifically from opportunities to reassess for high performance. For example:

> The kids that really, really want to learn it will do anything that they can to achieve that high performance. In the traditional class, they've got that one shot to master it, and they can't revisit it. So, they can't ever get to that achievement. They can't push themselves to reach that next level. It's either one or none. [With] the PARLO, they can go back and push themselves to get it. (Teacher # 32)

Clearer Focus on Learning. Sixty percent of teachers interviewed (21/35) also noted that providing feedback and reassessment opportunities for each learning outcome helped students focus their learning efforts fruitfully, asking better questions and showing better understanding of what they needed to work on—for example:

> They're coming to me with better questions. They're not just coming to me and saying, "What can I do?" They come in knowing where they need to focus, knowing what they need to work on. (Teacher #15)

Peer Interactions. Although the interviews did not explicitly ask about student peer support, 49% of teachers interviewed (17/35) brought up peer collaboration under PARLO as one contributor to better student learning—for example:

> They tend to actually be getting involved more with other students, because … they also recognize quickly that in teaching others, they're getting a better understanding themselves because now they can present it and defend. (Teacher #6)

Revisiting Topics. Thirty-one percent of teachers interviewed (11/35) noted the pedagogical value of allowing students to revisit topics over time—for example:

> One big thing is that PARLO allows me time to come back and go over things that we already covered … And by going over the learning outcomes … I see students that have gaps in their learning. Like, one student, she was truant in February and March, but she is a very high-level student, so now I see her learning all of that stuff and connecting the dots between our most recent learning outcomes and stuff that we learned three months ago. And also, just kids filling in the gaps in their own learning as we go through it. "Oh, I remember this" or "I forget how we do this" and then re-remembering it … I noticed that they're getting it very easily now the second time around. (Teacher #23)

Increased Engagement

While our quantitative analysis did not detect any program impact on motivation, teacher interviews indicated that student engagement, which is often associated with motivation, did increase. Eighty percent of teachers (28/35) interviewed said that under PARLO, their students were more likely to participate in discussions about content with both the teacher and their fellow students. The same percent of teachers said that students were more likely to ask questions when concepts were not well understood.

When asked to describe "successes" under the PARLO program, 63% of teachers (22/35) reported increased student persistence in the face of initial difficulty, increased ownership of the learning process, and increased responsibility—for example:

> Students are taking ownership in what they are learning … They are learning how to organize and keep up with it, each marking period they are getting better at it. (Teacher #15)

> Earlier this year, the K-8 math coordinator visited my classroom, to follow up on some of her students from last year. After class she said to me, "Was that Rachel [pseudonym] In the front row, with her head up?" I said, "Yes." And the coordinator said she could not believe it was the same girl because last year, Rachel, when she came to class, sat in the back of the classroom with her head on her desk. This year, Rachel told me during one-on-one conference that she wants high performance on every Outcome. And she's always asking for more and more tests. She's also begging me to teach her next year as she wants to stay in a PARLO classroom. (Teacher #4)

Examining Potential Reasons behind PARLO's Non-Significant Effects on Quantitative Measures of Student Motivation

Posner's (2011) pilot study of the PARLO system found positive effects on intrinsic value, utility value, expectancy for success, and motivation in college statistics. In contrast, the current study found no effects of PARLO on the motivation-related constructs we measured. Our qualitative analysis investigated why this might have occurred.

Decreased Engagement for Some Students

Despite teachers' reports that under PARLO overall engagement increased, a minority of teachers expressed concern that the PARLO system might interfere with some students' engagement. They identified two potential problems, which we labeled "contentment" and "procrastination."

Contentment. Twenty-nine percent of teachers interviewed (10/35), indicated that PARLO might discourage some students from doing their best. Specifically, they noted that some students seemed content with proficient performance and reluctant to try high cognitive demand problems in order to achieve high performance—for example:

> What I stopped doing was marking the problems with [an] asterisk if it were blue [high performance], because students that are not that motivated would just choose not to do those … (Now) they don't know if it's green [proficient] or blue [high performance] until we get the assessments back. (Teacher #28)

Procrastination. Twenty-six percent of teachers interviewed (9/35) indicated that the PARLO system might encourage some students to put off studying under the assumption that if they did poorly at first, they could always reassess later—for example:

> As far as kids that don't take advantage, we have so many. They're the kids that aren't motivated… John [pseudonym] has gotten to the point where he pretty much hands in tests with nothing on them. And I think that he does subconsciously think, "Oh, I can reassess," but he doesn't follow through. (Teacher #36)

Potential Net Zero Impact on Expectancies for Success

Re-assessment opportunities under PARLO appeared to increase students' expectancies for success, but this positive effect was perhaps balanced by a negative effect on expectancies caused by student uncertainties about their final grades.

Positive Effects of Re-Assessment Opportunities. In 2011–12, 678 PARLO students completed an open-ended survey about their experiences with the new assessment system. When asked "What do you like about PARLO?" the most common word or phrase students mentioned was "retake" ($N = 90$), followed by "make up" ($N = 52$) and "another chance" ($N = 48$). Students made comments like, "It gives you the chance to retake tests, quizzes, or homework. You can work at your own pace without having to rush to learn something new," and "I think PARLO is very good for math students struggling."

Negative Effects of Uncertainties about Final Grades. When asked, "What do you not like about PARLO?" 41% of students (277/678) indicated that they felt they never knew their exact grade until report card time or the end of the year, which they found confusing. When asked, "How would you describe PARLO to next year's ninth grade students?," the single most commonly mentioned word was "confusing," mentioned by 73 of the 678 students interviewed. Even students who found the system helpful complained about uncertainty in grading, making comments like, "PARLO is very helpful. You can always bring your grade up. However, one downer is that you never know what your grade is until report cards arrive."

No Detected Impact on Intrinsic Value and Utility Value

As noted in our methods section, the first step in our qualitative analysis was inductive, identifying major themes that emerged from the data. Changes in students' intrinsic value for mathematics or changes in students' beliefs about mathematics utility value did not emerge as themes. Unlike student expectancies for success, which seemed to have a balance of positive and negative impacts, we found no evidence that the PARLO system had either a positive or a negative impact on students' valuing of mathematics.

Unmeasured Motivational Constructs

A majority of the teachers interviewed (24 out of 35, or 69%) indicated that students' experiences with PARLO enhanced their motivation. The discrepancy between this finding and our quantitative results (i.e., no program effect on motivational antecedents or long-term motivation) may have occurred because, by "motivation," teachers were referring to constructs we did not measure quantitatively.

Mastery Goals. Of the 35 teachers interviewed, 24 (69%) indicated that students' experiences with PARLO promoted the adoption of mastery goals. For example:

> … last year all they were concerned about were points and grades. This year they are talking about math more than my students last year. They know their content. (Teacher #26)

Growth Mindsets. Of the 35 teachers interviewed, 20 (57%) also reported that, under PARLO, students were more likely to adopt a growth mindset—for example:

> They are finally, really understanding that knowledge is gained and built over time, you don't just know something, you don't just get something and that's it. You have to work at it, and if you want to keep it you have to continue to work at it. (Teacher #24)

> I think their work ethic is better. They realize they need to keep on top of stuff, keep working. If you fail one thing, don't just say, 'Oh, it's over.' It's not over! You can still learn stuff. (Teacher #1)

Autonomy. A smaller but still substantial number of teachers (34%; 12/35) described a different mechanism by which PARLO increased motivation: that under PARLO, some students felt that they had autonomy and could control their own mathematics destinies (Ryan & Deci, 2020)—for example:

> They have the ownership of their grade now. It's no longer "What can he give me?" So, the ball is in her court now. It's no longer, "You failed me, or you gave me this grade." They talk about getting their grade up to where they want it to be. (Teacher #16)

It should be noted, however, that some teachers highlighted one characteristic of the PARLO assessment system that caused some students to perceive *decreased* autonomy: Under PARLO, a student's grades are based solely on evidence of content understanding. Compliance measures like homework completion and attendance no longer count toward grades. Some students expressed concern that they could control their compliance, but not necessarily ensure their learning.

Relatedness. Finally, PARLO may also have helped meet students' needs for relatedness (Ryan & Deci, 2020). As noted above, teachers reported that increased peer support was one of the reasons that mathematics performance improved under PARLO. This increased peer support may have been effective because it improved students' sense of relatedness. For example, a teacher described classmates' response to a PARLO student coming up to the chalkboard to demonstrate proficiency on a learning outcome:

> They were still supportive … And they're like, "Come on, Jenny (pseudonym), you can do it! You just have a few more steps to go! You're almost there!" Never did they yell out, "Hey stupid, wrong step … " They're using the language [of PARLO]. (Teacher # 4)

The Emergence of Student Motivation as a Potential Moderator of PARLO's Impact

We turn now to an unexpected finding that emerged from our qualitative analysis. We had theorized that the PARLO system might improve students' math motivation. In contrast, teacher interviews suggested that the converse might also be true: higher student motivation might improve the effectiveness of the PARLO system.

Why did some students respond to the PARLO system with active and enthusiastic engagement, whereas other students responded with apathy and work avoidance? As illustrated in the quote from Teacher #36 in our discussion (above) of procrastination, teachers attributed the different student behaviors to differences in student motivation. In fact, 71% of teachers (25/35) noted that "motivation," "caring," "willingness" or a similar concept was important in determining the effectiveness of PARLO, making comments like the following:

> Motivation is huge. And if the kids don't have the motivation, then who cares if they can take a test? They don't care. (Teacher #2)

> The thing I really like about PARLO is the kids that are willing to work and seem to actually care, are the ones who really seem to benefit from it. (Teacher #9)

Quantitative Verification of Motivation as a Moderator of PARLO Impact

As noted in our Methods section, when we designed this study we did not plan to conduct a quantitative analysis to see whether student motivation might moderate the PARLO treatment's effect on student mathematics learning: However, after the qualitative data indicated that this might be the case, we conducted a quantitative analysis to see whether we could corroborate the unanticipated qualitative finding. If student motivation did moderate treatment effects on student achievement, we would expect to see a positive interaction between PARLO treatment and each of our three measured psychological antecedents of motivation, as well as between PARLO and our measure of long-term motivation. Table 3 summarizes the results of the analysis.

As can be seen in the table, all four interactions were positive and three of the four were statistically significant. To interpret the table substantively, the positive impact of PARLO on the end of the year mathematics content score increased by 0.09, 0.11, 0.13, and 0.07 *SD* for each 1-point increase on the 5-point measure of intrinsic value, utility value, expectancies, and long-term motivation, respectively. Thus, PARLO's impact on individual students' mathematics achievement was positively affected by their levels of motivation in mathematics.

We used the BH adjustment for multiple comparisons to account for the fact that Table 3 employs four statistical tests. The BH procedure confirmed the statistical significance of intrinsic value, utility value, and expectancies for success as moderators. Note that the analysis reported in Table 3 averaged the baseline and end-of-year scores to measure each of the four motivation-related constructs. As a sensitivity check, we also ran the moderation analysis that instead used only baseline scores for all four motivation-related moderators. The results pointed in the same direction as those reported in

Table 3. Interactions with treatment condition: do students' perception of intrinsic value, utility value, expectancy of success, and long-term motivation moderate PARLO's impact on mathematic achievement?

Interaction	n	Effect size Estimate	Standard Error	Degrees of Freedom	95% conf. interval	p-value
PARLO × Average Intrinsic Value	2,529	0.09	0.04	2,482	(+0.01, +0.17)	0.026
PARLO × Average Utility Value	2,532	0.11	0.05	2,483	(+0.01, +0.21)	0.028
PARLO × Average Expectancy	2,529	0.13	0.04	2,481	(+0.05, +0.22)	0.002
PARLO × Average Longterm Motivation	2,533	0.07	0.04	2,485	(−0.02, +0.15)	0.13

Notes: This analysis utilized data from 65 teachers, and 29 schools. "Mathematics Achievement" was defined as the Algebra Post-test z-score for algebra students and the Geometry post-test z-score for geometry students. n = number of students with data available for the model testing each specific moderator. Because the analysis utilized average of baseline and end-of-year scores, students who were missing end-of-year scores were not included in the analysis. Reported results controlled for the following covariates: Assignment block for randomization; School-level proportion disadvantaged; course assignment (geometry or algebra), student-level gender, race, pretest, pretest-squared, main effects of the motivation subscale being studied and main effects of the PARLO treatment.

Table 3 but were not statistically significant. See the online appendix for details about the BH procedure and the sensitivity check.

Discussion

The present work investigated the impact of the PARLO standards-based grading system on ninth graders' learning of algebra and geometry, hypothesizing that the reengineered grading system would positively impact student learning both by directly providing students additional opportunities to learn, and in a mediated manner by enhancing student motivation. Partly supporting these predictions, our quantitative analysis indicated that PARLO had a positive impact on ninth graders' learning of mathematics content, with an estimated effect size of 0.33 *SD*. In normally distributed data, an effect size of 0.33 *SD* would move a student from the 50th percentile on a test up to the 63rd percentile. Comparing the PARLO impact to typical student growth from pre- to post-test in our data set, we estimate 0.33 *SD* to be the equivalent to about 36% to 45% of a year's learning. The program was effective regardless of students' race, gender, or prior achievement.

As discussed, we had also hypothesized that one mechanism through which PARLO might improve achievement would be the mediating effect of motivation: the PARLO program would improve motivation, which in turn would improve student engagement and learning. Our quantitative analysis did not support this hypothesis. We found no significant effect of PARLO on any of our four motivation-related measures.

Our qualitative analysis provided insights into the reasons behind our quantitative findings. As predicted, teachers reported a number of ways that increased opportunities to learn under PARLO improved student achievement. These opportunities included the direct effects of encouraging students to re-study and reassess, helping students to focus on the content they needed to learn, increased support from students' peers, and providing opportunities to revisit topics over time.

Qualitative data also provided insight into why students did not report increased expectancies for success under PARLO. There appear to have been two countervailing tendencies, potentially leading to a net zero impact. As predicted, students' appreciation of the opportunities to reassess appears to have strengthened their expectancies. Balancing this, however, one aspect of our PARLO implementation may have had the unintended consequence of decreasing students' expectancies. The assessment system at YWLCS, which inspired PARLO, did not assign students letter grades, instead using long report cards that described proficiency by learning outcome. In our professional development sessions, we showed these report cards to teachers and suggested, "When all outcomes have been rated, students' overall proficiency can be converted to a letter grade if necessary." As a result, many PARLO teachers provided students with little information about their overall grade until all outcomes had been rated at the end of the semester or year. This appears to have created anxiety and confusion among some students about their grades, which may have reduced their expectancies for success. We anticipate that future implementations of the PARLO system will correct this problem, instead encouraging teachers to help students understand where they stand relative to a final mathematics grade and what they need to do if they want to improve it.

Our quantitative analysis did not indicate any program effects on students' valuing of math class, and the qualitative data seldom mentioned the topic. While we had hypothesized that changing the assessment system to focus less on quick learning and ranking and more on mastery of course content would positively impact students' value of mathematics, this expectation may have been unreasonable. Making mathematics more enjoyable (i.e., increasing intrinsic value) or increasing the perceived utility value of mathematics may be most directly impacted by designing engaging instructional activities or by explicitly connecting curriculum to potential applications, neither of which was a focus of PARLO.

Teachers interviewed did, however, report that the PARLO system positively impacted student engagement and motivation in ways that our quantitative analysis did not address. Motivational antecedents that may have been positively affected include growth mindsets and mastery goals, autonomy, and relatedness.

An additional unanticipated and noteworthy finding emerged from our qualitative analysis. Teacher interviews indicated that student motivation moderated PARLO's impact—that is, that more motivated students benefited more strongly from PARLO. Once this finding emerged from our qualitative analysis, our quantitative data provided some corroboration. Students who scored higher on year-average scores for intrinsic value, utility value, or expectancies for success in mathematics benefited more from the PARLO system than did students who scored lower on those same measures.

To unify these distinct but interconnected findings, we return to Figure 1, which depicts our current conceptual model regarding how the PARLO system may influence students' academic experiences and learning. Solid lines in the figure indicate relationships that have been supported either by prior research or that were supported both by our present quantitative *and* qualitative analyses. Dashed lines indicate relationships that we hypothesize based on patterns that emerged in our qualitative data, but have not received robust quantitative support in the present or past research, and the lack of a connection indicates relationships that did not emerge in the present or prior research.

We note that this current conceptual model includes two refinements to our initial hypotheses. First, we now theorize that the PARLO system impacts motivation indirectly through improving some, but not all, of the motivational antecedents we have considered. Specifically, the current study provides evidence that the PARLO standards based grading system impacts expectancies, autonomy, relatedness, growth mindset, and mastery goals–but not perceived intrinsic value or utility value. Second, motivation is believed to moderate the relationship between PARLO-provided opportunities to learn and actual learning. Thus, supplementing the PARLO program with additional support for motivational antecedents would in theory magnify the impact of the program on learning.

Study Limitations

We designed our study using an expectancy-value framework (Wigfield & Eccles, 2000), and our quantitative measures reflect that design. Findings that emerged from our qualitative analysis of teacher interviews caused us to expand the framework to include constructs from growth mindset theory (Dweck, 2007), self-determination theory (Ryan & Deci, 2020), and achievement goal theory (Senko, 2016). However, because these connections emerged through inductive qualitative analyses conducted following data collection, quantitative data on growth mindsets, autonomy, relatedness, and mastery goals were not collected. This may have contributed to the discrepancy between teacher interviews, which reported increased motivation under PARLO, and quantitative analyses, which did not detect a program impact on motivation.

The study would have been stronger had we interviewed the control teachers. Doing so might have provided important insight into the ways in which treatment and control students' experiences in the classroom differed.

The PARLO system works by changing the grading system to provide students with more opportunities to learn. To be effective, students must understand the system. We did not provide direct assistance to help teachers explain the PARLO grading system to their students, and our data indicate that some students did not understand it. In a free response survey, 73 out of 678 students described PARLO as "confusing." Future implementations might be able to increase PARLO's effectiveness by providing materials and techniques teachers can use to explain the new system to their students.

Informal aspects of our intervention may have influenced outcomes in important but untracked ways. For example, some but not all teachers encouraged peer-to-peer support by grading a student as high performance on a learning outcome if they helped another student attain proficiency. "Flashback days" were an innovation developed by some participating teachers and shared with all PARLO teachers through PLCs, but not implemented by all of them. Teachers also developed techniques to reduce workload requirements under PARLO. Some of these innovations may be key ingredients that greatly enhance program success, but we were not able to determine which innovations were the most important. Future research should seek to track how such elements—in addition to factors like the level of transparency with which proficiency ratings are translated into final grades, the way the program is introduced to students, the amount and types of supports provided to parents and guardians, and the number of courses or grade levels in a school that implement the PARLO program—might affect program results.

Finally, we note that treatment and control teachers had all volunteered, if assigned to treatment, to change a central pillar of their classroom instruction: the grading system. Implementing the PARLO system required considerable effort from the teachers involved. Results may not generalize to teachers or schools who are less willing to make such a change.

Conclusion: The Future

Working with teachers who were already implementing formative assessment and mathematics instruction designed around clearly state learning outcomes, the PARLO standards-based grading program reinforced those teaching practices and added two new elements: reassessment for full credit after further learning; and basing a student's final grade on the number of learning outcomes the student learned at a proficient level and the number of learning outcomes the student learned at a high performance level. This treatment improved the amount of mathematics students learned by 0.33 standard deviations, which is roughly 36% to 45% of the effect an entire year's learning had for our sample.

Moving forward, we anticipate that future work will not replicate the current project precisely, but rather build on lessons learned. For example, we recommend that future implementations correct elements of the program that may have unintentionally reduced program effectiveness. Specifically, we recommend that procedures for computing final grades be made transparent and explained clearly to students; and that the program provide materials and procedures to assist teachers in introducing the PARLO program to students. If possible, future implementations should provide optional-use assessment, reassessment, and re-teaching materials to reduce the amount of work required of PARLO teachers. Future work might also consider implementing standards-based grading in more courses or more grade levels (vs. ninth grade math only), so that the PARLO grading system does not seem so unusual (and thus potentially confusing) compared to grading in other courses.

Our data also provides evidence that the PARLO standards-based grading system, while effective regardless of race, gender, or prior knowledge, is especially effective for students who are more highly motivated. This finding makes sense: the PARLO system provides students additional opportunities to learn, and more motivated students are more likely to take advantage of those opportunities. If one views motivation as a fixed attribute of students, then this finding could be seen as troubling. A standards-based grading system would have positive effects, but it would be especially helpful for the "haves" who are already motivated and engaged.

Fortunately, however, research has consistently shown that motivation is malleable: the mindsets that are psychological antecedents to motivation, including beliefs in intrinsic and utility value of subject matter, expectancies for success, growth mindsets, autonomy, and relatedness are greatly impacted by teacher behaviors and classroom learning conditions (Binning & Browman, 2020; Farrington et al., 2012). Furthermore, educational psychologists have developed interventions that are designed to target the specific underlying psychological processes that can promote student motivation and engagement. Importantly, a growing body of research has found that such interventions can have strong positive impacts on student learning and well-being, but only in contexts that are fertile ground for supporting the target psychological factors (Walton & Yeager, 2020, Yeager & Walton, 2011). To create such "fertile ground" researchers have suggested that it may be especially important to reform grading systems in order to make them less focused on summative assessment and ranking (Farrington et al., 2012). This raises an exciting avenue for future work: implementing the PARLO standards-based grading system, while simultaneously supporting PARLO with interventions that foster student motivation. Such a combined intervention could theoretically magnify the impact of PARLO by supporting student motivation, while simultaneously magnifying the impact of the motivation interventions by creating a grading environment in which they can have a larger impact.

Finally, we note that there is no theoretical reason to expect the PARLO standards-based grading system to be less successful in other grade levels or in other disciplines than it was in ninth grade mathematics. We encourage other researchers to build on the potential identified in our work, finding effective ways to implement standards-based grading. We hope that the results will have large benefits for cultivating student learning.

Open Research Statements

Study and Analysis Plan Registration

There is no study and analysis plan registration associated with this manuscript.

Data, Code, and Materials Transparency

The data, code, and materials underlying the results reported in this manuscript are available on the Open Science Framework: https://osf.io/htkfs.

Design and Analysis Reporting Guidelines

This manuscript is accompanied by a completed JREE Randomized Trial Checklist.

Transparency Declaration

The lead author (the manuscript's guarantor) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Replication Statement

This manuscript reports an original study.

Acknowledgments

The authors thank Helen Kramer, Holly Bozeman, John Baker, Teresa Harrison, Maurice Bun and two anonymous reviewers for their feedback. This research was supported by the National Science Foundation under grant number DRL-0918474. Research involving human subjects was reviewed and approved by the Institutional Review Board (IRB) at Research for Better Schools. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Open Scholarship

This article has earned the Center for Open Science badges for Open Data and Open Materials through Open Practices Disclosure. The data and materials are openly accessible at 10.35542/osf.io/pzc3f.

ORCID

Steven L. Kramer http://orcid.org/0000-0003-2900-6484
Alexander S. Browman http://orcid.org/0000-0002-2957-3262

References

Akaike, H. (2011). Akaike's information criterion. In M. Lovric (Ed.), *International encyclopedia of statistical science* (pp. 25–25). Springer. https://doi.org/10.1007/978-3-642-04898-2_110

Allison, P. D. (2012). Handling missing data by maximum likelihood. In *SAS Global Forum*, Paper 312-212. Retrieved September 8, 2021, from https://statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf.

Binning, K. R., & Browman, A. S. (2020). Theoretical, ethical, and policy considerations for conducting social–psychological interventions to close educational achievement gaps. *Social Issues and Policy Review*, *14*(1), 182–216. https://doi.org/10.1111/sipr.12066

Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *92*(1), 81–90. https://doi.org/10.1177/003172171009200119

Bloom, B. S. (1968). Learning for mastery. Instruction and curriculum. Regional education laboratory for the Carolinas and Virginia, topical papers and reprints, number 1. *Evaluation Comment*, *1*(2), 1–12. https://eric.ed.gov/?id=ED053419

Bun, M. J. G., & Harrison, T. D. (2018). OLS and IV estimation of regression models including endogenous interaction terms. *Econometric Reviews*, *38*(7), 814–827. https://doi.org/10.1080/07474938.2018.1427486

Campbell, J. L., Quincy, C., Osserman, J., & Pedersen, O. K. (2013). Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research*, *42*(3), 294–320. https://doi.org/10.1177/0049124113500475

Clymer, J. B., & Wiliam, D. (2007). Improving the way we grade science. *Educational Leadership*, *64*(4), 36–42. https://eric.ed.gov/?id=EJ766296

Covington, M. V., & Omelich, C. L. (1984). Task-oriented versus competitive learning structures: Motivational and performance consequences. *Journal of Educational Psychology*, *76*(6), 1038–1050. https://doi.org/10.1037/0022-0663.76.6.1038

Dweck, C. S. (2007). *Mindsets: The new psychology of success*. Ballantine.
Farrington, C. A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T. S., Johnson, D. W., & Beechum, N. O. (2012). *Teaching adolescents to become learners. The role of noncognitive factors in shaping school performance: A critical literature review*. University of Chicago Consortium on Chicago School Research. Retrieved March 18, 2013, from http://ccsr.uchicago.edu/publications/teaching-adolescents-become-learners-role-noncognitive-factors-shaping-school.
Farrington, C. A., & Small, M. (2008). *A new model of student assessment for the 21st century*. American Youth Policy Forum. https://www.researchgate.net/profile/Camille-Farrington/publication/253449779
Haley, J. M. (2015). *To curve or not to curve? The effect of college science grading policies on implicit theories of intelligence, perceived classroom goal structures* and self-efficacy [Unpublished doctoral dissertation]. Boston College. http://hdl.handle.net/2345/bc-ir:104165
Jury, M., Smeding, A., & Darnon, C. (2015). First-generation students' underperformance at university: The impact of the function of selection. *Educational Psychology*, *6*, 710. https://doi.org/10.3389/fpsyg.2015.00710
Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, *30*(4), 28–37. https://doi.org/10.1111/j.1745-3992.2011.00220.x
Kuznetsova, A., Brockhoff, P., & Christensen, R. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26. https://doi.org/10.18637/jss.v082.i13
Leahy, S., Lyon, C., Thompson, M., & William, D. (2005). Classroom assessment: Minute by minute, day by day. *Educational Leadership*, *63*(3), 19–24. https://eric.ed.gov/?id=EJ745452
Marzano, R. J. (2010). *Formative assessment & standards-based grading*. Marzano Research Laboratory.
May, H., Perez-Johnson, I., Haimson, J., Sattar, S., & Gleason, P. (2009). *Using state tests in education experiments: A discussion of the issues* (NCEE 2009-013) National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
Mills, V. L., & Silver, E. A. (2018). Putting it all together and moving forward: Concluding thoughts. In E. A. Silver & V. L. Mills (Eds.), *A fresh look at assessment in mathematics teaching*. National Council of Teachers of Mathematics.
Murphy, R., Roschelle, J., Feng, M., & Mason, C. A. (2020). Investigating efficacy, moderators and mediators for an online mathematics homework intervention. *Journal of Research on Educational Effectiveness*, *13*(2), 235–270. https://doi.org/10.1080/19345747.2019.1710885
National Governors Association Center for Best Practices, Council of Chief State School Officers. (2010). *Common core state standards*. National Governors Association Center for Best Practices, Council of Chief State School Officers.
Posner, M. A. (2011). The impact of a proficiency-based assessment and reassessment of learning outcomes system on student achievement and attitudes. *Statistics Education Research Journal*, *10*(1), 3–14. https://eric.ed.gov/?id=EJ925274 https://doi.org/10.52041/serj.v10i1.352
Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology*, *61*, 101860. https://doi.org/10.1016/j.cedpsych.2020.101860
Sanchez, C. E., Atkinson, K. M., Koenka, A. C., Moshontz, H., & Cooper, H. (2017). Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *Journal of Educational Psychology*, *109*(8), 1049–1066. https://doi.org/10.1037/edu0000190
Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*(6), 110–114. https://doi.org/10.2307/3002019
Scarlett, M. H. (2018). "Why did I get a C?": Communicating student performance using standards-based grading. *InSight: A Journal of Scholarly Teaching*, *13*, 59–75. http://files.eric.ed.gov/fulltext/EJ1184948.pdf https://doi.org/10.46504/14201804sc
Schneider, J., & Hutt, E. (2014). Making the grade: A history of the A–F marking scheme. *Journal of Curriculum Studies*, *46*(2), 201–224. https://doi.org/10.1080/00220272.2013.790480

Senko, C. (2016). Achievement goal theory: A story of early promises, eventual discords, and future possibilities. In K. R. Wentzel & D. B. Miele (Eds.), *Handbook of motivation at school* (2nd ed., pp. 75–95). Routledge. https://www.researchgate.net/profile/Corwin-Senko/publication/312591307

Smeding, A., Darnon, C., Souchal, C., Toczek-Capelle, M.-C., & Butera, F. (2013). Reducing the socio-economic status achievement gap at university by promoting mastery-oriented assessment. *PloS One*, *8*(8), e71678. https://doi.org/10.1371/journal.pone.0071678

Supovitz, J. A., Ebby, C. B., Remillard, J., & Nathenson, R. A. (2018). Experimental impacts of the ongoing assessment project on teachers and students. CPRE Research Report # RR 2018-1. Consortium for Policy Research in Education. https://eric.ed.gov/?id=ED593465

Tapia, M., & Marsh, G. E. (2004). An instrument to measure mathematics attitudes. *Academic Exchange Quarterly*, *8*, 16–22. Downloaded 8/30/2021 from http://www.rapidintellect.com/AEQweb/cho25344l.htm

U.S. Department of Education. (2020). *What works clearinghouse procedures handbook, v4.1*. Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse. Downloaded 8/30/2021 from https://ies.ed.gov/ncee/wwc/handbooks

U.S. Department of Education. (2022). *What works clearinghouse procedures and standards handbook, v5.0*. Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, What Works Clearinghouse. Downloaded 12/15/2022 from https://ies.ed.gov/ncee/wwc/handbooks

Walton, G. M., & Yeager, D. S. (2020). Seed and soil: Psychological affordances in contexts help to explain where wise interventions succeed or fail. *Current Directions in Psychological Science*, *29*(3), 219–226. https://doi.org/10.1177/0963721420904453

Webb, N. L. (2002). Depth-of-knowledge levels for four content areas. *Language Arts*, *28*, 1–9. http://facstaff.wcer.wisc.edu/normw/All content areas DOK levels 2032802.pdf.

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement and motivation. *Contemporary Educational Psychology*, *25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

Wiliam, D. (2011). *Embedded formative assessment*. Solution Tree.

Yeager, D. (2017). Social-emotional learning programs for adolescents. *The Future of Children*, *27*(1), 73–94. https://doi.org/10.1353/foc.2017.0004

Yeager, D. S., Lee, H. Y., & Dahl, R. E. (2017). Competence and motivation during adolescence. In A. J. Elliot, C. S. Dweck, & D. S. Yeager (Eds.), *Handbook of competence and motivation: Theory and application* (pp. 431–448). The Guilford Press. https://www.researchgate.net/publication/337856583

Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: They're not magic. *Review of Educational Research*, *81*(2), 267–301. https://doi.org/10.3102/0034654311405999

Zuur, Alain, Ieno, E. N., Walker, N., Saveliev, A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. Springer. https://doi.org/10.1007/978-0-387-87458-6_1