Version of record at: https://doi.org/10.1007/s11251-025-09773-0

Instructional Science (2026) 54:18
https://doi.org/10.1007/s11251-025-09773-0

ORIGINAL RESEARCH

Supports for preservice teachers' metacognitive monitoring and explanations of fraction comparisons

Taylor-Paige Guba[1] · **Christina Areizaga Barbieri**[1] · **Amanda Jansen**[1] · **Anne Morris**[1]

Received: 23 September 2024 / Accepted: 15 December 2025
© The Author(s) 2026

Abstract
The ability to provide clear and comprehensive instructional explanations on mathematical concepts is key to supporting student learning in the classroom. This is especially important when teaching content that students most often struggle with. A large body of literature has established the challenges that elementary students face with developing fraction magnitude concepts. More recent work points to the struggles that both pre- and in-service teachers have with fraction magnitude. Though much work exists on interventions for students in supporting their fraction understanding, there is a dearth of literature on how to support teachers' provisions of high quality explanations. We designed and evaluated the effectiveness of three instructional supports on elementary pre-service teachers' (PSTs) revision and improvements of their explanations of fraction concepts. We also examined their metacognitive monitoring and calibration skills as they relate to the revision process. Elementary PSTs were instructed to identify the larger of two fractions and then explain and justify their choice. They had an opportunity to revise after examining different instructional supports depending on their condition. We found that PSTs who were provided with instructional supports that provided a correct and complete exemplar of a fraction magnitude comparison showed greater improvements in their explanations than those who studied example explanations that displayed common misconceptions and errors. Correct and complete exemplars also supported greater detection and correction of errors and omissions in their explanations than those provided with a list of steps focused on structural aspects of the explanation. Hedges' g effect sizes ranged from medium to large ($g = 0.64$ to 0.95) for the three comparison tasks. Most PSTs also showed limitations in calibration, indicating over-confidence in the degree of completeness of their explanations at both time points. These findings can inform instructional materials design for pre-service teacher mathematics content courses as they learn how to monitor their own thinking and explain fraction comparisons to students.

Keywords Metacognition · Fractions · Pre-service teachers · Math education · Revision

Taylor-Paige Guba
guba@udel.edu

[1] University of Delaware, Newark, DE, USA

Published online: 08 January 2026

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

Introduction

There is increased emphasis on supporting K-8 students' abilities to "construct viable arguments and critique the reasoning of others" (CCSSI, 2010, p. 6). It stands to reason that a teacher's own skill in explanation is necessary to support that of their students. Lachner and Nückles (2016) studied and compared the nature and effectiveness of mathematical explanations offered by mathematics teachers and mathematicians. They found that mathematics teachers' explanations focused primarily on the steps required to reach the solution (i.e., product-oriented), whereas mathematicians' explanations also included the *reasoning* behind *why* taking specific steps to solve effectively leads to the solution (i.e., process-oriented). The authors found that students who learned with process-oriented explanations outperformed students who learned with product-oriented explanations. They also found that mathematical content knowledge impacted their ability to provide process-oriented explanations. Ball et al. 's (2008) *Math Knowledge for Teaching* (MKT) domains include a subskill referred to as *Specialized Content Knowledge.* This is knowledge of content that is essential for teachers to know while teaching students, but this knowledge is not necessarily representative of what the student will learn from the teacher (Ball et al., 2008). Preservice teachers (PSTs) receive considerable instruction and support throughout teacher education programs to prepare them to teach their future students. However, even with taking multiple classes designed to enhance their mathematics conceptual understanding, PSTs may need more support in their ability to both provide correct and complete mathematical explanations as their ability to evaluate students' explanations. This may be even more important in relation to their teaching of math content that has been historically challenging for students, such as fraction understanding.

A large body of literature highlights the struggles that children have with fractions in general as well as fraction magnitude understanding in particular (see Newton et al., 2020 for a review), which is often traced back to the whole number bias (Siegler et al., 2011). These challenges are not only experienced by elementary grade students. Copur-Gencturk (2022) examined fraction magnitude understanding in U.S. in-service teachers who were teaching fractions in grades 3–7. Teachers demonstrated only partially accurate representations of fraction magnitude and teachers of earlier grades were less accurate than those of later grades. As fraction magnitude concepts are most commonly introduced in the third grade in the U.S., this finding is particularly concerning as these teachers will be tasked with elementary students' first exposure to fraction magnitude concepts. To prepare PSTs to support their future students' explanations and conceptual understanding of mathematical concepts (e.g., developing conceptual understanding of fractions, including magnitude judgements), we must develop effective instructional practices and materials with this goal in mind. However, there is no published research to date that experimentally assesses the effectiveness of different instructional materials on PSTs' mathematical explanations or their skill in evaluating explanations. In the current study, we evaluate the effectiveness of different instructional materials on changes in PSTs' explanations of a mathematical concept that their future elementary students are likely to struggle with and that some work has suggested is still a challenge into adulthood: fraction magnitude comparisons.

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

Students' and preservice teachers' efforts to compare fractions

Being able to compare the magnitude (or relative size) of fractions poses several challenges. The *whole number bias*, or students' overreliance on whole number principles to work with fractions (DeWolf & Vosniadou, 2015; Siegler et al., 2011; Vamvakoussi & Vosniadou, 2010), can be used to explain some of these initial difficulties. Moreover, when determining the size of a fraction, the size of the denominator cannot be considered in isolation of the numerator or vice versa. Rather, both components need to be considered simultaneously to make accurate magnitude judgements. For example, larger numbers in one part of a fraction do not always signify larger overall values (e.g., $\frac{1}{5}$ is less than $\frac{1}{3}$). The ability to judge the size of a fraction seems to support student learning in other areas of mathematics such as pre-algebra and algebra concepts (Booth et al., 2014; Booth & Newton, 2012), and some of this work points to struggles that even PSTs have with these types of magnitude judgments.

PSTs need to understand fractions to effectively teach their future students (Olanoff et al., 2014). The ability to competently and flexibly compare fraction magnitudes requires deep knowledge of content, the ability and willingness to invent nonstandard reasoning strategies, and the ability to assess the validity of a variety of comparison strategies (Smith, 1995; Whitacre et al., 2019; Whitacre & Nickerson, 2016). For example, in Smith's (1995) study of middle school and high school students' fraction comparisons, he identified strategies that students invented and used to successfully compare fractions that contrasted with a common denominator approach. One strategy involved identifying a "reference point" or benchmark that was used in two ways: (a) using a reference number as a point *between* the two fractions to show one fraction was larger than the reference point and the other was smaller (reference point as benchmark, such as $\frac{2}{5}$ is smaller than $\frac{1}{2}$ and $\frac{3}{4}$ is larger than $\frac{1}{2}$, so $\frac{3}{4}$ is larger than $\frac{2}{5}$) or (b) using a reference number that was larger or smaller than both fractions to identify which of the two fractions was closer or further from that reference point (distance to or from a reference point, such as $\frac{3}{4}$ is further away from 1 than $\frac{4}{5}$, so $\frac{4}{5}$ is larger than $\frac{3}{4}$). Students also "transformed" the fractions to compare them, not only transforming the fractions to have a common denominator, but also transforming the fractions to have common numerators that they could use to then compare fractions ($\frac{3}{4}$ is larger than $\frac{3}{5}$ because $\frac{1}{4}$ is larger than $\frac{1}{5}$).

As these examples illustrate, PSTs need to be able to evaluate the validity of a wide variety of student invented strategies. Assessing the validity of a strategy involves assessing both the correctness of the mathematical content and the completeness of the reasoning. Studies indicate many preservice and in-service teachers lack the requisite content knowledge for teaching mathematics (Ball, 2012; Ma, 2010). PSTs tend to rely on standard procedures rather than reasoning flexibly with nonstandard procedures (Olanoff et al., 2014). PSTs also need opportunities to invent and engage with nonstandard reasoning strategies (Whitacre & Nickerson, 2016), and have difficulties differentiating between valid/complete and invalid/incomplete mathematical arguments, reasoning, and justifications (Harel & Sowder, 2007; Morris, 2007).

Reasons for PSTs' struggles with understanding fraction concepts are likely similar to hindrances experienced by elementary students when learning about fractions, such as *whole number bias* (Siegler et al., 2011; Van Steenbrugge et al., 2014; i.e., treating numerators and denominators as whole numbers contributing to the misunderstanding that $\frac{1}{4}$ is greater than $\frac{1}{3}$ because the whole number 4 is larger than the whole number 3). For

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

example, in the current study we asked PSTs to provide detailed explanations of magnitude judgements made on fraction comparison tasks. Although K-12 students may encounter similar tasks, the demonstrated knowledge required for a PST's answer to this task is on a higher level of complexity due to the mathematics knowledge needed to teach. A satisfactory elementary student answer may consist of a simple explanation such as *" $\frac{5}{8}$ is bigger than $\frac{5}{9}$ because the numerator is the same and so the one with the smaller denominator is the bigger fraction."* However, a future educator should develop the ability to provide more detailed and in-depth explanations for the purposes of supporting student understanding. One such explanation may be,

> The numerators are equal, which means the fractions both contain an equal number of portions (or pieces or units of measure), so we can compare the two denominators to determine which is larger. Because the denominator represents the size of each portion, I know that $\frac{5}{8}$ is larger than $\frac{5}{9}$ because $\frac{1}{8}$ is a bigger quantity compared to $\frac{1}{9}$. Therefore, $\frac{5}{8}$ is larger than $\frac{5}{9}$ because 5 portions of a larger quantity is a larger amount than 5 portions of a smaller quantity.

This more elaborated explanation provides a deeper and more explicit explanation of the component concepts involved in understanding fractions that may support students in making logical connections. It is especially important for PSTs to have a strong understanding of rational number magnitudes so that they can support their future students in overcoming a whole number bias when it is present.

In addition to understanding fractions, being able to explain them to their future students, and being able to evaluate student conceptual understanding, PSTs must also be able to evaluate their own level of competence and knowledge in mathematical tasks. To date, there is limited published research on PSTs' *metacognitive monitoring* (i.e., evaluating what you know and do not know) capabilities and *calibration* (i.e., how well you estimate your level of knowledge) while performing math tasks. However, the ability to identify one's own errors and omissions in explanations is an essential competency for a future educator to have when planning how to explain new concepts to students and help them address their own mistakes. We argue that, in addition to identifying one's own errors and omissions in explanations, future teachers should also consider their students' understanding of the explanations that they provide. In other words, reflective and effective teachers are able to take the perspective of the student in judging whether their explanation is complete. In the current study we aim to compare varying supports for PSTs with this end goal in mind.

Metacognitive monitoring, math achievement, and preservice teachers

Evaluating the flaws in one's own work and revising that work is a form of metacognitive monitoring. Metacognitive monitoring deals with one's ability to *be aware and in control of one's own cognitive processes* and the self-monitoring and regulation of cognitive processes (Flavell, 1976; Rhodes, 2019). Metacognitive skills are specifically important in the domain of mathematics because these skills facilitate key problem-solving steps (e.g., planning, checking work, and revising; Desoete & De Craene, 2019; Veenman et al., 2005; Zhao et al., 2019). These skills develop naturally throughout childhood and seem to take full form by the age of 12 (Veenman et al., 2005), however, even adults may lack critical

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

metacognitive monitoring skills (Pennequin et al., 2010a, b). Luckily, research has shown that improvements in metacognitive monitoring can be encouraged through offering cues or hints to remind the participant to utilize metacognitive monitoring techniques (Veenman et al., 2005). Metacognitive monitoring is an especially useful skill in the domain of mathematics according to Desoete and De Craene (2019). One reason for this could be calibration efficiency (i.e., a student's ability to recognize when they do or do not know an answer) which may lead to checking their work or asking for extra help.

We are concerned with an ability beyond *being aware and in control of one's cognitive processes*. Our study investigates the extent to which a person can correct their own errors and omissions when provided with guided and scaffolded instructional supports. This *assessment and correction* (i.e., revision) ability is an essential skill for learners, as simply identifying an issue without being able to correct it is likely not helpful to learning. This skill seems to be particularly important to develop in future reflective teachers who will be required to support students in refining their own misunderstanding and naive misconceptions of mathematics in particular (Kramarski, 2008). As a teacher, an awareness of one's own knowledge state could be helpful in determining when to seek help such as outside resources or additional training to support their ability to address student needs. Additionally, teachers act as metacognitive role models and can help their students develop essential metacognitive skills by modelling metacognitive processes out loud in the classroom (Wall & Hall, 2016).

Revision thinking

One instructional practice that may help students reason mathematically is promoting revisions. When we write, we make our thinking and reasoning public, which affords opportunities to revisit and revise our thinking and reasoning (cf., Moschkovich, 2024). Revising can support success in problem solving because revising occurs when students evaluate and improve their solutions (e.g., Verschaffel et al., 1999; Zimmerman, 2000). However, students may not necessarily engage in revising if they are not explicitly asked to do so. In a study of fifth and sixth graders, students who solved mathematics problems successfully incorporated revising when problem solving, although across the sample, evidence of engaging in revising in their problem solving process was scarce and sporadic (García et al., 2019).

Combining revision thinking with direct confrontation of mathematical content for which students commonly have misconceptions may be a particularly fruitful use of the approach if used in combination with a method that directly targets the incorrect or naive conception through either provision of correct thinking and modeling or through confrontation and resolution of common errors. However, when asked to revise, students tend to comply, but their revisions are not always improvements. When writing mathematical arguments, in a study of high school students, after the students revised arguments in response to peer feedback, there was an increase in the *number* of justifications that students provided in support of claims, but these justifications were not always mathematically *correct* (Conner & Krejci, 2022). In Prasad's (2020) self-study of her work to incorporate specifications grading in an undergraduate mathematics course, undergraduates received feedback that they either passed a writing assignment or had to revise it. However, their revisions were usually only minor improvements (Prasad, 2020). Although metacognitive monitoring and calibration

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

were not explicitly measured in this particular study, it is possible that the request for revisions alone was not enough to provoke proper monitoring and that this fault is what resulted in students not meaningfully revising their work. Thus, although engaging in the process of revising can increase the likelihood of success, revising alone is not enough. The current study combines revisions with three distinct instructional supports to encourage students to deeply consider and revise their work when comparing fraction magnitudes.

Overview of the current study

Many students struggle to learn fractions and many teachers struggle to teach fractions (Siegler & Pyke, 2013; Whitacre et al., 2019). Many teachers themselves also need support in refining their knowledge of fraction concepts such as fraction magnitude (Copur-Genturk 2022). Teachers are in need of more supports to help them foster their students' development of fraction knowledge. Though there is a substantial body of work on how best to teach fraction magnitude judgments to elementary and middle school students in particular (Barbieri et al., 2020; Dyson et al., 2020; Fuchs et al., 2017), the study of how best to support teachers in their learning and teaching of fractions is more limited. Continuing to study how best to teach fractions to children without also studying how best to teach future teachers to support student learning of fractions is akin to providing only half of a solution. Without this knowledge, we lack the crucial evidence needed to determine effective practices for improving PSTs' explanations and understandings of fraction magnitude comparisons.

Specifically, it is unknown how different instructional methods can help metacognitive monitoring. This goes beyond identifying which instructional supports help lead students to a correct answer, but rather focuses on which instructional supports lead students to evaluate their thinking and conclude where their explanations need improvement (i.e., revision). Additionally, it is unknown whether and how metacognitive monitoring can improve explanations of math concepts (i.e., fractions). If metacognitive monitoring is promoted, PSTs will be more willing and able to determine how they can improve their understanding and instruction on such concepts. Not examining and assessing ways to continuously improve the professional learning opportunities for future teachers may lead us to lack the crucial knowledge needed to improve their training as well as the learning opportunities afforded to their future students.

We conjectured there are many ways to support elementary pre-service teachers' revising of their mathematical explanations, but we focus on three in the current study. We note each here and then provide supporting literature for each conjecture below. One method would be (a) to provide them with a fully correct and complete explanation that they could compare and contrast with their own explanations, which would be a proxy for getting direct feedback from their instructor on their explanation (see Epstein, 2024; Loch et al., 2014) and is similar in design to correct examples within the worked examples literature (e.g., Durkin et al., 2023). Another is (b) to provide them with criteria for evaluating their explanations based on key components needed for a rational argument (see Yee et al., 2018). A third is (c) to analyze common errors made by a fictitious learner of the content that may highlight a conceptual hurdle that one must overcome to truly understand the content, modeled

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

similarly to incorrect or erroneous examples (e.g., Barbieri & Devlin, 2024). Each of these conjectures are supported further below.

First, we support the conjecture that providing learners with criteria for a strong explanation may help to support both metacognitive monitoring and revisions. Students can be taught criteria for evaluating strengths and weaknesses in their work. Teaching students such criteria can serve as a cognitive apprenticeship (Collins et al., 2018) for learning to self-evaluate and revise. A cognitive apprenticeship occurs when teachers make their thought processes visible and explicit to students and then students have opportunities to try to adopt and enact them. Yee and colleagues (2018) studied undergraduate students in mathematics education courses. In their study, the students struggled to revise their arguments to align with class-developed criteria for high quality arguments (Yee et al., 2018). Epstein (2024) found that when undergraduates discussed their mathematics writing with their professor in small group meetings, their revisions were stronger than when students revised their work without these discussions. These students' self-evaluations indicated they were aware of how their proofs aligned with or diverged from criteria for valid proofs, but they struggled to create improved proofs through revising (Epstein, 2024). However, as previously stated, seeing an example of a correct explanation to contrast with one's own explanation may further support students' deep revision thinking. For example, in this same study, when undergraduate students received explicit feedback on mathematics problems via screencasts (i.e., "live" feedback), their revised solutions were improved over their original solutions (Loch et al., 2014).

We also conjectured that revising an explanation could be supported by analyzing an error in another person's explanation, taking what they learned from that analysis, and applying it to analyze and improve their own explanations. That is, when considering the completeness of an explanation, it may be critical to ensure the most challenging components of the concepts are included and addressed in the explanation itself. If learners are shown what is ineffective in an explanation, or knowledge of what does not work (Gartmeier et al., 2008), they may be more likely to avoid such errors in their own work. This is related to the negative knowledge literature and has been found to be an important skill for problem-solving more generally (Gartmeier et al., 2008). Recent work (Barbieri et al., 2021) has demonstrated that an instructional tool which included prompting Algebra I students to study and explain examples of fictitious students' work, some of which included errors, increased learners' ability to anticipate errors in later work. However, the impact of reflecting on errors has not been studied in relation to revision by PSTs more generally.

Here we note that another component that may be considered an instructional feature to support learning is the act of explaining itself. Literature on self-explanation prompts lends support to this claim. Self-explanations prompt students to explain a strategy or approach and have been shown to reduce errors (Siegler & Chen, 2008). Aural self-explanation prompts are effective (e.g., Roy & Chi, 2005), and written prompts are also beneficial when paired with worked examples (Berthold et al., 2009). Classic work in self-explanation reveals that the quality of self-explanations reflects the depth of learning (Renkl, 1997); high-quality explanations focus on deep, principled features rather than superficial ones.

In the current study, PSTs explain either their own reasoning or that of a fictitious student both before and after examining varying instructional supports. The focus of the current study is on the effect of the experimental manipulation or instructional support that is presented between drafts or explanations and so we view the explanations provided as a

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

reflection of PSTs understanding of the concepts assessed. However, it is very likely that the act of explanation at both first draft and revision helps PSTs to deepen their knowledge of the concepts of study. As all three conditions were prompted to explain their magnitude comparison, it is unlikely that explanation prompts would be the cause of our differential findings by condition. However, it is possible that the improvements found in explanation completeness were not solely due to the instructional scaffolds between drafts but perhaps the combination of a draft or explanation of magnitude comparisons combined with or followed by the varying supports used.

The current study investigates the effectiveness of three possible instructional supports based on prior research on scaffolding strategies as well as current practices in the authors' university courses on rational numbers for PSTs. We investigate not only the effectiveness of each instructional support at improving explanation completeness in fraction comparison tasks, but also their effectiveness at sparking metacognitive monitoring and improving calibration over time. To study metacognitive monitoring and calibration, we must evoke reflection of one's thought processes while engaging in an instructional or learning task. To do so, in this study, we invited participants to reflect upon and revise their explanations. Results will inform teacher educators' instructional decisions to best prepare PSTs to teach their future students how to reason mathematically and to identify and reflect on incomplete or incorrect conceptions.

In summary, the current study has several aims. First, how does the completeness of a PST's fraction comparison explanation vary when they are provided with three different types of instructional supports aimed at improving their judgements and explanations of fraction magnitude comparisons? Thus, we investigate (1) the effect of three different instructional scaffolds on the accuracy and completeness of pre-service teachers' revisions to their explanations of fraction comparisons. We determine which instructional supports promote PSTs revisions of their explanations for fraction comparison tasks that make them more complete.

Second, within two of these conditions that ask PSTs to reflect upon *their own* mistakes, omissions, or ambiguities in their initial reasoning (i.e., our operationalization of metacognitive monitoring), do PSTs show differential detection by condition? That is, does this form of metacognitive monitoring vary depending on whether PSTs study supports that provide correct examples of judgements and explanations or supports that focus more on the schema for providing a correct and complete explanation? Thus, we 2a) examine how calibrated teacher candidates are in their judgments of their own explanations and we 2b) determine whether varying instructional supports foster differential changes in calibration between their first draft explanations and their revised explanations.

Third, we 3a) examine PSTs' metacognitive monitoring on this series of fraction comparison tasks, or the extent to which they successfully detect mistakes, omissions, or ambiguities in their own initial reasoning after studying instructional supports and 3b) determine whether this varies by the specific support used.

Within our experimental approach, we can determine which type of instructional supports are most effective at yielding more correct and complete explanations about fraction magnitude comparisons. We can also observe differences in PST's reflections on their revisions and omissions based on condition differences. Finally, we can get a better understanding of whether one type of support fosters stronger calibration and awareness of gaps in

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

knowledge or explanations and whether this awareness helps to explain condition differences in explanations from first draft to revision.

Methods

Participants

Participants include PSTs ($N = 87$) enrolled in an elementary teacher education program with the School of Education at a Predominantly-White Institution (PWI). More specifically, it is a privately governed, state-assisted land-grant research University in the Mid-Atlantic Region of the United States. Our sample was representative of the University's student body which is predominantly white, traditional college-aged, and female. However, we did not have access to student-level demographic data. We do have access to the demographic information of all students enrolled in the School of Education during one semester of the current study. It is important to note that these demographics are not sample-specific but rather represent the student body of the program at the time of data collection. According to institutional records, undergraduate students enrolled in the School of Education were 93.50% female and 6.50% male. The racial and ethnic breakdown of the student body at the time of study was 86.62% white, 1.53% Black, 1.34% Hispanic, 1.91% Asian, and 8.6% Other. The total number of students enrolled in the course sections from which students were recruited was 184; 47.3% of the enrolled students participated in the study. All participants provided informed consent, and all study procedures were approved by the university's Institutional Review Board (Study ID # 1924371-1).

Students in the Elementary Teacher Education program at the participating university are required to take three mathematics courses specifically designed for future teachers in grades K – 8. These courses include one focused on number and operations, another on rational numbers and probability, and a third on geometry, algebra and measurement. Participants in the current study were enrolled in the course focused on rational number concepts. More specifically, the course content focused on developing *Mathematics Knowledge for Teaching* (Ball et al., 2008) related to rational number concepts focusing on developing conceptual understanding of fractions, meanings underlying rational number operations, and proportional reasoning. Participants completed the study tasks outside of class but after receiving instruction in fraction concepts within this course. They were thus familiar with fraction comparisons as they had prior exposure to the comparison strategies in the tasks we designed.

Procedures

The current study uses a 2 (within, time point) $\times$ 3 (between, condition) design. PSTs were exposed to varying instructional supports designed to provoke revisions of their explanations of fraction magnitude comparisons. We provided three different types of instructional supports for PSTs to use while contrasting with their own explanations that could support them to improve their explanations with a revision: (1) a fully correct and complete explanation to compare with theirs, (2) a list of specific features of a correct and complete explanation to use to evaluate their own explanation, and (3) a fictitious student's explanation that

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

had two errors to have in mind when revising their explanation. We wondered if students needed a fully correct and complete explanation to support their revisions or if a list of features of a correct and complete explanation would be enough to support revising. The student error was considered to be a support because the PST could revise their explanation to provide instruction to the student with an error. Thus, our participants were practicing explaining fraction comparison and evaluating and revising their own explanations after encountering these supports.

PSTs were randomly assigned to one of three instructional support conditions to scaffold their analysis and revision of their explanations for fraction comparisons: (1) Correct examples (participants were given a complete and correct explanation of the fraction comparison at hand; CE), (2) Defining Features (participants were given a list of features and guidelines outlining what a correct and complete explanation should include; DF), and (3) Error Reflection (participants were given an incorrect explanation of the fraction comparison that showcased common misconceptions; ER). Participants across all three conditions completed a fraction comparison task. Participants completed three fraction magnitude comparison tasks before studying these supports and were offered the opportunity to revise their response on each task. This task included three pairs of fractions for which they had to determine which of the two fractions had a larger magnitude or whether they were equal. They also had to provide a written explanation to support their choice. They completed these fraction comparisons both before (i.e., first draft) and after (i.e., revision) being exposed to one of three condition-specific instructional support (described in the Materials section below). The only difference between the DF and CE conditions were the instructional supports presented. The ER condition additionally differed in that participants were asked to identify errors in fictitious students' answers, not their own answers. Participants were also asked to reflect on their omissions or refinements between first draft and revision. Participants also reported their confidence in their responses provided in their first drafts and revisions.

Materials

Fraction comparison tasks

PSTs completed three paper-and-pencil tasks in which they were asked to compare three pairs of fractions: $\frac{7}{8}$ and $\frac{6}{7}$ (Task A); $\frac{5}{8}$ and $\frac{5}{9}$ (Task B); $\frac{13}{27}$ and $\frac{18}{35}$ (Task C). The order of tasks completed was partially counterbalanced to control for order effects. Fraction pairs were chosen because they evoke three important fraction comparison strategies which are described below.

All comparison strategies taught in the course emphasized a measurement interpretation of the fraction $\frac{m}{n}$. In the measurement interpretation of $\frac{m}{n}$, the denominator represents a unit of measure of size $\frac{1}{n}$ that is created by dividing the quantity that represents the numeral 1 into *n* equal pieces and the numerator tells us how many units of size $\frac{1}{n}$ that we have. The fraction $\frac{m}{n}$ thus represents *m* units of size $\frac{1}{n}$.

We asked students to compare $\frac{7}{8}$ and $\frac{6}{7}$ to evoke the *distance-to-one strategy*. In this strategy, the reasoner determines which fraction is larger by determining how far away each fraction is from one, and then compares the two distances to identify the larger fraction. For example, $\frac{7}{8}$ is $\frac{1}{8}$ away from 1, and $\frac{6}{7}$ is $\frac{1}{7}$ away from 1. Since $\frac{1}{8}$ is less than $\frac{1}{7}$, $\frac{7}{8}$ must

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

be larger than $\frac{6}{7}$. We asked students to compare $\frac{5}{8}$ and $\frac{5}{9}$ to evoke the *common numerator strategy*. In this strategy, one determines which fraction is larger by comparing the relative sizes of the unit fraction represented by their denominators (e.g., $\frac{1}{8}$ and $\frac{1}{9}$), because each fraction had the same number of pieces (5). We asked students to compare $\frac{13}{27}$ and $\frac{18}{35}$ to evoke the *benchmark-to-one-half strategy* in which one determines which fraction is larger by first determining which fraction is less than one-half and which fraction is greater than one-half and then uses the transitive property to make their final conclusion (e.g., if $a < b$ and $b < c$ then $a < c$). The PSTs were exposed to these strategies in their course to expand their fraction comparison strategies beyond the common denominator strategy.

Instructional supports by condition

Experimental instructional supports varied by condition (Fig. 1 for the study design and procedure). The instructional support procedure proceeded as follows. In all conditions and on each task, for their first drafts, students were asked to identify the larger fraction. They were then prompted to explain how they knew that was the larger fraction. Next, they were asked to rate their confidence in their comparison and then in their explanation on a Likert Scale (1: Not at all confident; 5: Very confident). Only the end-points of the scale were marked with descriptors. All task requirements up until this point were identical across all tasks and all conditions. After rating their confidence in their explanation, this completed their first drafts. After their first drafts, condition experiences varied. That is, after this point, the condition-specific instructional supports were presented per task.

Correct examples condition

For the CE condition, to scaffold their analysis of their first drafts and support their revisions, participants were presented with a correct and complete explanation of the fraction comparison task for each of the three fraction pairs after they completed their first drafts. The correct and complete explanation included the correct answer, a complete explanation of how the correct answer was obtained, and a visual diagram that represented each fraction with an equally partitioned rectangle (fraction strip) to use to visually compare the two fractions. The instructional support was marked as correct with a checkmark. This condition was designed to mirror an in-class instructional situation when a teacher would explain and model a correct and complete example explanation for the class.

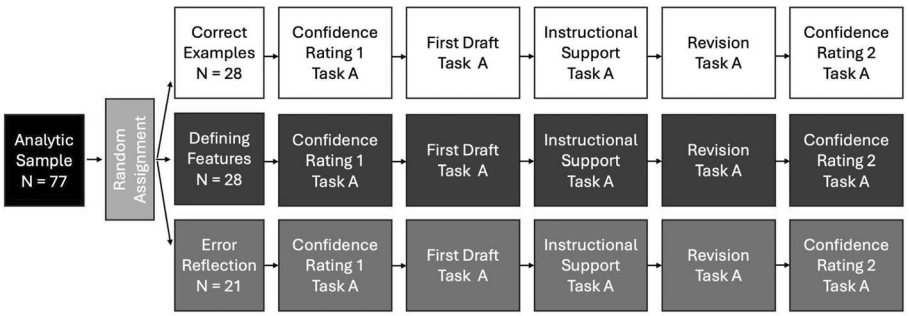

Fig. 1 Study design

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

After participants were asked to think about how their first draft explanation compared to the correct and complete explanation that was provided, they were then prompted to answer, "Were there any differences between your explanation and the explanation above? If so, list and describe them below. Number each difference." After answering this question, participants were then asked to rewrite their explanation below if they wanted to make it more correct or complete. If they did not want to revise their explanation, they were asked to simply copy their first draft explanation. This was done to confirm that participants understood the task and to ensure that they wouldn't be encouraged to choose to skip revising to save time. After providing their revised explanation (or a copy of their original explanation), participants were asked to highlight any changes they made from their original explanation and label each revision with the corresponding number in their numbered list of omissions or differences. Finally, they were asked to provide a confidence rating on their revised explanation.

Defining features condition

For the DF condition, participants were presented with a list of steps on how to completely explain the fraction comparison for each of the three fraction pairs. This instructional support did not include a final answer, a full explanation, or a visual of the fraction comparison. Rather, it served as a more abstract set of guidelines meant to trigger students' prior knowledge to accurately complete the task. This condition was designed to model an instructional situation when a teacher would make explicit to students a set of criteria to use to determine a high-quality explanation.

After participants were asked to think about how their first draft explanation compared to the example shown, they were then prompted to highlight and explain any features they omitted in the same manner that the CE condition was asked. Then the revision prompts and confidence ratings followed and they were identical to the CE condition. Finally, the same confidence rating system described above was used.

Error reflection condition

For the ER condition, participants were presented with an explanation of the fraction comparison from a fictitious student for each fraction pair with one incorrect component and one omission in the explanation. Aside from the errors and omissions in these explanations, they looked largely similar to the instructional supports shown in the CE condition. These explanations were clearly indicated as incorrect with an X. Errors displayed were errors students commonly make when working with fractions (e.g., ignoring or misinterpreting the measurement unit, not using equal-sized parts, using benchmarks incorrectly, etc.). This condition was designed to reflect an opportunity for PSTs to learn by analyzing (fictitious) students' work. It is important to note that the explanations and diagrams provided were representative of the errors students make when comparing fractions. However, the specific wording of the explanations was not designed to perfectly mirror student language. That is, it is possible that students may explain the prompted comparisons in the ways that we presented them. However, our focus in designing the materials was primarily to make clear to the PSTs the common errors that students make rather than to perfectly align language used with children's typical language use. Thus, clarity was deemed more important in the design

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

of these materials than age-appropriate language in explanations (Fig. 2 shows examples of instructional supports by condition).

After participants were asked to think about how their first draft explanation compared to the example shown, they were then prompted to answer, "List and describe ANY incorrect and/or missing statements in the student's work above. Number each concern." After answering this question, participants were asked to rewrite their explanation below if they wanted to directly address the fictitious student's incorrect thinking and then provide a confidence rating in the revised explanation. They could also choose not to revise and copy their original explanation. They then proceeded to report their confidence rating on the revised explanation.[1]

Coding schemes

Each research question was answered by a unique coding scheme that allowed us to operationalize (1) explanation completeness, (2) calibration, and (3) metacognitive monitoring based on process data derived from participant responses to different prompts within the instructional materials.

Explanation completeness

To determine to what extent teacher candidates' revisions to their explanations of fraction comparisons vary by the type of instructional support they were exposed to, we created a measure of explanation completeness score for participants' first draft and revision by task. We operationally defined explanation completeness as whether the explanation included three key concepts for the explanations for each task. Each concept consisted of sub-components. For example, for Task B, the three key concepts were (1) recognizing and beginning to execute the common numerator strategy, (2) comparing unit fractions and explaining why one is larger than the other, and (3) connecting the reasoning presented in step 2 to

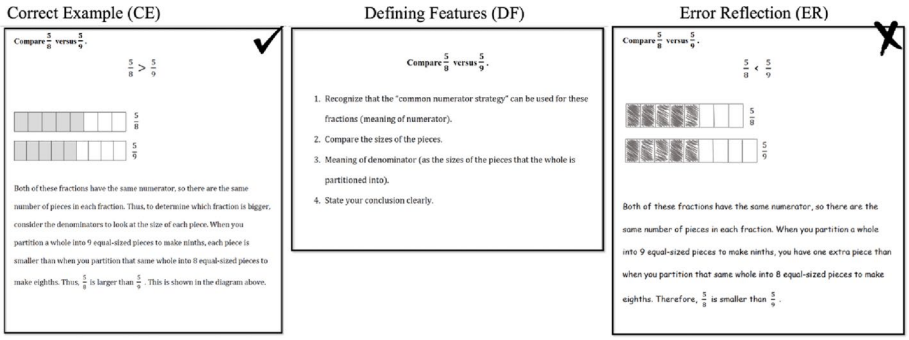

Fig. 2 Instructional Supports by Condition. *Note* All instructional supports for each condition for each task can be found in the supplemental file

[1] Note that PSTs were not asked to reflect upon their own errors and omissions in this condition and were instead prompted to reflect on the fictitious students' errors and omissions. Therefore, though we expect PSTs in this condition to revise their explanations based on this information, we do not expect this form of reflection to impact metacognitive monitoring in particular. Further exploration of PSTs' reflections on fictitious student errors is presented in Gesuelli et al. (in prep).

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

the final answer. See Table 1 for Task B's coding scheme. The number of subcomponents used to assess addressing each of the three main concepts varied by task. The rationale for this is that some fraction comparisons require more complex explanations than others. For example, comparison of $\frac{5}{8}$ vs. $\frac{5}{9}$ may be addressed by the common numerator strategy which is arguably simpler than the benchmark to $\frac{1}{2}$ strategy as it focuses primarily on comparing two denominators to each other. The benchmark to $\frac{1}{2}$ strategy requires comparing each of the two fractions to a third fraction ($\frac{1}{2}$) and that is more appropriate for the $\frac{13}{27}$ vs. $\frac{18}{35}$ comparison. However, even if sub-components varied by task, each task had three key concepts that were specifically assessed for explanation completeness.

Coding scheme development and revision

An independent coder (blind to the study hypotheses) completed all coding. A second independent coder coded 30% ($n=27$) of the sample. The double-coding procedure was used in the development and refinement of the coding scheme. That is, first, an initial coding scheme was developed by authors and implemented by two independent coders (non-authors) to determine replicability and transparency of the codes. In this initial round of coding, coders had a high level of disagreement with only 60% of their codes matching. After this initial round of coding, the coding scheme was further revised and clarified. Coders received additional training and practice rounds with feedback. Discrepancies were resolved through discussion and another round of coding was completed. Training was successful as the discrepancy rate was much lower in this second round of coding. As each

Table 1 Sample codes for task B (Compare 5/8 and 5/9)

Key component	Explanation piece	Example of response rating 2	Example of response rating 1
Part 1: recognizing and beginning to execute the common numerator strategy	Explains meaning of numerator reflecting same number of pieces in these fractions.	"Which means that the fractions have the same number of pieces."	"The numerator is the number of pieces"
	Recognizes that denominators can be compared to determine sizes of the fractions.	"This means I can determine the relative sizes of the two fractions by comparing the denominators."	"So I can compare eighths and ninths"
Part 2: comparing unit fractions and explaining why one is larger than the other	Recognizes which piece is larger - Compares sizes of remaining pieces.	"1/8 > 1/9"	"Eighths are larger than ninths"
	Explains rationale for comparison of 1/8 > 1/9	"…because when you partition a whole into 8 pieces, each piece would be larger than if you partitioned a whole into 9 pieces."	Either "…because the more pieces you partition a whole into, the smaller each piece." OR Draws a diagram showing two same sized wholes partitioned into eights and ninths.
Part 3: connecting the reasoning presented in step 2 to the final answer	Connects prior rationale to conclusion by explaining that having more pieces that are larger results in a larger fraction.	"Therefore, 5/8 > 5/9 because 5 pieces of a larger size is a greater quantity than 5 pieces of a smaller size."	"Since 5/8 has bigger pieces than 5/9, it is the larger fraction."

Note All coding schemes for all tasks can be found in the supplemental materials

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

component code could be scored on a 0–2 scale, our composite scores by Task and Overall (collapsed across all three tasks) were continuous measures. Thus, we relied on percent agreement and percent disagreement to assess our transparency and replicability of the finalized coding scheme. Our final double-coded subset reached a high level of agreement. To compute this, we converted our 0–2 scores to a range of 0–100% completeness. We then found the difference in percent completeness for First Draft between Coder 1 and Coder 2 as well as for Revision between Coder 1 and Coder 2. At First Draft, the average discrepancy in completeness scores was 3.73% (which translates to a difference of 0.075 on a scale of 0–2). At Revision, the average discrepancy in completeness scores was 0.01% (which translates to a 0.013 on a scale of 0–2). The second author observed that disagreement was lower at Revision likely because explanations were more complete, thus leaving less room for disagreement between coders. That is, it is easier to disagree on a scale of 0–2 when explanations are vague or brief. However, longer and more precise responses were more in line with the sample correct responses in our coding scheme and this left less room for error. The second author reviewed the discrepancies and determined that Coder 1's scoring was slightly more aligned with the finalized coding scheme than Coder 2. However, both coders were highly accurate. We also note that a 0.013 difference and a 0.075 difference on a scale of 0–2 are not meaningfully different. In sum, the two coders had a high level of agreement with the average percent agreement of 98.13% on first draft and revision explanation coding. Coder 1 completed coding for the entire dataset and these codes were used in the final data analysis.

We now explain the codes used in the current study. First, each comparison was coded by whether the correct fraction was chosen, and the target strategy was used in students' explanation. Strategies other than the ones targeted were not coded or used in subsequent analyses. We made this decision so that we could use a consistent coding scheme for each task that would enable us to compare explanations across participants and across conditions. At first draft, 72% ($n=63/87$) of the sample used the targeted strategy for Task A (i.e., Distance-to-one), 94% ($n=82/87$) of the sample used the targeted strategy for Task B (i.e., Common Numerator Strategy), and 62% ($n=53/85$) of the sample used the targeted strategy for Task C (i.e., Benchmark-to-Half). At revision, 100% of participants used the targeted strategy for Task A (i.e., Distance-to-one) and C (i.e., Benchmark-to-Half) and 99% used the target strategy for Task B (i.e., Common Numerator Strategy). Chi-square analyses revealed that target strategy use did not vary by condition at first draft or revision. Thus, although there was some variability in rates of target strategy usage, each target strategy was the most common strategy used in our sample. The analytic sample was reduced to $n=77$ to allow us to focus in on changes in explanation completeness from first draft to revision when using the target strategy. Thus, 8% ($n=10$) of the sample was not included in the coding of this work. Usage of target strategy did not vary by condition for any of the three tasks (Task A: χ^2 [2] = 0.274, $p=.872$; Task B: χ^2 [2] = 2.886, $p=.236$; Task C: χ^2 [2] = 4.857, $p=.088$). As there were no systematic differences by condition, this does not impact interpretation of our results in relation to condition effects. Further, we note the purpose of the current analysis and the coding scheme that was developed to investigate whether and how our learning materials differentially supported refinement of the use of the targeted strategies and so those who did not use the targeted strategies would not be able to be coded using the same coding scheme. Creating new coding schemes for those students' work would move beyond the scope and research aims of the current manuscript. However, future work with

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

broader aims may investigate the full range of strategies used regardless of alignment to instructional materials.

For each sub-component, responses were rated as 2 (the maximum value) if it included the target sub-component in the explanation and it was correct and complete, 1 if it included the target sub-component but needed improvement (e.g., vague, incomplete, incorrect), and 0 (the minimum value) if it did not include the target sub-component in the explanation. The scores for the sub-components within each key concept were averaged for a score out of 2 for each key concept. The key concept scores were then averaged for a total score of explanation completeness out of 2. This was computed for each task. Then, completeness scores for all three tasks were averaged at first draft and at revision to derive an overall explanation completeness score by participant at first draft and at revision.

Calibration

To determine how calibrated PSTs were in their judgments of their own explanations of fraction comparisons, we compared their completeness scores on each task to their confidence ratings of their explanations (i.e., *How confident are you that your explanation is correct and complete*?). Confidence ratings ranged from 1: Not at all Confident to 5: Very Confident. Values 2, 3, and 4 were not labeled, but we interpreted them to be 2: Not Confident, 3: Neutral Confidence, and 4: Confident. Participants were generally confident with mean confidence ratings of M = 4.28 (SD = 0.56) at first draft and M = 4.51 (SD = 0.60) at revision, collapsed across all tasks and conditions. Confidence ratings were broken into three categories: *Underconfident* (meaning they were less confident than their explanation was complete), *Accurate* (meaning their confidence matched their explanation completeness), and *Overconfident* (meaning they were more confident than their explanation was complete. Completeness scores ranged from 0 to 2 with 2 being the most complete. We developed a coding scheme to approximate what confidence rating would be generally accurate based on completeness score and developed the following benchmarks. We determined that completeness scores ranging from 0 − 0.5 should likely align with "1. Not at all Confident". Completeness scores ranging from 0.5 to 1 should align with "2–3: Not confident to Neutral". Completeness scores ranging from 1 to 2 should align with "4–5: Confident to Very Confident". If a participant's confidence level was within the range for their matching completeness score, we coded them as "accurate". If their confidence level was below what their completeness score should align with, we coded that as "underconfident". If their confidence level was above what their completeness score warranted, we coded that as "overconfident". We developed these ratings for all three tasks at first draft and revision in the CE and DF condition only.

Metacognitive monitoring

To determine the extent to which PSTs successfully detected mistakes, omissions, or ambiguities in their own initial reasoning after studying instructional supports, we devised a coding scheme for each task. The coding schemes were created by breaking down each task into component parts where each component part represents a possible omission or mistake. Then, we used their responses to the prompt, "Were there any differences between your explanation and the explanation above (i.e., CE supports or DF supports)? If so, list

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

and describe them below" to develop a measure of metacognitive monitoring. That is, we interpreted them as a reflection of their awareness of the differences between their first draft and the instructional support.

The coding schemes for metacognitive monitoring were developed iteratively using a subsample (27%) of the participants in the CE and DF conditions. Reflections were first independently reviewed by two of the authors of the current manuscript to produce a list of potential codes. Then, two authors conducted three rounds of coding and revision until codes accurately represented PSTs' metacognitive monitoring strategies and 100% agreement between the two coders was reached.

For each component code, responses were coded as 1 if the PST realized their error, omission, or ambiguity or 0 if the PST did not note the mistake. For example, in Task A ($\frac{7}{8}$ vs. $\frac{6}{7}$), the CE and DF conditions either displayed (CE) or summarized (DF) the steps for using the distance-to-one strategy. Thus, two component metacognitive monitoring codes ("distance of first fraction" and "distance of second fraction") were used to determine whether the PST detected that in their first explanation, they did not include a note about the distance of each fraction ($\frac{7}{8}$ and $\frac{6}{7}$ respectively) from one. This was a common omission in their first draft explanations. A sample metacognitive monitoring coding scheme is presented in Table 2 with the codes for Task A.

Composite metacognitive monitoring scores that reflect PSTs' awareness of their own accuracy in reasoning were calculated by comparing their omissions in their first draft explanations to their written reflection of these omissions as revealed in the coding. That is, each PST has a score for each task that reflects the proportion of omissions, mistakes, or ambiguities they explicitly noted in their written reflections of their first draft explanations compared to the omissions, mistakes, or ambiguities displayed in their first draft explanations. For example, if a PST omitted ten pieces of the explanation but identified eight of these omissions during the metacognitive monitoring process, they would receive a score of 0.8. The overall score was computed by collapsing across all three tasks.

Table 2 Metacognitive monitoring coding scheme for task A (Compare 7/8 and 6/7)

Code name	PST realizes that in their first explanation they:
Wrong Fraction	Chose the incorrect fraction (6/7).
Distance to 1	Did not include a note about the fractions being one unit fraction (or one piece) less than 1 (one whole).
Distance of First Fraction	Did not note the distance of the first fraction from 1.
Distance of Second Fraction	Did not note the distance of the second fraction from 1.
Comparison of Distances	Did not note which distance is larger.
Partition Rationale	Did not note the rationale for comparison of $1/8 < 1/7$
Equal-sized-pieces	Did not note need for equal-sized pieces in rationale.
Same-whole	Did not note need for same whole (MU) in rationale.
Connected Rationale	Did not connect prior rationale to conclusion using distance to 1 by relating the unit fractions back to the original fractions.
Different Method	Did not use the distance to 1 strategy.
Language	Did not use precise language.
No Difference	Noted no difference between the support and their first draft explanation.

Note Each component was coded for whether the error/omission/ambiguity was noted (1 = mentioned; 0 = not mentioned). All coding schemes for all tasks can be found in the supplemental materials

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

Results

Data analytic plan

The data analytic plan is described below by research question. Thus, we first remind the reader of our research questions.

1. To what extent do PSTs' revisions to their explanations of fraction comparisons vary by the type of instructional support they are exposed to?

2.

a. How calibrated are PSTs in their judgements of their own explanations of fraction comparisons?
b. Do varying instructional supports foster differential changes in calibration between PSTs' first draft explanations and their revised explanations?

3.

a. To what extent do PSTs successfully detect mistakes, omissions, or ambiguities in their own initial reasoning after studying instructional supports?
b. Does this metacognitive monitoring vary by the specific support used?

To determine the differential effects of instructional supports on improvements in explanation completeness from first draft to revision, we conducted a 3 (CE, DF, ER) × 2 (First Draft or Revision) mixed (or split-plot) ANOVA on explanation completeness scores which were computed as average completeness across all tasks at first draft and at revision by participant. A mixed ANOVA, sometimes referred to as a split-plot ANOVA, is a general linear model similar to a combination between a repeated measures ANOVA (RMANOVA) and a one-way ANOVA (Murrar et al., 2018). The mixed ANOVA explains variance in the dependent variable between conditions, within conditions, and, most importantly, between conditions over time. The main focus of a mixed ANOVA is the Time × Condition interaction, which examines whether differences over time in the dependent variable were significantly different between conditions. This analysis is appropriate for our design as assignment to condition was completed between-subjects. That is, each participant experienced only one condition. However, each participant also completed a first draft and a revision. This is the time component of the analysis handled by the within-subjects component of the mixed ANOVA. We acknowledge here that there were three tasks that each had a first draft and a revision opportunity for a total of six measurement opportunities over the study session. However, as the administration order of the tasks were counterbalanced to reduce the influence of order effects, we would not be able to treat these statistically as time points one through six, for example. Thus, our time factor in the mixed ANOVA focuses on first draft and revision as the two overall time points of interest. The standard assumptions that must be met prior to conducting a mixed ANOVA include normality and homogeneity of variance for the between-subjects factor. Each of these assumptions were met and are detailed in the supplemental files.

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

As per APA 7 guidelines (American Psychological Association, 2020), we also report effect sizes for all comparisons made with continuous scores. We use Hedges' g to estimate the standardized mean differences between conditions and these are displayed in Table 5. Hedges' *g* is a measure of effect size that indicates the standardized difference between two means. It is similar to Cohen's d, but includes a correction for small sample sizes to reduce bias or overestimation of effects. The formula for Hedges' g is the difference between two means (or two mean differences in the case of repeated measures designs), divided by the pooled standard deviation and then multiplied by the correction factor J. This is displayed as $g = \frac{M_1 - M_2}{SD_{pooled}} \times J$. This correction or scaling factor is calculated as $J = 1 - \frac{3}{4N-9}$, where N is the total sample size. Hedge's *g*, which is in standard deviation units, is interpreted in the same way that Cohen's *d* is, with 0.2 being a small effect, 0.5 being a medium effect, and 0.8 being a large effect.

To examine how calibrated PSTs are in their judgments of their own explanations and to determine whether varying instructional supports foster differential changes in calibration between their first draft explanations and their revised explanations, we relied on the calibration categories derived from first drafts and revisions. Frequencies in calibration categories, as described in the methods, are reported overall as well as by condition with corresponding chi-squared analyses to determine whether there are condition differences.

The third aim was to examine PSTs' metacognitive monitoring on this series of fraction comparison tasks, or the extent to which they successfully detect mistakes, omissions, or ambiguities in their own initial reasoning after studying instructional supports and determine whether this varies by the specific support used. To address this aim we examined the proportion of omissions PSTs detected when reflecting on the differences between the instructional supports and their first drafts overall and then ran a one-way ANOVA to compare these omission scores by condition. Descriptive statistics are presented in Table 3. Prior to conducting analyses, continuous data were screened for normality.

Aim 1: differential effects of instructional supports on explanation completeness

First, we conducted a one-way ANOVA on first draft explanation completeness by condition to establish baseline balance. This ANOVA revealed that conditions varied in their explanation completeness at first draft ($F[2, 83] = 7.26$, $p = .001$). Follow-up Tukey's HSD comparisons revealed that students in the ER condition had lower completeness scores at first draft than both the CE condition and the DF conditions ($p = .010$ and $p = .002$, respectively). First draft completeness was included in further analyses to account for this difference.

Participants were more likely to improve their explanation completeness if they were in the CE condition or the DF condition rather than those who were in the ER condition. A 2 (Time) × 3 (Condition) mixed ANOVA on explanation completeness revealed a signifi-

Table 3 Descriptive statistics by condition for explanation completeness and metacognitive monitoring

Measure	Correct examples (CE)		Defining features (DF)		Error reflection (ER)	
	Mean	SD	Mean	SD	Mean	SD
First draft explanation completeness	0.98	0.24	1.03	0.28	0.79	0.30
Revision explanation completeness	1.16	0.40	1.17	0.28	0.759	0.39
Metacognitive monitoring (proportion)	29.10%	21.00%	11.14%	15.90%	n/a	n/a

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

cant Time × Condition interaction, $F(2, 74) = 3.48$, $p = .036$. Tukey's HSD post hoc analyses revealed that both CE and DF showed significantly greater improvements than ER (Mean Differences = 29.82% and 32.04%; $p = .001$ and $p < .001$ respectively). CE and DF did not significantly differ in their improvements of explanation completeness from first draft to revision (see Fig. 3) and each had medium effect sizes ($p_{CE:\ DF} = 0.954$; $g_{CE} = 0.54$; $g_{DF} = 0.53$).

Aim 2: calibration of teacher candidates' judgment of their explanation completeness

To address Aim 2, we report frequencies in calibration categories below by condition and by task in Table 4. First, we established that the calibration category at first draft was correlated with revision completeness at revision, $r(75) = 0.45$, $p < .001$. As the data were categorical, we analyze these non-parametrically with a series of chi-squared analyses with Bonferroni corrections for multiple tests. Second, we compared whether the proportion of participants who were underconfident, overconfident, and accurate varied by condition on each task at both first draft and revision. No apparent differences were statistically significant (all ps > 0.05; see Table 4). Then, we analyzed the data as a whole, collapsing across conditions, to determine whether the proportion of participants who were underconfident, overconfident, or accurate significantly varied for each task from first draft to revision. Very few participants were categorized as being underconfident. Comparison of these proportions indicates that there are general improvements from first draft to revision in terms of more participants moving from overconfident to accurate in their judgements. There was a large change in Task A from first draft to revision. That is, 59% of PSTs in the CE condition were accurate at first draft and 85% were accurate at revision. There was over a 25% increase in accuracy rates for CE while there was only a 10% increase in PSTs accuracy rates in the DF condition on Task A. The movement from overconfident to accurate was generally

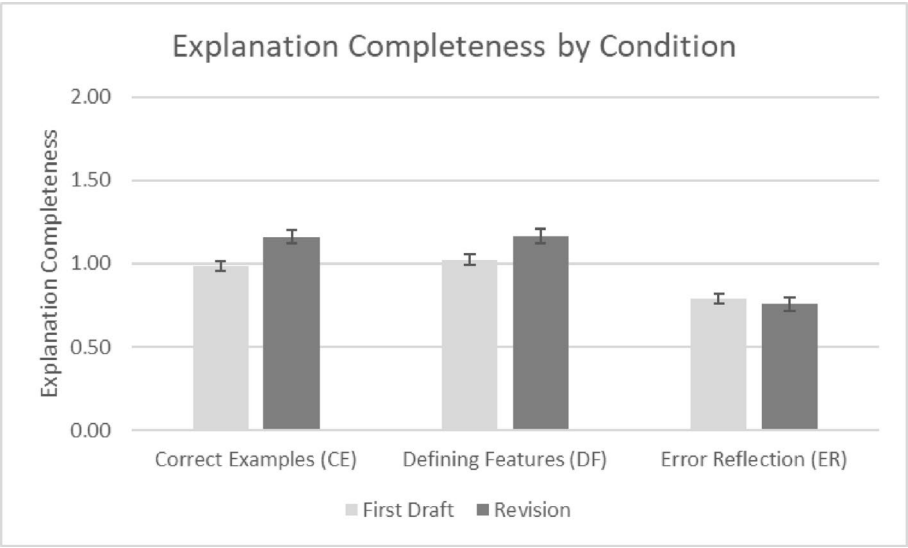

Fig. 3 Explanation completeness by condition for first draft and revision

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

consistent across Tasks B and C for both conditions. Here, although we observed that the proportion of participants that were accurate increased by about ~ 20% and ~ 10% for Tasks A and B respectively, and the proportion of participants that were overconfident decreased by about 14% for both Tasks A and B, these improvements did not reach statistical significance (see Table 4). Although statistical significance was not reached, general trends in these preliminary data suggest that calibration is malleable and may be improved over time with repeated opportunities to revise.

Aim 3: PSTs metacognitive monitoring

To determine whether and to what extent PSTs successfully detected mistakes, omissions, or ambiguities in their own initial reasoning after studying instructional supports, we present descriptive statistics on the percentage of mistakes or omissions participants detected overall as well as by task in Table 5. We also report the percent of participants that reported no omissions or mistakes. This analysis explores a potential mechanism (i.e., differential improvements in metacognitive monitoring by condition) that may help to explain differences in performance between the CE and DF condition. Thus, we then compare these measures by condition to determine whether they vary. Descriptive statistics by condition and results from independent samples t-tests along with effect sizes are also presented in Table 5. A visual representation of the data is presented in Fig. 4. In the previous section, we established that the CE and DF conditions did not vary in explanation completeness on their first draft, ensuring baseline balance between these two groups. Thus, we do not control for prior explanation completeness in the comparisons by condition for Aim 3.

Table 4 Calibration categories by task and condition at first draft and revision

Calibration	Correct examples condition ($n=28$)		Defining features condition ($n=28$)		Overall ($n=56$)	
	First draft	Revision	First Draft	Revision	First draft	Revision
Task A						
Underconfident	0%	0%	5%	0%	2%	0%
Accurate	59%	85%	65%	75%	61%	80%
Overconfident	40%	15%	30%	25%	36%	20%
Chi-square results	First draft: $\chi^2(2)=1.508$, $p=.470$ Revision: $\chi^2(1)=0.625$, $p=.429$				$\chi^2(2)=3.705$, $p=.157$	
Task B						
Underconfident	0%	0%	0%	0%	0%	0%
Accurate	44%	56%	36%	46%	40%	51%
Overconfident	56%	44%	64%	54%	60%	49%
	First draft: $\chi^2(1)=0.437$, $p=.509$ Revision: $\chi^2(1)=0.484$, $p=.487$				$\chi^2(1)=1.304$, $p=.253$	
Task C						
Underconfident	0%	6%	5%	0%	3%	3%
Accurate	56%	43%	41%	46%	47%	45%
Overconfident	44%	50%	55%	55%	50%	53%
	First draft: $\chi^2(2)=1.403$, $p=.496$ Revision: $\chi^2(2)=1.417$, $p=.492$				$\chi^2(2)=0.054$, $p=.973$	

Note Percentages reported are within condition

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

Table 5 Descriptive statistics overall and by condition

	Sample (N = 57)		Correct examples (*n* = 28)		Defining features (*n* = 28)		Difference tests
	Mean	*SD*	Mean	*SD*	Mean	*SD*	t-statistics
Task A: Proportion of omissions noted	.19	.24	.26	.27	.11	.20	$t(53) = -2.41$, $p = .020$; $g = .64$ (CE > DF)
Task B: Proportion of omissions noted	.09	.25	.16	.33	.01	.05	$t(29.19) = -2.51$, $p = .017$; $g = .65$ (CE > DF)
Task C: Proportion of omissions noted	.35	.39	.52	.35	.18	.36	$t(53) = -3.58$, $p < .001$; $g = .95$ (CE > DF)
Overall: Proportion of omissions noted	.20	.21	.29	.35	.11	.16	$t(55) = -3.59$, $p < .001$; $g = .94$ (CE > DF)
Overall: Proportion of tasks on which participants noted no difference	.21	.28	.06	.13	.37	.31	$t(35.95) = 4.99$, p < .001; $g = 1.32$ (CE < DF)

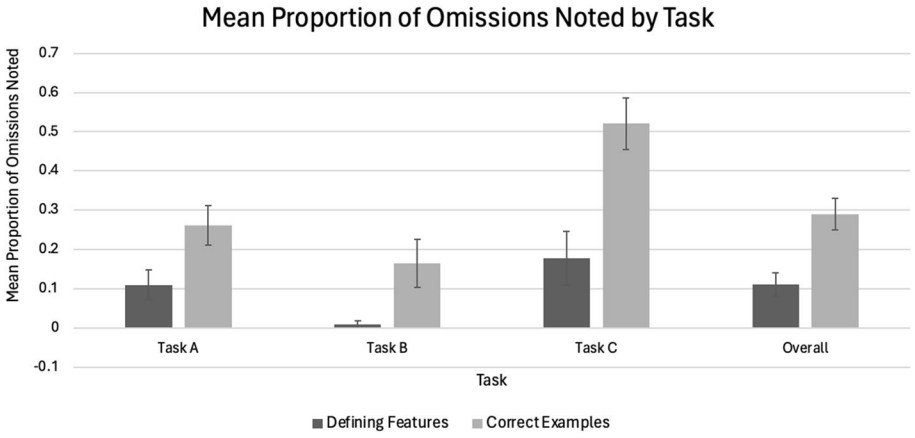

Fig. 4 Mean proportion of omissions noted by ask and overall for each condition

We conducted a series of independent samples t-tests comparing proportion of omissions noted by task and overall as well as participants' overall proportion of tasks on which they reported no difference between first draft and explanations or guidelines provided within instructional materials. Their first draft explanations and the explanations or guidelines presented in each of the conditions all revealed significant differences by condition (all $ps < 0.05$, see Table 5). Participants in the CE condition detected a higher proportion of omissions in their explanations than the DF condition by task and overall, with medium to large effect sizes ranging from Hedge's $g = 0.64 - 0.95$ depending on the measure (see Table 5). Further, a larger proportion of participants in the DF condition reported that they made no mistakes or omissions in their first draft explanations across the three tasks compared to the CE condition ($g = 1.32$). To interpret this finding it is important to remember that these two conditions did not vary in the completeness of their first draft explanations ($t[54] = -0.83$, $p = .413$). Further, no participant's first draft explanation was coded as complete (i.e., all participants omitted or made an error within at least one component of a full

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

explanation), so all participants who reported no omissions or mistakes overestimated the completeness of their first draft explanations. Thus, the perception that the DF condition had fewer omissions/mistakes than the CE condition at first draft did not reflect their true performance.

Discussion

In the current study, we provide several important insights into how PSTs can improve both their metacognitive monitoring and the completeness of their explanations regarding fraction comparisons.

Aim 1: differential effects of instructional supports on explanation completeness

In response to our first research question, which examined the extent to which PSTs' revisions to their explanations of fraction comparisons varied by the type of instructional support they are exposed to, we found that PSTs' explanation completeness improvement from first draft to revision differed by condition such that the CE condition and the DF condition were both more effective than the ER condition. These two conditions were likely more effective in yielding more complete revisions than the ER condition due to the nature of the task demands. Specifically, the ER condition was asked to revise their explanation to specifically target the error of the fictitious student displayed. The CE and DF conditions were asked to revise their explanations to be more correct and complete. Thus, these findings are understandable and more detailed analysis of the learning process in the ER condition is discussed in AUTHORS (in prep). More surprising is the finding that the CE and DF conditions were equally effective at promoting more complete explanations at revision. It is interesting that displaying a concrete example of a correct explanation is no more effective than distilling said explanation into its component parts. One could argue that the DF condition required a greater transfer of skills than the CE condition, as participants were asked to revise explanations based on principles rather than example explanations themselves. Transfer of skills is normally quite challenging for learners (e.g., Barnett & Ceci, 2002; Hajian, 2019). Yet our sample did not find the DF condition requirements more challenging than the CE condition, at least in terms of using these supports to revise their explanations. However, it is important to note that most revised explanations across conditions still had room for improvement in terms of completeness.

Aim 2: calibration of teacher candidates' judgment of their explanation completeness

Through our examination of our second research question, which examined how calibrated PSTs are in their judgements of their own explanations of fraction comparisons as well as differential changes in calibration between PSTs' first draft explanations and their revised explanations by instructional support, we found that PSTs' assessment of their ability to fully and correctly explain fraction comparison tasks was generally overconfident at first draft, meaning that they rated their confidence as high when their explanation completeness was often low. There were slight (although non-significant) improvements in calibration

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

from first draft to revision across all conditions ($g = 0.08$). However, there were no significant differences in calibration scores or improvement between conditions. Some PSTs may have limited pedagogical content knowledge, which creates challenges for them when they attempt to interpret students' errors (Maher & Muir, 2013; Runnalls & Hong, 2020). It is possible that participants needed more direct feedback on their own errors and omissions to bring about substantial change in their own calibration. Clerjuste et al. (under review) found that scaffolded feedback in the form of specifically highlighting PSTs' errors within a fraction arithmetic task was more effective at bringing about revision than corrective feedback (i.e., only noting whether the final solution was correct or not). Although they did not measure calibration specifically, they did employ revisions and study PSTs in a related mathematical content area - fraction arithmetic. Additionally, Cardelle-Elawar (1992) found that when teachers were able to identify and explain errors to their low-achieving middle school students, the students would learn better mathematical strategies rather than making the same mistake over and over again. Thus, we may speculate that PSTs need more concrete and specific information about their own performance to yield substantial improvements.

The frequent overconfidence in our sample is consistent with prior research which has shown that overconfidence is common, especially in mathematics (Erickson & Heit, 2015). Similarly, confidence judgments after performing a task are generally more accurate than those made before performing a task (Lingel et al., 2019; Rutherford, 2017) but our study yielded equally poor calibration both before and after revision of PSTs' explanations. Another reason for this disconnect may be that the model used in this study does not use both predictions (i.e., confidence rating before a task) and post-dictions (i.e., confidence ratings after a task). Rather we used post-dictions after the "first draft" attempt and again after "revision" attempt. Therefore, there is some precedent as to why both ratings, done after finishing a task, would be similar. Though there were no significant improvements in calibration overall, we do know that calibration accuracy was positively correlated with explanation completeness both within and between time points. These findings suggest that calibration plays a role in promoting explanation completeness such that more accurate calibration is associated with greater explanation completeness.

Aim 3: PSTs metacognitive monitoring

Our third research question, which examined the extent to which PSTs successfully detect mistakes, omissions, or ambiguities in their own initial reasoning after studying instructional supports; and whether this metacognitive monitoring vary by the specific support used, yielded findings such that a difference *was* found between the CE condition and the DF condition within our metacognitive monitoring measure. The CE condition was more effective than the DF condition in supporting metacognitive monitoring on the fraction comparison tasks. A potential explanation of this finding is that the CE condition provided an actual example of a correct and complete answer, allowing the participant to directly compare their answer to the model. In contrast, the DF condition was more abstract and did not allow this direct comparison of explanation to explanation. This suggests that the CE supports may be the most beneficial overall as they provide explanation completeness at an equivalent level to the DF condition as well as greater metacognitive monitoring. This conclusion is consistent with prior literature that suggests that instructional explanations

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

paired with worked examples is an effective way of developing conceptual mathematical knowledge (Wittwer & Renkl, 2010).

Utility and difficulty of instructional supports

It is interesting that the conditions varied in their effects differently for different outcomes, but upon further reflection, it seems that the supports are on a spectrum of both difficulty and utility. For example, CE supports would make it quite easy to identify errors (either one's own errors or a student's errors) because it exemplifies the specific components necessary to form a complete and correct explanation. On the other hand, this support is low in utility as teachers rarely have this type of specific correct example to compare student work to unless they create it themselves for every unique student task or rely only on tasks from a given curriculum with detailed teacher guides. The DF support is at the midpoint of both spectrums as it requires the PST to generalize guidelines to an actual response (in this case their own response), which may be potentially more cognitively demanding than simply comparing specific answers word for word. However, DF might be more useful than CE in that DF support could be a feasible resource for teachers to have on hand as they grade student work and can act as a sort of check list for giving feedback or partial credit. The ER condition may be the most difficult in terms of reflection because the PST must recall the defining features scaffolds (which they learned about in class rather than in the moment of the study) and then apply them to the errors of a student. In retrospect, the ER support has the potential to hold high utility for PSTs because this is the task they must perform on the job: correcting student work and providing feedback for improvement. However, PSTs might not be getting enough opportunities to interpret students' errors in their teacher education programs (Leighton et al., 2022). Thus, our conditions form both a continuum of easiest to most difficult and a continuum of least to most utility (CE, DF, and ER for both continuums).

Scaffolding instruction

Interestingly, the CE and DF conditions differed significantly in promoting metacognitive monitoring and revision, but not explanation completeness. That is, the CE condition demonstrated significantly more metacognitive monitoring than the DF condition. One possibility as to why this is the case stems from the differences in the types of instructional scaffolds used in the study. *Conceptual scaffolding* (i.e., scaffolding by emphasizing content, exemplified by the CE condition), has different prerequisites and outcomes than *heuristic scaffolding* (i.e., scaffolding by emphasizing processes, exemplified by the DF condition; Holton & Clarke, 2006). That is, conceptual scaffolding utilizes content knowledge whereas heuristic scaffolding relies on more domain-general skills. Conceptual scaffolding may be a more efficient mechanism for initiating the metacognitive monitoring and revision process in that it draws attention to what *content* must be revised more so than the heuristic scaffolding, which focuses on the *process* of crafting an explanation. Given that metacognitive monitoring scores were calculated based on identifying what specific content PSTs missed, the CE supports should be more supportive than the DF supports. However, heuristic scaffolding may be just as effective as conceptual scaffolding in eliciting a certain amount of explanation completeness given that they support different mechanisms (i.e., supporting explicit content knowledge and supporting the process of building an explanation) of producing

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

a complete answer. Given that explanation completeness was scored using a mixture of specific content and strategies/reasoning, it is reasonable that the two conditions performed equally well. Overall, Prior literature suggests that there are key steps to all metacognitive scaffolding in mathematics, namely following a strategic plan to identify what the problem is asking and interpreting relevant symbols or concepts, modeling the problem, choosing and implementing the appropriate solution, and making sure the solution satisfies the problem at hand (Oladunni, 1998).

Unexpected findings

Curiously, even though the DF condition and the CE condition were equally effective in evoking increased explanation completeness at revision, the CE condition was significantly more effective at promoting metacognitive monitoring. One may have expected more consistent findings in these two related outcomes as metacognitive monitoring may be a potential mechanism that brings about well-revised explanations. However, it may be the case that increases in metacognitive monitoring do not explain or account for the changes in explanation completeness, either overall or specifically for the DF condition. It is possible that the nature of the CE supports allowed for greater metacognitive monitoring than the nature of the DF supports due to the provision of more information for a participant to work with as they identify what may have been missing in their first draft explanation. This additional information may serve as indirect feedback on PSTs' own explanations supporting their metacognitive monitoring skills.

Another interesting and somewhat unexpected finding is that calibration was extremely limited (mostly overconfident across trials). Although we expected a high amount of overconfidence across participants based on prior work (Lingel et al., 2019), we did not expect it to be as prevalent as found here considering past research has shown large variation in calibration for math knowledge of PSTs (Ernst et al., 2023). For example, prior work suggests that post-diction (as opposed to prediction) yields more accurate confidence ratings (Lingel et al., 2019), so participants in our study were expected to follow this trend. Additionally, these participants were recruited directly from a class in which they were learning and practicing with these same materials and being asked to perform an elementary level task as college students. These findings suggest that more empirical and intervention work is needed to support PSTs' calibration. However, there are a few alternative explanations as to why the PSTs performed lower than expected on calibration. First, the PSTs may not have put in as much effort in this task as they would have on a graded assignment within a classroom or in a teaching scenario with their own students. Second, they may have been experiencing more math anxiety than they would normally experience due to the focus on fractions, as higher mathematics anxiety among PSTs is associated with poorer performance on fraction arithmetic, which suggests an association between fractions and anxiety among PSTs (Rayner et al., 2009). Math anxiety generally hinders performance and calibration (Erickson & Heit, 2015). Fractions are notably difficult for learners (Jordan et al., 2023) and are a particular stressor which may evoke more anxiety than other math topics. Thus, calibration related to fractions performance may be hindered more so than other mathematics topics and future work should examine effects in other math domains.

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

Limitations

The primary limitation of the current study is that this study was conducted over a single session and so we cannot establish the effects of the instructional supports if provided repeatedly in a classroom over time. Thus, we can only make claims about the effects of single session administration. Findings may be different with increased dosage of each support. Thus, we may be underestimating the effects of such supports. Future work should examine multiple sessions of these promising techniques on PSTs understanding of fractions and their metacognition. Additionally, we did not achieve baseline balance across conditions for explanation completeness, suggesting that random assignment did not efficiently control for individual differences prior to the treatment. We appropriately controlled for this in the analysis by including first draft explanation completeness as a covariate in all analyses. A third limitation of the study was the confidence measure used in the calculation of calibration. Traditionally, Likert scales are labeled at every interval (e.g., 1 = not at all confident, 2 = somewhat not confident, 3 = neutral, 4 = somewhat confident, 5 = very confident). However, our measure was only labeled at each end point, leaving some room for interpretation and ambiguity.

Another limitation of the current experiment is that we could not assess the effect of error reflection on calibration or metacognitive monitoring. That is, the parallel reflection prompts provided within the ER condition asked PSTs to focus on a fictitious students' understanding rather than their own as in the CE and DF conditions. This hinders us from understanding the full extent of the effects of such instructional supports. One final limitation is that this study focused on a very specific sample of predominantly white female traditional college-aged PSTs. Although these characteristics are representative of many PSTs in education programs in the U.S., it limits the generalizability to more diverse samples in terms of age, gender, and ethnicity. We recommend that future research in this area conduct a larger longitudinal experimental study such that students in mathematics content courses in PST programs across the country be evaluated using these instructional supports across a full semester, perhaps even including a condition that scaffolds the supports across the difficulty and utility continuums discussed earlier.

Strengths

There were several methodological strengths of the current study. The experimental design (randomization of participants into conditions, counterbalancing of tasks) supports the validity of results as individual differences should be balanced across conditions and there would be no differential effects of order or fatigue by condition or task. Additionally, although the randomization of participants reduced the likelihood of systematic differences between conditions, administration of a first draft allowed us the ability to account for individual differences in the analysis. Another advantage of our design is that the instructional supports used in each condition were inspired by real techniques used in PST math content classes. This means that the instructors of such courses can use these results to plan their instructional supports for future classes to best support PSTs' explanation completeness and metacognitive monitoring. Lastly, the current study used fine-grained iteratively developed coding schemes which were individualized to each task and measure to capture the qualitative differences in explanations, metacognitive monitoring, and calibration.

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

Implications

Our findings suggest that revising may be important aspects of mastering *Specialized Content Knowledge*, which is an essential part of PSTs' training (Ball et al., 2008). Given that many PSTs struggle with understanding foundational math concepts such as fractions, implementing revision thinking into PST training may improve the acquisition of *Specialized Content Knowledge* (Ball et al., 2008; Reid & Reid, 2017; Ribeiro & Jakobsen, 2012). As future materials are created to assist teachers in creating and implementing lessons on fraction comparisons, we suggest publishers include more examples that follow the patterns of the instructional supports used in our study that help evoke metacognitive monitoring, revision, and explanation completeness. This could include a list of key features of complete explanations for different comparison strategies (mimicking the DF condition) as well as several examples of correct and complete explanations for different comparison strategies (mimicking the CE condition). By including more of these types of supports for in-service teachers, we can better scaffold their metacognitive monitoring in the classroom.

Future directions

One important finding to consider is that improvements in explanation completeness were minimal in the error reflection condition which is, arguably, most similar to what PSTs may one day see in their classrooms. The ER condition showed the work of a fictitious student with errors displayed that are commonly made in fraction comparison tasks. It is possible that due to the task decision that PSTs gave more precise explanations targeted to the error and so explanation completeness was lower. However, future work may be conducted to confirm that speculation as well as consider whether that is an effective approach in addressing student errors in the classroom. That is, is providing a more precise and concise explanation of an error more beneficial for PST learning than providing a more complete explanation?

Additionally, future research could investigate in-service teachers' drafts and revisions of mathematical explanations when they have different forms of support. It is possible that the DF condition could be more informative to in-service teachers than PSTs, as it is likely that in-service teachers have higher mathematics knowledge for teaching that would align with stronger interpretations of the meanings of each of the defining features. It is also possible that an in-service teacher would produce stronger revisions in the ER condition, as a teacher with more experience is likely to have stronger mathematical knowledge for teaching aligned with interpreting students' errors due to experience with encountering them. Empirical data is needed to see if these possibilities hold.

A final recommendation for future research regarding mathematical calibration of preservice teachers is to comparatively test calibration in other domains (e.g., science or english). This comparison could allow researchers to determine if the low mathematical calibration from our study is an indication of low calibration in general or if it is specific to the domain of mathematics. Similarly, researchers could choose a variety of mathematical tasks to determine if the low calibration is indeed specific to the domain of mathematics or specific to the subdomain of fractions.

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

A final suggestion for further research is to reflect on strategies used by students who did not use the target strategies. In the current study, we excluded a small number of participants from the analysis who deviated from the expected strategies ($N = 10$) to ensure we were able to use the same coding scheme. It is interesting to note that the targeted strategies were those being taught in the course that the participants were enrolled in. Thus, choosing to not use the targeted strategy may reflect something unique about these particular individuals that may be captured in future work. That is, this data could provide important insights into how PSTs think about fraction comparisons and how they may relate to students' interpretations of these same problems. We recommend research into the range of strategies that PSTs might use, particularly in terms of what additional strengths they bring to this work.

Conclusion

The current study showed that different types of instructional supports yield different levels of metacognitive monitoring and different levels of explanation completeness in three fraction comparison tasks. Specifically, PSTs who revised their response after seeing an example of a correct and complete response improved their responses more than those who revised after seeing a list of key components to an explanation. Additionally, PSTs who saw either of the above-described supports had more complete explanations than those who saw an incorrect example response. The implications of this research can be found primarily in mathematics content courses for PSTs in which similar tasks are part of the curriculum. Through these results, mathematics teacher educators can make decisions about which instructional supports will best accomplish their goals at any given time. For example, instead of only using one type of support, they might initially give correct and complete examples and gradually fade those examples away until PSTs only need a list of key components to effectively explain the concepts involved in fraction comparison questions.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11251-025-09773-0.

Acknowledgements Aneysha De Coteau, Kelly-Ann Gesuelli, and Courtney Martin.

Funding This work was not grant-funded.

Declarations

Conflicts of interests The authors declare no conflicts of interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

References

American Psychological Association, issuing body. (2020). *Publication manual of the American Psychological Association : The official guide to APA style* (Seventh edition). American Psychological Association.

Ball, D. L. (2012). Halves, pieces, and twoths: Constructing and using representational contexts in teaching fractions. *Rational numbers* (pp. 157–195). Routledge. https://doi.org/10.4324/9780203052624-15

Ball, D., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, *59*(5), 389–407. https://doi.org/10.1177/0022487108324554

Barbieri, C. A., & Devlin, B. L. (2024). Targeting fraction misconceptions and reducing high confidence errors in an online tutor. *Journal of Computer Assisted Learning*, *40*(1), 254–268.

Barbieri, C. A., Rodrigues, J., Dyson, N., & Jordan, N. C. (2020). Improving fraction understanding in sixth graders with mathematics difficulties: Effects of a number line approach combined with cognitive learning strategies. *Journal of Educational Psychology, 112*(3), 628–648. https://doi.org/10.1037/edu0000384

Barbieri, C. A., Booth, J. L., Begolli, K. N., & McCann, N. (2021). The effect of worked examples on student learning and error anticipation in algebra. *Instructional Science*, *49*(4), 419–439. https://doi.org/10.1007/s11251-021-09545-6

Barnett, S., & Ceci, S. (2002). When and where do we apply what we learn? A taxonomy for Far transfer. *Psychological Bulletin*, *128*(4), 612–637. https://doi.org/10.1037//0033-2909.128.4.612

Berthold, K., Eysink, T. H., & Renkl, A. (2009). Assisting self-explanation prompts are more effective than open prompts when learning with multiple representations. *Instructional Science*, *37*(4), 345–363.

Booth, J. L., & Newton, K. J. (2012). Fractions: Could they really be the gatekeeper's doorman? *Contemporary Educational Psychology*, *37*(4), 247–253. https://doi.org/10.1016/j.cedpsych.2012.07.001

Booth, J. L., Newton, K. J., & Twiss-Garrity, L. K. (2014). The impact of fraction magnitude knowledge on algebra performance and learning. *Journal of Experimental Child Psychology*, *118*, 110–118. https://doi.org/10.1016/j.jecp.2013.09.001

Cardelle-Elawar, M. (1992). Effects of teaching metacognitive skills to students with low mathematics ability. *Teaching and Teacher Education*, *8*, 109–121. https://doi.org/10.1016/0742-051X(92)90002-K

Collins, A., Brown, J. S., & Newman, S. E. (2018). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In *Knowing, learning, and instruction* (pp. 453–494). Routledge. https://doi.org/10.4324/9781315044408-14

Conner, K. A., & Krejci, B. (2022). Developing secondary students' understanding of proof through constructing, critiquing, and revising arguments. *Investigations in Mathematics Learning, 14*(2), 101–116. https://doi.org/10.1080/19477503.2021.2000201

Copur-Gencturk, Y. (2022). Teachers' knowledge of fraction magnitude. *International Journal of Science and Mathematics Education*, *20*(5), 1021–1036.

Desoete, A., & De Craene, B. (2019). Metacognition and mathematics education: An overview. *ZDM, 51*(4), 565–575. https://doi.org/10.1007/s11858-019-01060-w

DeWolf, M., & Vosniadou, S. (2015). The representation of fraction magnitudes and the whole number bias reconsidered. *Learning and Instruction*, *37*, 39–49. https://doi.org/10.1016/j.learninstruc.2014.07.002

Durkin, K., Rittle-Johnson, B., Star, J. R., & Loehr, A. (2023). Comparing and discussing multiple strategies: An approach to improving algebra instruction. *The Journal of Experimental Education*, *91*(1), 1–19.

Dyson, N. I., Jordan, N. C., Rodrigues, J., Barbieri, C., & Rinne, L. (2020). A fraction sense intervention for sixth graders with or at risk for mathematics difficulties. *Remedial and Special Education*, *41*(4), 244–254. https://doi.org/10.1177/0741932518807139

Epstein, R. (2024). Revisiting past assignments: Do precalculus students learn more from revising or discussing their work? *Primus (West Point, N.Y.), 34*(2), 135–151. https://doi.org/10.1080/10511970.2023.2250764

Erickson, S., & Heit, E. (2015). Metacognition and confidence: Comparing math to other academic subjects. *Frontiers in Psychology*, *6*, 742. https://doi.org/10.3389/fpsyg.2015.00742

Ernst, H. M., Wittwer, J., & Voss, T. (2023). Do they know what they know? Accuracy in teacher candidates' self-assessments and its influencing factors. *British Educational Research Journal, 49*(4), 649–673. https://doi.org/10.1002/berj.3860

Flavell, J. H. (1976). Metacognitive aspects of problem solving. *The nature of intelligence* (pp. 231–235). Lawrence Erlbaum Associates.

Fuchs, L. S., Malone, A. S., Schumacher, R. F., Namkung, J., & Wang, A. (2017). Fraction intervention for students with mathematics difficulties: Lessons learned from five randomized controlled trials. *Journal of Learning Disabilities*, *50*(6), 631–639. https://doi.org/10.1177/0022219416677249

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

García, T., Boom, J., Kroesbergen, E. H., Núñez, J. C., & Rodríguez, C. (2019). Planning, execution, and revision in mathematics problem solving: Does the order of the phases matter? *Studies in Educational Evaluation*, *61*, 83–93. https://doi.org/10.1016/j.stueduc.2019.03.001

Gartmeier, M., Bauer, J., Gruber, H., & Heid, H. (2008). Negative knowledge: Understanding professional learning and expertise. *Vocations and Learning*, *1*(2), 87–103. https://doi.org/10.1007/s12186-008-9006-1

Hajian, S. (2019). Transfer of learning and teaching: A review of transfer theories and effective instructional practices. *IAFOR Journal of Education*, *7*(1), 93–111. https://doi.org/10.22492/ije.7.1.06

Harel, G., Sowder, L., Information Age Publishing. (2007). Toward comprehensive perspectives on the learning and teaching of proof. *Second Handbook of Research on Mathematics Teaching and Learning, 2*, 805–842.

Holton, D., & Clarke, D. (2006). Scaffolding and metacognition. *International Journal of Mathematical Education in Science and Technology*, *37*(2), 127–143. https://doi.org/10.1080/00207390500285818

Jordan, N. C., Dyson, N. I., Devlin, B. L., & Gesuelli, K. A. (2023). Developing fraction sense in students with mathematics learning difficulties: From research to practice. In K. M. Robinson, A. K. Dubé, & D. Kotsopoulos (Eds.), *Mathematical cognition and understanding: Perspectives on mathematical Minds in the elementary and middle school years* (pp. 259–280). Springer International Publishing. https://doi.org/10.1007/978-3-031-29195-1_14

Kramarski, B. (2008). Promoting teachers' algebraic reasoning and self-regulation with metacognitive guidance. *Metacognition and Learning*, *3*(2), 83–99. https://doi.org/10.1007/s11409-008-9020-6

Kramarski, B., & Zoldan, S. (2008). Using errors as springboards for enhancing mathematical reasoning with three metacognitive approaches. *The Journal of Educational Research*, *102*(2), 137–151. https://doi.org/10.3200/JOER.102.2.137-151

Lachner, A., & Nückles, M. (2016). Tell me why! Content knowledge predicts process-orientation of math researchers' and math teachers' explanations. *Instructional Science*, *44*(3), 221–242.

Leighton, J. P., Guo, Q., & Tang, W. (2022). Measuring preservice teachers' attitudes towards mistakes in learning environments. *Learning Environments Research*, *25*(1), 287–304.

Lingel, K., Lenhart, J., & Schneider, W. (2019). Metacognition in mathematics: Do different metacognitive monitoring measures make a difference? *ZDM: The International Journal on Mathematics Education*, *51*(4), 587–600. https://doi.org/10.1007/s11858-019-01062-8

Loch, B., Jordan, C. R., Lowe, T. W., & Mestel, B. D. (2014). Do screencasts help to revise prerequisite mathematics? An investigation of student performance and perception. *International Journal of Mathematical Education in Science and Technology*, *45*(2), 256–268. https://doi.org/10.1080/0020739X.2013.822581

Ma, L. (2010). *Knowing and Teaching Elementary Mathematics: Teachers' Understanding of Fundamental Mathematics in China and the United States*. Routledge.

Maher, N., & Muir, T. (2013). I know you have to put down a zero, but i'm not sure why: Exploring the link between pre-service teachers' content and pedagogical content knowledge. *Mathematics Teacher Education and Development, 15*(1), 72–87.

Morris, A. K. (2007). Factors affecting pre-service teachers' evaluations of the validity of students' mathematical arguments in classroom contexts. *Cognition and Instruction, 25*(4), 479–522. https://doi.org/10.1080/07370000701632405

Moschkovich, J. N. (2024). Academic Literacy in Mathematics to Frame Mathematical Writing Research and Practice: Writing and Revising for Mathematical Practices. *Illuminating and Advancing the Path for Mathematical Writing Research* (pp. 264–290). IGI Global. https://doi.org/10.4018/978-1-6684-6538-7.ch013

Murrar, S., Brauer, M., & Frey, B. B. (2018). Mixed model analysis of variance. *The SAGE encyclopedia of educational Research, Measurement, and evaluation* (Vol. 4, pp. 1074–1078). SAGE Publications, Inc. https://doi.org/10.4135/9781506326139.n436

Newton, K. J., Barbieri, C. A., & Booth, J. L. (2020). Key mathematical competencies from arithmetic to algebra. In *Oxford Research Encyclopedia of Education*.

Oladunni, M. O. (1998). An experimental study on the effectiveness of metacognitive and heuristic problem solving techniques on computational performance of students in mathematics. *International Journal of Mathematical Education in Science and Technology*, *29*(6), 867–874. https://doi.org/10.1080/0020739980290608

Olanoff, D., Lo, J. J., & Tobias, J. (2014). Mathematical content knowledge for teaching elementary mathematics: A focus on fractions. *The Mathematics Enthusiast*, *11*(2), 267–310. https://doi.org/10.54870/1551-3440.1304

Pennequin, V., Sorel, O., & Mainguy, M. (2010a). Metacognition, executive functions and aging: The effect of training in the use of metacognitive skills to solve mathematical word problems. *Journal of Adult Development*, *17*(3), 168–176. https://doi.org/10.1007/s10804-010-9098-3

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

Pennequin, V., Sorel, O., Nanty, I., & Fontaine, R. (2010b). Metacognition and low achievement in mathematics: The effect of training in the use of metacognitive skills to solve mathematical word problems. *Thinking & Reasoning*, *16*(3), 198–220. https://doi.org/10.1080/13546783.2010.509052

Prasad, P. V. (2020). Using revision and specifications grading to develop students' mathematical habits of mind. *Primus, 30*(8–10), 908–925. https://doi.org/10.1080/10511970.2019.1709589

Rayner, V., Pitsolantis, N., & Osana, H. (2009). Mathematics anxiety in preservice teachers: Its relationship to their conceptual and procedural knowledge of fractions. *Mathematics Education Research Journal*, *21*(3), 60–85.

Reid, M., & Reid, S. (2017). Learning to be a math teacher: What knowledge is essential? *International Electronic Journal of Elementary Education*, *9*(4), 851–872.

Renkl, A. (1997). Learning from worked-out examples: A study on individual differences. *Cognitive Science*, *21*(1), 1–29.

Rhodes, M. G. (2019). Metacognition. *Teaching of Psychology*, *46*(2), 168–175. https://doi.org/10.1177/0098628319834381

Ribeiro, C. M., & Jakobsen, A. (2012). Prospective teachers' mathematical knowledge of fractions and their interpretation of the part-whole representation. *Generalization in Mathematics at All Educational Levels*, 289–298. Wydawnictwo Uniwersytetu Rzeszowskiego.

Roy, M., & Chi, M. T. (2005). *The self-explanation principle in multimedia learning* (pp. 271–286). The Cambridge handbook of multimedia learning.

Runnalls, C., & Hong, D. S. (2020). Well, they understand the concept of area: Pre-service teachers' responses to student area misconceptions. *Mathematics Education Research Journal, 32*(4), 629–651.

Rutherford, T. (2017). The measurement of calibration in real contexts. *Learning and Instruction*, *47*, 33–42. https://doi.org/10.1016/j.learninstruc.2016.10.006

Siegler, R. S., & Chen, Z. (2008). Differentiation and integration: Guiding principles for analyzing cognitive change. *Developmental Science*, *11*(4), 433–448.

Siegler, R. S., & Pyke, A. A. (2013). Developmental and individual differences in understanding of fractions. *Developmental Psychology, 49*(10), 1994–2004. https://doi.org/10.1037/a0031200

Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology*, *62*(4), 273–296. https://doi.org/10.1016/j.cogpsych.2011.03.001

Smith, J. P., III (1995). Competent reasoning with rational numbers. *Cognition and Instruction*, *13*(1), 3–50. https://doi.org/10.1207/s1532690xci1301_1

Vamvakoussi, X., & Vosniadou, S. (2010). How many decimals are there between two fractions? Aspects of secondary school students' understanding of rational numbers and their notation. *Cognition and Instruction, 28*(2), 181–209. https://doi.org/10.1080/07370001003676603

Van Steenbrugge, H., Lesage, E., Valcke, M., & Desoete, A. (2014). Preservice elementary school teachers' knowledge of fractions: A mirror of students' knowledge? *Journal of Curriculum Studies*, *46*(1), 138–161. https://doi.org/10.1080/00220272.2013.839003

Veenman, M. V. J., Kok, R., & Blöte, A. W. (2005). The relation between intellectual and metacognitive skills in early adolescence. *Instructional Science*, *33*(3), 193–211. https://doi.org/10.1007/s11251-004-2274-8

Verschaffel, L., De Corte, E., Lasure, S., Van Vaerenbergh, G., Bogaerts, H., & Ratinckx, E. (1999). Learning to solve mathematical application problems: A design experiment with fifth graders. *Mathematical Thinking and Learning*, *1*(3), 195–229. https://doi.org/10.1207/s15327833mtl0103_2

Wall, K., & Hall, E. (2016). Teachers as metacognitive role models. *European Journal of Teacher Education*, *39*(4), 403–418. https://doi.org/10.1080/02619768.2016.1212834

Whitacre, I., & Nickerson, S. D. (2016). Investigating the improvement of prospective elementary teachers' number sense in reasoning about fraction magnitude. *Journal of Mathematics Teacher Education*, *19*(1), 57–77. https://doi.org/10.1007/s10857-014-9295-2

Whitacre, I., Atabaş, Ş, & Findley, K. (2019). Exploring unfamiliar paths through familiar mathematical territory: Constraints and affordances in a preservice teacher's reasoning about fraction comparisons. *Journal of Mathematical Behavior, 53*, 148–163. https://doi.org/10.1016/j.jmathb.2018.06.006

Wittwer, J., & Renkl, A. (2010). How effective are instructional explanations in example-based learning? A meta-analytic review. *Educational Psychology Review*, *22*(4), 393–409. https://doi.org/10.1007/s10648-010-9136-5

Yee, S. P., Boyle, J. D., Ko, Y.-Y.(winnie), & Bleiler-Baxter, S. K. (2018). Effects of constructing, critiquing, and revising arguments within university classrooms. *The Journal of Mathematical Behavior, 49*, 145–162. https://doi.org/10.1016/j.jmathb.2017.11.009

Zhao, N., Teng, X., Li, W., Li, Y., Wang, S., Wen, H., & Yi, M. (2019). A path model for metacognition and its relation to problem-solving strategies and achievement for different tasks. *ZDM : the International Journal on Mathematics Education, 51*(4), 641–653. https://doi.org/10.1007/s11858-019-01067-3

Version of record at: https://doi.org/10.1007/s11251-025-09773-0

Zimmerman, B. J. (2000). Chapter 2 - Attaining Self-Regulation: A Social Cognitive Perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of Self-Regulation* (pp. 13–39). Academic Press. https://doi.org/10.1016/B978-012109890-2/50031-7

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.