Teacher preparation *does* matter: Relationships between elementary mathematics content courses and graduates' analyses of teaching

Siobahn Suppa
Stockton University

Joseph DiNapoli
Montclair State University

Robert Mixell
University of Delaware

Received: 31 January 2017 Accepted: 8 January 2018
© Mathematics Education Research Group of Australasia, Inc.

In the United States, teacher preparation programs are under increased pressure to demonstrate their effectiveness in producing graduates with knowledge, abilities, and competencies to be quality teachers. However, very little research shows this kind of evidence. In a rare exception, Hiebert, Miller, and Berk (2017) found significant positive results of the influences of an elementary teacher education program on graduates' knowledge. Due to the rarity of these kinds of findings, we replicated their analyses with a different cohort of graduates from the same preparation program. Graduates completed a video analysis task correlated with high quality mathematics teaching for topics taught during their program and topics not taught during their program. Our results corroborate theirs, showing that graduates performed better on topics taught during their program versus topics not taught. These findings suggest that teacher education programs can have a significant positive and lasting effect on graduates' knowledge several years after graduation.

Keywords teacher education · teacher preparation · elementary mathematics · replication study · analysis of teaching

Teacher preparation programs in the United States are under increased pressure to demonstrate their effectiveness (Feuer, Floden, Chudowsky, & Ahn, 2013; Levine, 2006). National accrediting agencies, such as The Council for the Accreditation of Education Programs (CAEP), have been pressed to develop new standards that demonstrate concretely that accredited programs produce graduates with critical teaching competencies. However, there is very little evidence that attending a teacher preparation program helps prospective teachers develop the knowledge and skills needed to teach that subject effectively (Floden & Meniketti, 2005).

In fact, there is mixed evidence about the effects of teacher preparation programs. For example, there is evidence that teachers who have completed a preparation program do not perform significantly differently from teachers who have not completed a preparation program (e.g., Greenberg, McKee, & Walsh, 2013). These types of studies suggest that preparation does not make a difference in graduates' abilities to teach. Other studies suggest that teachers trained in an alternate route program, such as Teach for America (one of the most popular non-traditional preparation programs in the United States), do not perform any differently compared to teachers prepared in traditional university-based preparation programs (e.g.,

Constantine, Player, Silva, Hallgren, Grider, & Deke, 2009). These kinds of studies suggest that the type of teacher preparation does not matter, but do not address whether any preparation at all makes a difference in graduates' teaching abilities.

On the other hand, there is evidence that teachers with standard certification (graduating from university-based preparation programs) have significantly positive effects on student learning gains, while teachers with no certification being prepared in alternative route programs such as Teach for America had significantly negative effects on student learning gains (Darling-Hammond, Holtzman, Gatlin, & Vasquez Heilig, 2005). These types of studies suggest that preparation does matter and furthermore that the type of preparation makes a difference in graduates' teaching quality. Because of these contradictory results, there is mixed evidence regarding the effectiveness of teacher preparation programs.

However, much of this research relies on student performance measures, such as standardized tests, and teacher characteristics, such as teacher licensure and years of experience in the field (e.g., Darling-Hammond et al., 2005; Goldhaber et al. 2013). These types of measures indirectly measure the effect of preparation programs on graduates' knowledge. Due to the complexity of the education system, using measures such as student assessments to measure the effectiveness of a teacher preparation program is comparable to holding medical schools accountable for the multitude of issues in the healthcare system of the United States (Zeichner, 2014).

In addition, studying the vast array of teacher preparation programs in the country in such broad ways ignores the fact that the quality of preparation programs in the United States is extraordinarily variable (Greenberg et al., 2013; Levin, 2006; Zeichner, 2014). As Greenberg et al. (2013) summarize, "The explanation for why teacher preparation in the United States seems to make no impact on the whole is variability: First, in the aggregate, there are not enough high-quality teacher preparation programs; and second, their impact is diluted by the preponderance of weak programs" (p. 10). Even though states can regulate evaluation and accreditation of teacher preparation programs, the quality control mechanisms of these systems are weak (Goldhaber, Liddle, & Theobald, 2013). As Arthur Levin (2006) laments, "under the existing system of quality control, too many weak programs have achieved state approval and have been granted accreditation" (p. 61).

In sum, there is very little evidence regarding *specific* qualities of teacher preparation programs or *specific* approaches to teacher training that affect graduates' *specific* teaching competencies and capabilities (National Research Council, 2010). Even studies that do not use student performance measures to determine the quality of a teacher preparation program typically rely on broad measures, such as whether the program includes courses that train prospective teachers in how to design and plan lessons (Greenberg, McKee, & Walsh, 2013). As a field, we lack knowledge about connections between specific aspects of a teacher preparation program and measures of graduates' abilities and competencies more closely related to the direct skills of teaching.

In a rare exception, Hiebert, Miller, and Berk (2017) report that preservice teachers who studied mathematics for teaching elementary school as freshmen and sophomores demonstrated significantly better teaching skills, four to seven years later, in topics they studied, compared to those they did not study. In other words, Hiebert et al. (2017) claim that studying a specific topic in considerable depth enables graduates not only to remember the material but also to apply it to perform teaching-related tasks. If this is true, it would have significant policy implications for teacher preparation. It would mean that teacher preparation *does* matter but only for the particular topics studied in depth during preparation.

Because the findings reported by Hiebert et al. (2017) are relatively rare in teacher education research, and because these findings have significant policy implications, we believe the

findings warrant replication. Although replication has been a continuously neglected aspect of research (Schmidt, 2009; Smith, 1970; Yong, 2012), we assert that "the most defensible test of the reliability of data is provided by the replication or cross-validation study" (Smith, 1970, p. 971). Therefore, if Hiebert et al.'s (2017) findings are replicated, additional strength would be added to the argument that teacher preparation can make a difference, even years later, but only for those subject matter topics specifically studied as prospective teachers. As we carried out this study, our research team extensively consulted Hiebert and colleagues (2017) to better ensure close replication of methodology. However, our analyses and dissemination practices are completely independent of this group.

The purpose of this study is to analyse relationships between specific mathematical content taught in an elementary teacher education program and graduates' analyses of teaching using the same task administered in Hiebert et al.'s (2017) study. We focus on the relationships between several topics taught in two mathematics content courses for preservice teachers and graduates' performances analysing classroom teaching episodes pertaining to these topics. In this paper, we address the research question: *Do graduates perform better analysing teaching for topics taught during their education program compared to topics not taught during their program?*

Next, we will briefly describe the courses from the teacher education program under investigation. Following this brief description, we explain the theoretical perspective driving our analyses and our hypotheses. Because our hypotheses reference aspects of the content courses graduates experienced, we describe the content courses prior to our hypotheses.

Setting for the study

For this replication study, we report data from one cohort of the elementary teacher education program in the School of Education at the University of Delaware. This is the same setting as the work of Hiebert and colleagues (2017), but we studied a different cohort of students. In the United States, a standard mathematics preparatory program for future elementary teachers includes coursework addressing both mathematics content and mathematics pedagogy. At the University of Delaware, students in this program complete three mathematics content courses and one mathematics methods course, often during their freshman and sophomore years.

This study assesses content taught during the first two courses: whole number and decimal operations, and fractions and proportional reasoning. The third content course focuses on early algebra and geometry, but is not the focus of this study. Activities in the first two courses offer preservice teachers opportunities to explore the structure of place-valued numeration systems, model the four basic arithmetic operations using multiple representations (such as base-ten blocks, area models using graph paper, and fraction strips), and watch videos of young children solving problems to analyse and critique their solution strategies. The videos do not contain teachers helping elementary students. The curricula for these courses are based on a constructivist theory of learning (von Glasersfeld, 1995) focused on developing conceptual understanding through making connections between concepts and procedures explicit and allowing students to productively struggle (Hiebert & Grouws, 2007; Morris, 2012).

Though students in this program are spread out amongst different sections with different instructors, we are confident these students received similar learning opportunities for several reasons. First, every instructor uses the same lesson plans with the same activities, common classwork and homework problems, and common exams and grading rubrics. Moreover, the lesson plans have undergone continuous improvement (e.g., Bryk, 2015; Lewis, Perry, & Hurd, 2009) for several years and contain precise learning goals for students, rich descriptions of how students are expected to engage in the classroom activities, and detailed rationales for why the

activities are hypothesized to help students achieve the learning goals (see Morris, 2012 for an example of an improvement iteration). Second, recent empirical observation data from two first-time instructors teaching during the same semester shows that instructors provide similar learning opportunities to their students (Suppa, in preparation).

A third reason for the consistency of the learning opportunities provided by different instructors is their continual collaboration throughout the semester. The instructors of these courses typically meet once a week to discuss the upcoming lesson plans and their expectations for how each lesson should unfold. At least one experienced instructor (someone who has taught the course at least once before) typically teaches a section during each semester and thus also attends these meetings. Therefore, first-time instructors have the opportunity to ask questions regarding the lessons. As a result of these various supports for instructors, we are confident that students in different sections and across different years received similar learning opportunities.

Theoretical perspective: Knowledge needed to analyse teaching

The theoretical perspective guiding our study consists of the kinds of knowledge needed to effectively analyse and critique classroom teaching. We believe that three specific kinds of knowledge are needed to effectively analyse mathematics classroom teaching: specialized content knowledge (Ball, Thames, & Phelps, 2008), knowledge of content and teaching mathematics (Ball et al., 2008), and what we term analytical knowledge (Hiebert, Morris, Berk, & Jansen, 2007). We review the definitions of each of these next followed by our hypotheses for this study.

Specialized content knowledge (SCK) refers to mathematical knowledge and skills that are unique to teaching (Ball et al., 2008). For example, understanding a non-standard approach to solving a subtraction problem or finding an appropriate example to make a specific mathematical point requires SCK. Other professions, even those that use mathematics every day, do not typically require the kind of depth unpacking mathematical ideas that is required while teaching. This kind of knowledge is separate from pedagogical knowledge or knowledge of students and teaching. It is a form of pure subject matter knowledge that has a substantial amount of depth not required in fields other than teaching.

As Ball and colleagues describe, interpreting a student's work when completing a two-digit subtraction problem and knowing whether the student's method is mathematically sound and generalizable, and why, involves SCK. If the student made an error, understanding precisely what misconception the student likely possesses involves a deep level of pure specialized mathematical knowledge to understand the mathematical processes that likely underlie the student's thoughts. Knowing what to then say to the student is a form of pedagogical knowledge (a different type of knowledge from SCK). Mathematicians engage in this type of error-analysis frequently in *their own* work; however, teachers are required to engage in this type of analysis for *students'* work, many times during a lesson, and very quickly in order to maintain a smooth flow of instruction.

Knowledge of content and teaching (KCT) refers to "an interaction between specific mathematical understanding and an understanding of pedagogical issues that affect student learning" (Ball et al., 2008, p. 401). KCT consists of knowledge of mathematical content involved in teaching (SCK) in combination with pedagogical knowledge of teacher moves or pedagogical decisions that influence student learning. For example, understanding the advantages and disadvantages of using circles or rectangles to represent fractional quantities to develop students' understandings of fractions requires KCT.

Knowing which examples to choose to build on students' knowledge and how these examples will affect students' understanding involves KCT. For example, the problem 307 - 168 can be solved in several different ways. For instance, this problem requires "borrowing" when using the standard algorithm, which could elicit a common misconception in elementary students to subtract the digits in the wrong order (e.g., 7 - 8 = 1, 0 - 6 = 6, 3 - 1 = 2) producing the answer of 261, or misunderstanding how to "borrow" correctly and why borrowing is mathematically justified, producing an answer such as 169. Students may also develop their own methods for solving this problem, such as adding on to 168 until they reach 307 (2 + 30 + 107 = 139). Understanding these various methods and possible misconceptions requires SCK, a deep knowledge of the mathematical topic. Knowing where to place this example within the sequence of students' learning is a pedagogical decision, which requires KCT. Knowledge of which examples are appropriate to meet one's goals, how to sequence certain solution strategies, and how to choose which examples to use to deepen students' understandings all require KCT. Understanding the effects of these pedagogical decisions on students' learning is precisely what KCT captures.

Finally, *analytical knowledge* (AK) refers to reasoning skills necessary for supporting causal claims about teaching and learning (Hiebert et al., 2007). Specifically, AK enables teachers to assess whether the learning goals were achieved in a lesson (the "effect") and to develop hypotheses about why or why not (the "cause"). AK involves SCK and KCT. It involves SCK because knowledge of the learning goals and whether they were achieved requires deep knowledge about key underlying mathematical ideas unique to teaching. It involves KCT because knowledge of why the learning goals were achieved or not requires knowledge about how the pedagogical decisions of the teacher affected student learning. Therefore, we assume that there is some overlapping knowledge between AK and SCK and some overlap between AK and KCT.

These three kinds of knowledge—specialized content knowledge (SCK), knowledge of content and teaching (KCT), and analytical knowledge (AK)—comprise the foundation of our theoretical perspective. In order to effectively analyse mathematics classroom teaching, one must possess deep knowledge of the mathematical content, knowledge of how pedagogical decisions affect student learning, and knowledge of how to identify evidence that supports causal claims about teaching and student learning. Thus, we believe that this combination of knowledge is the minimum knowledge needed in order to effectively analyse mathematics teaching. Our hypotheses describe relationships between the kinds of knowledge that the content courses focused on developing and graduates' hypothesized performances analysing teaching videos.

Hypotheses

Because our hypotheses focus on the content preservice teachers studied during their preparation program and the kinds of knowledge we suspect they will exhibit when analysing classroom videos, we briefly elaborate on the learning opportunities preservice teachers experienced during their preparation program. Throughout the entire program, preservice teachers are provided with limited opportunities to analyse teaching. In the three mathematics content courses, preservice teachers are regularly given the opportunity to observe and analyse their peers' presentations and explanations of conceptual solutions. However, no explicit attempt is made by instructors to strengthen their knowledge of producing evidenced-based causal claims relating teaching to student learning. The focus of these critiques is typically on clarity of language and mathematical representations. In addition, when watching videos of

young children solving problems, the focus of the class discussion is on categorizing different ways children tend to solve problems and uncovering whether children's invented strategies are mathematically sound or not. Again, the videos do not contain teachers helping elementary students. Thus, these critiques focus more on developing SCK within preservice teachers than on AK or KCT.

In the elementary methods course, preservice teachers study pedagogical moves (Kazemi & Hintz, 2014) and number talks (Humphreys & Parker, 2015) regarding content taught in their first and second content courses. Therefore, preservice teachers analyse the types of solution strategies one example might elicit from students during a number talk and what types of teacher moves to engage in when leading a number talk. Preservice teachers have opportunities to lead number talk sessions in front of their peers inside and outside of class and observe and critique one another's pedagogical practices. They also watch videos of classroom teachers leading number talks and discuss what aspects of the teaching in the video they wish to incorporate into their own number talks. These experiences provide preservice teachers with opportunities to continue to develop SCK, begin to develop KCT, and perhaps indirectly begin to develop AK. Again, the focus of instruction in the methods course is on preparing preservice teachers to lead number talks when they enter the classroom. The focus is not on producing evidence-based causal claims regarding teaching and student learning, although preservice teachers may attempt to do this as a natural extension of these experiences.

We hypothesized that graduates would perform better analysing teaching for topics taught during their program (target topics) compared to topics not taught during their program (control topics) because we predicted graduates would have greater SCK and KCT in target topics than in control topics. We hypothesized that graduates would have greater SCK in target topics than in control topics due to their mathematics content courses focusing explicitly on developing SCK. We hypothesized that graduates would have greater KCT in target topics than in control topics due to their methods course and the pedagogy used in the mathematics content courses. Even though pedagogical decisions were not frequently discussed explicitly in the content courses, we believed preservice teachers' experiences as learners in the courses would influence their KCT.

We hypothesized that graduates would possess only slightly more AK for target topics than for control topics at best. Our main reason for this hypothesis is the fact that AK was not a focus of the graduates' teacher education program. The curriculum of their mathematics courses did not explicitly develop skills for analysing classroom teaching. Therefore, the only reason we hypothesized graduates might possess slightly greater AK for target topics than for control topics is because SCK and KCT are both required in AK. Therefore, since we believed that graduates would possess greater SCK and KCT in target topics than in control topics, we hypothesized they would be more likely to have higher AK in target topics than in control topics as well. In other words, graduates' would have a "head start" in developing AK for target topics because we predicted that they would have greater SCK and KCT, which are both prerequisites to possessing AK.

Our last hypothesis concerns how graduates would perform over time analysing teaching tasks for target and control topics. Because graduates typically study this content during their freshman year, we were unsure whether they would remember this knowledge and apply it to a teaching task several years later. However, if they were actively using this knowledge, then this may allow their knowledge to continue to develop over time. Therefore, we might expect graduates' scores on all topics to increase over time, with their scores on target topics increasing at a faster rate than control topics because of their greater initial SCK and KCT of target topics.

On the other hand, we also thought that this knowledge could deteriorate over time. Therefore, graduates might perform worse over time on target topics because they likely begin

with high initial SCK and KCT in target topics, but over time they might start to forget some of this knowledge if they are not actively using it. To conclude, we were unsure what to expect in terms of graduates' performances over time in all topics.

Methods

As previously mentioned, this study replicates Hiebert et al.'s (2017) study using analogous analyses on a different cohort of graduates from the same preparation program. For additional methodological information, please refer to Hiebert et al. (2017).

Sample

Our sample from the elementary teacher education program graduated in 2010. All 132 students in this cohort were invited to participate in a longitudinal study. We had no direct contact with the participants. Of the 59 graduates that participated in the first year (one year post-graduation), 25 participated all four years. These 25 participants comprise the sample for this study. All participants were paid to participate.

The combined grade-point average (GPA) for the two content courses of the 25 participants for this study was 3.08. For comparison purposes, the combined GPA for the remaining 107 graduates in the cohort was 2.91. Since the difference between our sample and their nonparticipating peers is not significant ($U = 1109,\ p = 0.23$), our sample should not be considered significantly more mathematically prepared than their non-participating peers.

Research design

Like Hiebert et al. (2017), we used a two-way repeated measures design so each participant would serve as his/her own control. Each graduate in our sample analysed four video clips, each concerning a different mathematical topic. Three of these topics were included in the elementary teacher education program curriculum (target topics) and one of these topics was not (control topic). The target topics are multiplying two-digit whole numbers, subtracting fractions, and dividing fractions. The control topic is finding the mean for a small set of whole numbers. All four topics are associated with a standard algorithm whose meaning derives from several underlying concepts and are considered fundamental topics in the U.S. elementary school curriculum.

Tasks. Our sample was assessed using an online video analysis task very similar to one developed and validated by Kersting and colleagues (Kersting, 2008; Kersting, Givvin, Sotelo, & Stigler, 2010; Kersting, Givvin, Thompson, Santagata, & Stigler, 2012). This type of assessment measures "teachers' knowledge of teaching mathematics in concrete teaching situations, emphasizing the contextual and situational nature of teaching" (Kersting, 2008, p. 857). These concrete teaching situations give teachers the opportunity to demonstrate their ability to critique mathematical teaching episodes, which assesses their SCK, KCT, and AK.

The only difference between our task and the one developed and used by Kersting and colleagues was the content of each video. We used the same prompts for participants to respond to each video as Kersting and colleagues. Each online video analysis task contained a brief video clip from a classroom lesson on one of the four elementary mathematics topics selected. Each clip showed a teacher interacting with students in the context of a mathematics lesson and demonstrated some deficiencies, both mathematical and pedagogical in nature, in the interaction. Since these video-based assessments have been correlated with high quality

teaching and student learning (Kersting et al., 2012), we used them to assess graduates' knowledge related to teaching.

The video clip for multiplying two-digit whole numbers captured a third-grade teacher demonstrating the standard algorithm for solving 52 × 36. The subtracting fractions video clip depicted a fifth-grade teacher modelling the problem 9/12 – 1/3 using blocks. The video clip for dividing fractions showed a sixth-grade teacher soliciting student solutions to the problem 1/2 ÷ 2/3. The video clip for finding the mean portrayed a teacher working with two students to find the number of pets each of seven families could have if the mean number of pets was four, but no family had exactly four pets. For more details about each video, please see Hiebert et al. (2017). Participants were asked to view each video clip and respond to the following prompt for each task: "View the clip and discuss how the teacher and the student(s) interact around the mathematical content." This is the same prompt used by Kersting and colleagues (Kersting, 2008; Kersting et al., 2010; Kersting et al., 2011).

Our sample completed these video tasks one, two, three, and four years post-graduation. However, the prompt during the first year of data collection was different than the remaining years and it only asked participants to consider two video clips, not four. Therefore, we disregard data from this first year for our sample and only include data from years two, three, and four, which aligned with the prompt used by Kersting and colleagues.

Coding. Responses were coded using a rubric focused on what participants noticed and critiqued in each video clip, as well as what suggestions participants made to improve each lesson. The rubric, a modified version of that used by Kersting (2008), incorporated both a pedagogical and mathematical dimension. Specifically, our rubric focused on what participants noticed, critiqued, and/or suggested about both pedagogical moves and mathematics in each video clip. The rubric also accounted for two key pedagogical moves shown to support students' conceptual understanding: allowing students to productively struggle with the mathematics, and making the key mathematical ideas in the lesson explicit for students (Hiebert & Grouws, 2007). As mentioned earlier, these pedagogical moves were characteristic of the written lesson plans of all mathematical content courses in the elementary teacher education program.

For each video clip response, two mathematical scores and two pedagogical scores were assigned, each ranging from 0 to 2 (see Appendix A for a full description and examples of each score). The first mathematical score (mathematical-descriptive) measured participants' observations of the specific mathematics in the classroom interactions. Prior to data collection, Hiebert et al. (2017) identified components of the key mathematical ideas relevant to each video clip. A score of 2 was assigned if participants described all components of the key mathematical idea at stake in the video clip. A score of 1 was assigned if participants described the mathematics present in the clip, but did not describe all components of the key idea. A score of 0 was assigned if participants did not describe any of the mathematics apparent in the video.

The second mathematical score (mathematical-critique) measured any critiques of the mathematical interactions and suggestions for improvement. All videos had room for improvement in this regard. Participants earned 2 points if they suggested ways to more clearly address the key mathematical idea at stake in the video. Participants earned 1 point if they suggested a change in the mathematics, but not about the key mathematical idea present. Participants earned 0 points if they made no suggestions for mathematical changes in the video.

The first pedagogical score (pedagogical-descriptive) assessed if participants made an inaccurate claim that the teacher's pedagogical moves directly helped students *understand* the mathematics. None of the videos contained evidence connecting teacher moves to students' understanding. A score of 2 was assigned if participants made no such claims. A score of 1 was assigned if participants claimed the teacher helped the students understand the mathematics in

some way. A score of 0 was assigned if participants claimed the teacher not only helped the students understand mathematical content, but did so by supporting students' productive struggle and/or making the key mathematical ideas explicit during instruction.

Lastly, the second pedagogical score (pedagogical-critique) assessed whether participants suggested any pedagogical changes to improve the lesson, and if these changes would help support students' conceptual understanding. All videos had room for this type of improvement. Participants earned 2 points if they suggested the teacher in the video should allow more time for students to productively struggle with the mathematics and/or make more explicit the key mathematical ideas in the lesson. Participants earned 1 point if they critiqued any teacher moves and/or made any pedagogical suggestions other than those identified as worth 2 points. Participants earned 0 points if they made no critiques or suggestions about the teacher's instructional moves.

Participant responses were coded by the second and third author. The second and third author first independently coded a random selection of 10% of the responses and compared the number of agreements to the total number of decisions. Interrater reliability was 90.0% for mathematical-descriptive, 95.0% for mathematical-critique, 97.5% for pedagogical-descriptive, and 92.5% for pedagogical-critique.

Analyses. These four coding dimensions (mathematical-descriptive, mathematical-critique, pedagogical-descriptive, and pedagogical-critique) comprise three total scores that we use in all subsequent analyses: mathematical score, pedagogical score, and critique score. A participant's mathematical score is calculated by adding his/her mathematical-descriptive and mathematical-critique scores. A participant's pedagogical score is calculated by summing his/her pedagogical-descriptive and pedagogical-critique scores. Finally, a participant's critique score is calculated by adding his/her pedagogical-critique and mathematical-critique scores. Therefore, each of these three total scores (mathematical, pedagogical, and critique) range from 0 to 4 (see Table 1).

Table 1
Summary of the three total scores and how they were calculated

Total Score	Composed of Coding Dimensions	Scoring Range
Mathematical Score	Mathematical-Descriptive + Mathematical-Critique	0 – 4
Pedagogical Score	Pedagogical-Descriptive + Pedagogical-Critique	0 – 4
Critique Score	Pedagogical-Critique + Mathematical-Critique	0 – 4

These three total scores align with the three kinds of knowledge we hypothesize are necessary for analysing teaching: SCK, KCT, and AK. We suggest that a participant's total mathematical score captures SCK, total pedagogical score captures KCT, and total critique score captures AK. As a reminder, these three types of knowledge are integrated and expand upon one another. For instance, KCT requires SCK; and AK captures aspects of both SCK and KCT. Therefore, there is not a simple way to solely capture these three kinds of knowledge individually. Attempting to identify whether a participant possesses only one type of knowledge is a complex feat because of the ways in which these three knowledge types depend on one another. Therefore, although we singularly map each knowledge type (SCK, KCT, and AK) onto a total calculated score (mathematical, pedagogical, and critique), we believe that it is

difficult to separate the analyses of these types of knowledge and admit to the limitations of our methods in doing so. Still, we believe that certain scores capture certain knowledge types in greater detail than others. Therefore, in the following paragraphs, we explain that the mathematical score *mainly* captures SCK, the pedagogical score *mainly* captures KCT, and the critique score *mainly* captures AK. However, all three scores capture all three types of knowledge to some extent due to the integrated nature of these three knowledge types.

The mathematical score focuses on the key mathematical ideas involved in the classroom interactions and suggestions for how to improve the ways the mathematics is presented. This score focuses mostly on SCK because in order to discuss (and critique) the key mathematical ideas in the video, participants must possess deep content knowledge of the mathematics. For example, for division of fractions, Hiebert and colleagues identified two components of the key mathematical ideas in this video as knowing that (a) two thirds fits into one half *less than one time*; and (b) two thirds must be partitioned (divided into equal-sized parts) to make part of it fit into one half. This type of knowledge focuses on the repeated subtraction meaning of division (knowing how many copies of 2/3 fit into 1/2) and knowing how to show this using pattern blocks. This knowledge relies on a depth of understanding of what *division means*, what *division of fractions* means, how to *explain* this meaning, and how to *show* this process visually. To describe and critique these two key components thus requires much more knowledge than simply knowing how to divide 1/2 by 2/3 procedurally. An appropriate analysis would require knowledge of why the two answers (1 1/6 and 5/6) nominated by students in the video are incorrect and what misconceptions the students are invoking (e.g., commuting the number sentence). Therefore, this score mainly aligns with SCK.

The pedagogical score mainly captures KCT because this score focuses on what participants notice and critique regarding the pedagogical moves of the teacher in the videos that lead to deeper student conceptual understanding. This score captures whether participants possess knowledge of certain pedagogical decisions and how these decisions affect students' conceptual understandings. For example, critiquing the affordances and constraints of using pattern blocks to help students understand the idea of "fitting in" requires KCT. This connection between pedagogical decisions and how they affect student learning is precisely what KCT entails. Furthermore, offering suggestions for different pedagogical moves focused on providing students with more time to grapple with the mathematical ideas (e.g., the meaning of division) and/or ways to make the key mathematical ideas more explicit for students is evidence of KCT. These kinds of suggestions require KCT because, in order to offer pedagogical suggestions to improve students' conceptual understandings, one must know how to analyse the types of pedagogical moves present in the video and why these decisions did not lead to deep conceptual understanding. These kinds of observations and critiques thus require a substantial degree of KCT.

Finally, the critique score mainly captures AK because of the focus on analysing cause-effect relationships between the teaching and student learning. As mentioned previously, we assume that there is overlapping knowledge between AK and SCK and between AK and KCT. In order to effectively analyse causal relationships between teaching (mathematical representations and pedagogical decisions) and student learning, one must possess a certain degree of SCK *and* KCT. Then, one must be able to use this knowledge (SCK and KCT) in order to analyse what aspects of teaching led to certain types of student learning. For example, in the division of fractions video, suggesting that the teacher could have clarified the meaning of division in order to improve student learning would be evidence of AK that also overlaps with SCK. Using evidence in the video of the way the first student attempted to model 1/2 divided by 2/3 (by fitting in 1/2 into 2/3) to show that students do not yet possess deep conceptual understanding is one way to display AK. To then offer suggestions for improvement would be further

evidence of AK because of the explicit focus on what teaching aspects should be improved in order to deepen students' conceptual understandings. Similarly, critiquing the amount of time that the teacher allowed the first student to explain his reasoning (the student was essentially given no time to speak) would be evidence of AK that overlaps with KCT because providing students with more time to struggle is a pedagogical move.

To compare the separate effects of topic and time in addition to the interaction between time and topic, we conducted two-way repeated measures ANOVAs three separate times: once for mathematical scores, once for pedagogical scores, and once for critique scores. Each ANOVA included two within-subject factors—one with four levels: topic (multiplication of two-digit whole numbers, subtraction of fractions, division of fractions, and the mean) and one with three levels: time (two, three, and four years post-graduation). Post hoc tests for significant interactions were conducted by calculating simple main effects using a Bonferroni correction for multiple comparisons. To maintain conventional p values for significance, we multiplied the p values from the post hoc tests by the number of comparisons being conducted (six for the simple main effects for topic and three for the simple main effects for time) and then compared these values to conventional p values to determine significance. Mauchley's test was used to investigate sphericity. In the case of a violation of sphericity, the Greenhouse-Geisser correction was applied.

Our small sample ($N = 25$) reduces the power of the ANOVAs. While we set the p level at 0.05 for determining statistically significant results, we also considered marginally significant results ($p < 0.1$) and important patterns that emerged in the data with respect to our research question. See Carver (1978) and Cronbach (1957, 1975) for a discussion of small sample results that are not statistically significant at a standard level but still might be educationally significant, especially if the results display a consistent pattern.

Results

In this paper, we investigated the following research question: *Do graduates perform better analysing teaching for topics taught during their education program compared to topics not taught during their program?* As a reminder, the target topics are: multiplication of two-digit whole numbers, subtraction of fractions, and division of fractions. We will refer to these three topics henceforth as "multiplication," "division," and "subtraction" for short. The one control topic is: the mean. We first present our results separately for each of the three total scores: mathematical, pedagogical, and critique.

Means and standard deviations for participants' three different scores (mathematical, pedagogical, and critique) earned two, three, and four years after graduation are displayed in Table 2 according to each topic (multiplication, subtraction, division, and mean). In addition, pairwise comparisons between each of the target topics and the control topic every year for each score are shown in Table 3. In other words, Table 3 shows the difference between the average scores reported in Table 2 by subtracting the average score on the control topic from the average score on the target topics for each year and for each of the three scores. Recall our hypotheses that graduates would have greater KCT and SCK (as measured by mathematical and pedagogical scores, respectively) for topics taught during their program (target topics) compared to topics not taught during their program (control topics). Further, we predicted that graduates' AK (measured by critique scores) would be only slightly greater for target topics than for control topics. Lastly, we were conflicted about how graduates' KCT, SCK, and AK might vary over time. As such, we are interested in comparing mathematical, pedagogical, and

critique scores, over time for target topics (multiplication, subtraction, division) to the same scores for a control topic (mean).

Table 2
Mean (SD) for all three scores (maximum of four points)

Mathematical Scores = Mathematical Descriptive + Mathematical Critique			
Topic	Years since graduation		
	2	3	4
T1: Multiplication	1.72 (1.14)	1.96 (1.06)	1.76 (1.27)
T2: Subtraction	2.16 (1.38)	2.20 (1.41)	2.20 (1.41)
T3: Division	1.76 (1.17)	1.44 (1.12)	1.12 (0.78)
Control: Mean	0.92 (0.70)	1.08 (1.01)	1.04 (0.89)
Pedagogical Scores = Pedagogical Descriptive + Pedagogical Critique			
Topic	Years since graduation		
	2	3	4
T1: Multiplication	2.28 (0.74)	2.36 (0.70)	2.44 (0.65)
T2: Subtraction	2.44 (0.58)	2.56 (0.77)	2.60 (0.50)
T3: Division	2.60 (0.82)	2.48 (0.92)	2.64 (0.70)
Control: Mean	2.08 (0.57)	2.08 (0.81)	2.24 (0.60)
Critique Scores = Pedagogical Critique + Mathematical Critique			
Topic	Years since graduation		
	2	3	4
T1: Multiplication	1.08 (1.12)	1.28 (1.21)	1.16 (1.43)
T2: Subtraction	1.20 (1.19)	1.44 (1.08)	1.52 (1.01)
T3: Division	1.24 (1.23)	1.08 (1.19)	.80 (.71)
Control: Mean	.28 (.54)	.48 (.65)	.52 (.82)

Note that T1, T2, and T3 represent the three target topics, while Mean represents the one control topic.

Table 3
Test of simple main effect for topic for each score (row – column)

		Mathematical scores	Pedagogical Scores	Critique Scores
Years since graduation	Topic	Control: Mean	Control: Mean	Control: Mean
2	T1: Multiplication	0.80**	0.20	0.80***
	T2: Subtraction	1.24***	0.36	0.92**
	T3: Division	0.84*	0.52	0.96**
3	T1: Multiplication	0.88*	0.28	0.80**
	T2: Subtraction	1.12***	0.48	0.96***
	T3: Division	0.36	0.40	0.60
4	T1: Multiplication	0.72	0.20	0.64
	T2: Subtraction	1.16***	0.36	1.00**
	T3: Division	0.08	0.40	0.28

Note: $^{m}p < 0.1$ *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$; also T1, T2, and T3 represent the three target topics, while Mean represents the one control topic.

In reference to our research question (*Do graduates perform better analysing teaching for topics taught during their education program compared to topics not taught during their program?*), the answer appears to be yes. The data in Table 3 show significant differences between several target topics and the control topic for mathematical and critique scores over several years. For example, two years after graduation, participants' average mathematical score on multiplication was 1.72 and their average mathematical score on the mean was 0.92 (see Table 2). The difference between these two average scores (1.72 – 0.92 = 0.80) is significant ($p < 0.05$) according to Table 3, suggesting that participants performed significantly better analysing the mathematics in the multiplication video compared to the mean video. In addition, although there are not statistically significant differences between target topics and the control topic for pedagogical scores in any year, all pairwise comparisons favour the target topic. In fact, in every single case across all three scores and all three years, comparisons favour the target topic. This suggests that participants perform better analysing topics for which they were taught during their preparation program compared to topics they were not taught. Next, we elaborate on our results for each of the three main scores: mathematical, pedagogical, and critique scores.

Total mathematical scores

Mauchley's Test of Sphericity was not significant in any case, and therefore sphericity was not violated. Mathematical scores showed that there was not a statistically significant main effect for time ($F(2,48) = 0.458, p > 0.05$) or for the interaction between time and topic ($F(6,144) =$

$1.364, p > 0.05$). Thus, graduates' mathematical analyses do not appear to change over time for any of the four topics and their mathematical analyses for each topic do not depend on time and time does not depend on topic. However, there was a statistically significant main effect for topic ($F(3,72) = 13.759, p < 0.001$), indicating that graduates' mathematical analyses appear to differ based on topic, irrespective of time (see Figure 1). A test for the simple main effect for topic displayed significant differences between various target topics and the control topic in each year. Two years after graduation, participants' mathematical scores for all three target topics (multiplication, subtraction, and division) were significantly higher than their mathematical scores on the control topic ($p < 0.01, p < 0.001,$ and $p < 0.05$, respectfully; see Table 3). Three years after graduation, participants' mathematical scores for two target topics (multiplication and subtraction) were significantly higher than the control topic ($p < 0.05$ and $p < 0.001$, respectfully). And four years post-graduation, participants' mathematical scores on one target topic (subtraction) were significantly higher than their scores on the control topic ($p < 0.001$). When comparisons were not statistically significant, participants' mathematical scores appeared to favour target topics in any year. Thus, results show that participants tend to perform significantly better analysing the mathematics on target topics compared to the control topic several years post-graduation.

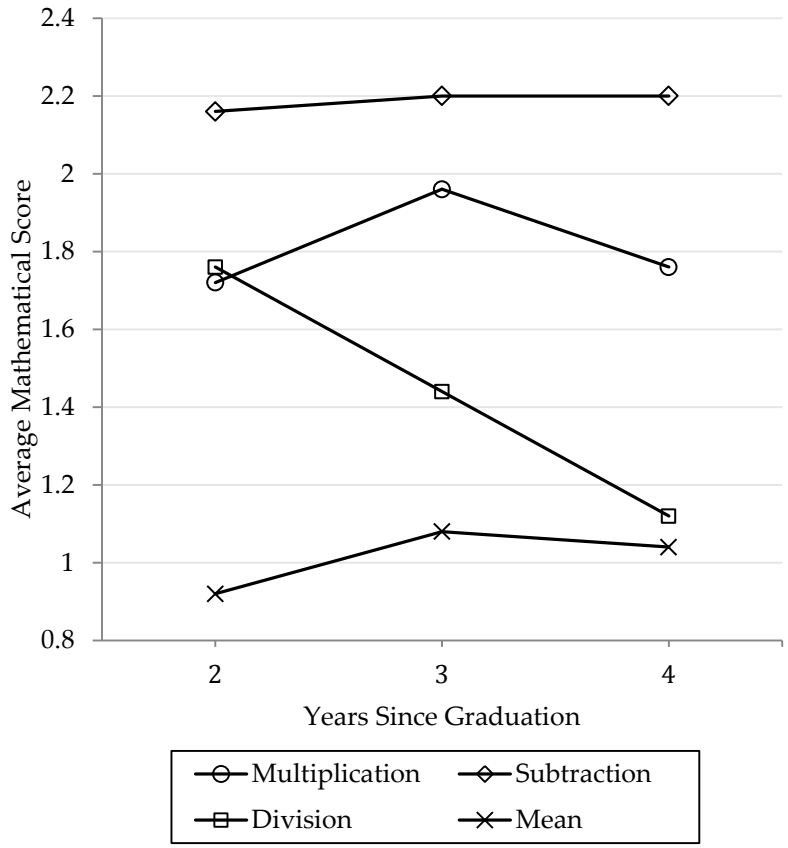

Figure 1. Changes in average mathematical scores on the four topics over time.

One surprising note was the decline of the average mathematical scores for division of fractions. As shown in Table 2, participants' average division scores declined from 1.76 (year 2) to 1.44 (year 3) to 1.12 (year 4). Such a decrease may be due to a number of reasons. First, working with fractions may be more difficult for students than working with whole numbers. However, since we do not see the same pattern of results with subtracting fractions, this decline appears to be unique to division of fractions. Upon further speculation, it may be the case that when the divisor is greater than the dividend, dividing fractions poses a unique challenge to participants that does not occur when subtracting fractions. This challenge involves subtly altering the meaning of division from "how *many copies* of the divisor fit into the dividend" to "how *much* of the divisor fits into the dividend." An analogous shift in meaning does not appear to occur with subtraction until the introduction of negative numbers, which does not occur until after elementary school in the United States curriculum. This unique challenge to division thus may have prevented participants from adequately providing a quality critique of what mathematics actually occurred in the video.

Second, it may also be the case that the types of curricula graduates were using while teaching may not develop teachers' and students' conceptual understandings of division of fractions while promoting others (subtraction of fractions, multiplication of two-digit whole numbers), thereby discouraging graduates to retain their knowledge of the topic. Whatever the reason, this result came as a surprise to our research team since the graduates' preparation program heavily focuses on the concept of division both as repeated subtraction and as partitioning. Furthermore, these two meanings of division are addressed not only in the first mathematics content course on whole numbers and decimals, but again in the second mathematics content course on fractions and operations. Therefore, it was surprising to observe a decrease in graduates' mathematical scores for division of fractions.

Total pedagogical scores

Mauchley's Test of Sphericity indicated that the sphericity assumption was not violated for time, topic, or the interaction term. Similar to the results for mathematical scores, there was not a statistically significant main effect for time ($F(2,48) = 0.775, p > 0.05$), indicating that graduates' pedagogical analyses do not appear to change over time several years post-graduation. There was also not a statistically significant effect for the interaction between time and topic ($F(6,144) = 0.223, p > 0.05$), indicating that graduates' pedagogical analyses for each topic do not appear to depend on time and any changes over time do not appear to depend on topic. However, there was a statistically significant main effect for topic ($F(3,72) = 4.721, p < 0.01$), indicating that graduates' analyses appear to differ based on topic, irrespective of time (see Figure 2).

A test for the simple main effect for topic two, three, and four years post-graduation did not reveal any statistically significant differences between topics in any single year (see Table 3). However, the overall main effect for topic along with the fact that the average scores always favoured the target topics (see Table 3) suggests better pedagogical analysis by graduates on the target topics than the control topic.

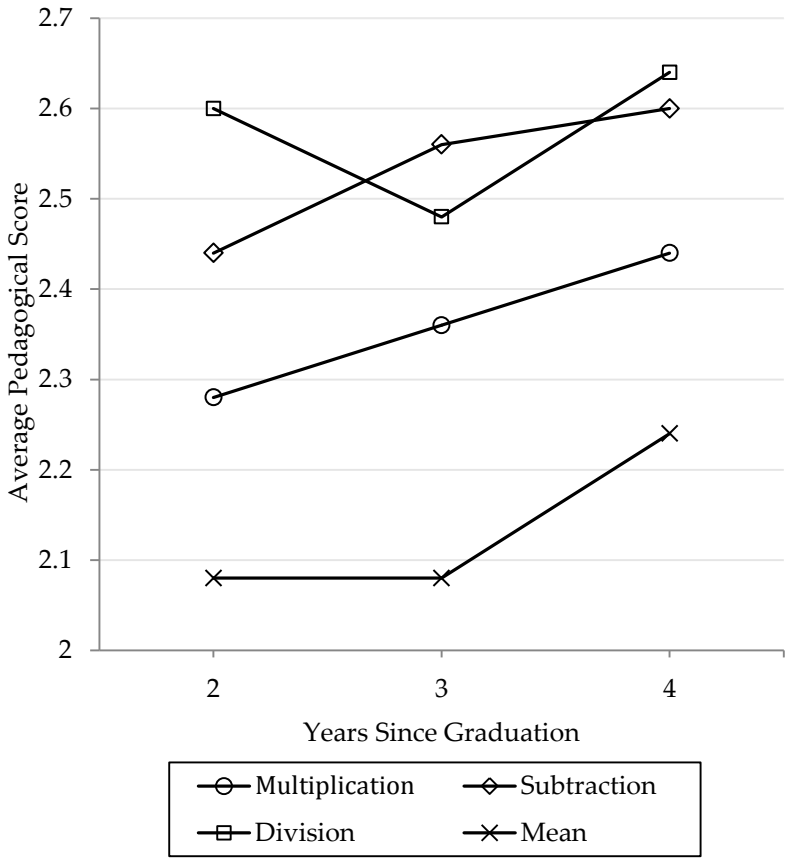

Figure 2. Changes in average pedagogical scores on the four topics over time.

Total critique scores

The sphericity assumption was not violated since Mauchley's Test of Sphericity was not significant for either main effect or the interaction term. As with the two previous scores, there was not a statistically significant main effect for time ($F(2,48) = 0.279, p > 0.05$) or for the interaction between time and topic ($F(6,144) = 1.14, p > 0.05$). However, there was a statistically significant main effect for topic ($F(3,72) = 9.633, p < 0.001$). Thus, graduates' critique scores across topics differed, yet their scores did not change significantly over time from two years post-graduation to four years post-graduation for any topic indicating that the effect of topic does not depend on time (see Figure 3). We recognize the same surprising decline in division of fractions regarding graduates' critique scores, which aligns with the results on graduates' mathematical scores since the critique score includes one component of participants' mathematical scores (mathematical-critique).

A test for the simple main effect for topic on graduates' critique scores two, three, and four years post-graduation reveals several significant differences between target and control topics in certain years (see Table 3). Specifically, two years after graduating, their critique scores on all three target topics (multiplication, subtraction, and division) were significantly higher than their control topic scores ($p<0.001$, $p < 0.01$, and $p < 0.01$ respectfully). Three years after

graduating, their critique scores on two target topics (multiplication and subtraction) were significantly higher than the control topic ($p < 0.01$ and $p < 0.001$ respectfully). Finally, four years after graduating, their critique scores on one target topic (subtraction) were significantly higher than the control topic ($p < 0.01$). When comparisons were not statistically significant, participants' critique scores appeared to favour target topics in any year. Thus, results show that participants tend to perform significantly better critiquing target topics compared to control topics several years post-graduation. It is noteworthy that all three scores (mathematical, pedagogical, and critique) showed statistically significant effects for topic ($p < 0.01$ in all three cases).

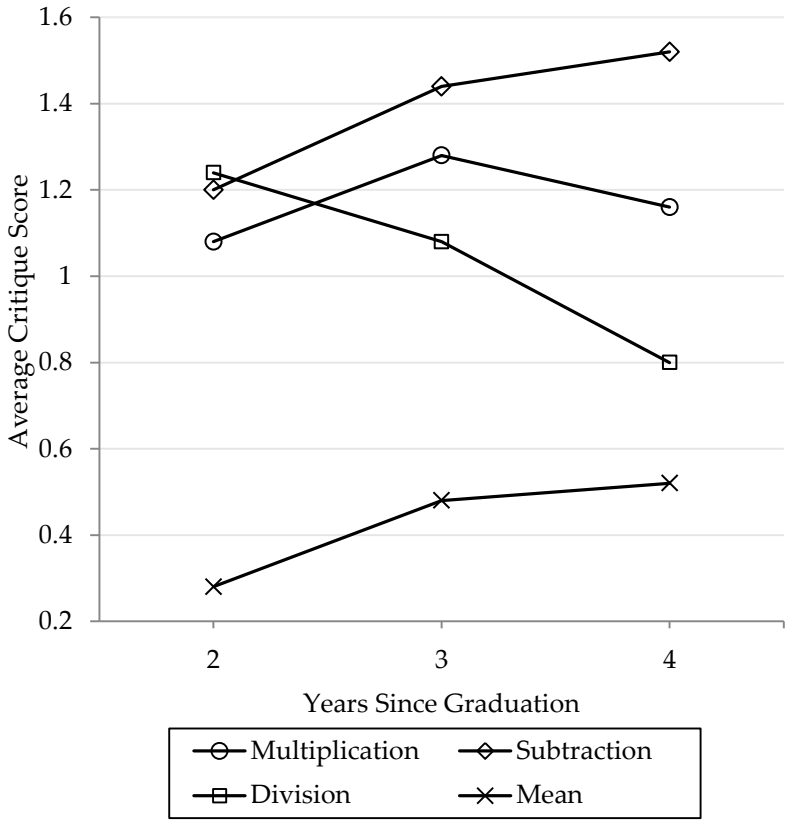

Figure 3. Changes in average critique scores on the four topics over time.

Discussion

Our results show that graduates' mathematical, pedagogical, and critique scores favoured topics taught during their program over a topic not taught during their program, even several years after graduating. In many cases, these comparisons were statistically significant. Thus, in response to our research question, our results support the claim that graduates performed significantly better on analysis-of-teaching tasks for topics taught during their program compared to topics not taught during their program. As a reminder, our small sample size merits caution in interpreting our results. However, because our main finding corroborates

Hiebert et al.'s (2017) results, the power of replication in this regard strengthens our claim. As such, we encourage other replication studies as a means to deepen the knowledge base for this type of work. Because longitudinal studies such as these are unlikely to include sample sizes affording power levels appropriate for many statistical analyses, it is important to rely on replications of similar findings under similar conditions to strengthen these types of claims.

Based on our theoretical perspective, results from both studies regarding graduates' mathematical scores suggest that graduates do indeed exhibit greater specialized content knowledge (SCK) for topics taught during their program than for topics not taught during their program, which supports our first hypothesis. However, conclusions regarding graduates' KCT are not as straightforward due to the non-statistically significant results for pedagogical scores. It is likely that graduates possess more KCT for target topics than for control topics based on the consistent pattern of findings between our study and Hiebert et al.'s (2017) study, but further research is warranted to support this claim.

In addition, we did not find compelling evidence to suggest that AK develops alongside SCK and KCT. On the one hand, graduates tended to perform significantly better on target topics than on control topics for their critique scores several years post-graduation. Yet on the other hand, graduates' critique scores were the lowest of the three scores (mathematical, pedagogical, and critique). This result aligns with our initial hypothesis. We predicted that graduates would be better equipped to develop AK in target topics because we hypothesized that graduates would have greater SCK and KCT in target topics than in control topics. However, we were unsure whether graduates would independently develop AK (as this knowledge was not directly emphasized in their education program) as a result of possessing the prerequisite kinds of knowledge needed to develop AK. It would seem, at least in the context of this study, that SCK and KCT can only take teachers so far in their AK without explicit attention to developing AK during their preparation. It follows that AK may not be an immediate consequence of solely possessing SCK and KCT, and therefore AK is not just the combination of SCK and KCT, but a more distinct type of knowledge - a kind of knowledge that must be explicitly addressed and developed in teacher preparation coursework. If the graduates' preparation program had focused more explicitly on developing AK, the differences between graduates' critique scores on target topics compared to the control topic may have been even more pronounced.

Finally, our results show that graduates' scores in any topic did not change significantly over time for any score (mathematical, pedagogical, or critique). Recall that our hypothesis regarding change over time was mixed because we were unsure whether graduates' knowledge would remain active and increase each year or deteriorate from inactivity over time. This finding reveals that neither hypothesis is supported. We suggest that further research is warranted in this area with regard to how graduates' knowledge changes over time.

In both studies, graduates consistently scored higher on topics that were taught in their program compared to topics not taught in their program when tested several years after graduating. This finding suggests that the content taught during graduates' freshman and sophomore years of their teacher education program makes a difference in their ability to analyse mathematics teaching in ways correlated with high quality instruction (Kersting et al., 2010; Kersting et al., 2012), even several years after graduating. If this is true, it would have significant implications for policies concerning the preparation of prospective teachers.

The ultimate question thus becomes whether the graduates' preparation program is the main reason for these findings or if another factor can better explain these findings. Hiebert et al. (2017) discuss four primary alternative hypotheses: (a) teaching the target topics more than the control topic; (b) engaging in professional development experiences involving target topics more often than those involving the control topic; (c) teaching from a curriculum that provided

better learning opportunities for graduates for the target topics than for the control topic; and (d) differences in the videos for each of the four topics making it easier to analyse the target topics. Hiebert et al. (2017) discuss each of these hypotheses and provide evidence to show that these alternative hypotheses are not supported.

One alternative hypothesis that Hiebert and colleagues (2017) do not discuss is the possibility of the curriculum emphasizing the target topics for more *time* over the control topic. While they showed that more graduates do not teach specific topics over others, they do not explore the time that graduates spent teaching each topic. It could be the case that the curricula graduates used emphasized the topics of multiplying two-digit whole numbers, subtracting fractions, and dividing fractions much more than the concept of finding the mean of a set of whole numbers, even if the curriculum addressed all four topics to some extent. Hiebert et al. (2017) also suggest that the graduates were teaching in different districts and schools across the region, suggesting that they likely used different curriculum materials. We concur with this reasoning. However, we do not have enough evidence to rule out this alternative hypothesis. Therefore, we suggest that future research replicate these methods but include more than one control topic to determine whether the teacher preparation program is the primary factor influencing graduates' knowledge. Using more than one control topic would increase the likelihood of various curricula emphasizing different topics for different amounts of time and would increase the odds of the results being explained by the graduates' teacher education program.

Concluding remarks

The most consistent finding shows that graduates performed better analysing teaching about topics taught in their teacher education program compared to topics not taught in their program. In many ways, this finding suggests that the graduates' teacher education had a significant effect on graduates' abilities to analyse teaching. One way to strengthen this type of claim is to investigate alternative competing hypotheses and the likelihood of each, as Hiebert et al. (2017) did in their study. Yet another method of strengthening this claim is through replication (Smith, 1970). If the same findings continue to result when employing similar methods of analysis with different cohorts, it is more likely due to reasons involving experiences common to all cohorts than differential experiences across cohorts. Since graduates likely have different professional development opportunities, different curricula, and other differing factors relevant to developing knowledge related to teaching, then the likelihood of these factors explaining the results is weakened substantially. However, since cohorts all experience the same teacher education program, the likelihood of this common experience explaining the results is substantially strengthened.

Our results speak to the accusations that teacher preparation programs are not useful for teachers. In particular, our results show that teacher preparation can make a substantial difference in graduates' knowledge and competencies related to teaching, but only for topics specifically studied during preparation. Graduates' knowledge did not transfer to the topic not studied during preparation. This suggests that teacher preparation needs to be organized around the specific competencies most valued for beginning teachers. It also means subject matter courses are important but probably make the most difference if they focus on SCK, KCT, or AK, or at least some form of content knowledge for teaching. These results show that teacher preparation programs can have a significant and lasting positive effect on graduates' knowledge, as measured by an analysis-of-teaching task shown to be correlated with high

quality teaching, which is exactly the kind of evidence teacher preparation programs are currently under pressure to demonstrate.

Acknowledgements

The study reported in this article was supported by the National Science Foundation (Grant #DRL-0909661). The opinions expressed in the article are those of the authors and not necessarily those of the Foundation. Special thanks to James Hiebert and Dawn Berk for their support, and to Emily Miller for her assistance with statistical analyses.

References

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

Bryk, A. S. (2015). 2014 AERA distinguished lecture: Accelerating how we learn to improve. *Educational Researcher*, *44*(9), 467–477. http://doi.org/10.3102/0013189X15621543

Constantine, J., Player, D., Silva, T., Hallgren, K., Grider, M., & Deke, J. (2009). An evaluation of teachers trained through different routes to certification. Final Report. NCEE 2009-4043. *National Center for Education Evaluation and Regional Assistance*.

Darling-Hammond, L., Holtzman, D. J., Gatlin, S. J., & Vasquez Heilig, J. (2005). Does teacher preparation matter? Evidence about teacher certification, Teach for America, and teacher effectiveness. *Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas*, *13*.

Feuer, M., Floden, R. E., Chudowsky, N., & Ahn, J. (2013). Evaluation of teacher preparation programs: Purposes, methods, and policy options. *National Academy of Education*. Washington, DC. Retrieved from http://www.martin.uky.edu/faculty/Toma/Evaluation_Teacher_Prep.pdf

Floden, R. E., & Meniketti, M. (2005). Research on the effects of coursework in the arts and sciences and in the foundations of education. In *Studying teacher education: The report of the AERA panel on research and teacher education* (pp. 261–308).

Greenberg, J., McKee, A., & Walsh, K. (2013). *Teacher prep review: A review of the nation's teacher preparation programs*.

Goldhaber, D., Liddle, S., & Theobald, R. (2013). The gateway to the profession: Assessing teacher preparation programs based on student achievement. *Economics of Education Review*, *34*, 29-44.

Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. In F. K. Lester (Ed.), *Second Handbook of Research on Mathematics Teaching and Learning* (pp. 371–404). Charlotte, NC: Information Age Publishing.

Hiebert, J., Miller, E., & Berk, D. (2017). Relationships between mathematics teacher preparation and graduates' analyses of classroom teaching. *The Elementary School Journal*, 117(4), 687-707. https://doi.org/10.1086/691685

Hiebert, J., Morris, A. K., Berk, D., & Jansen, A. (2007). Preparing teachers to learn from teaching. *Journal of Teacher Education*, *58*(1), 47–61. http://doi.org/10.1177/0022487106295726

Humphreys, C., & Parker, R. (2015). *Making number talks matter: Developing mathematical practices and deepening understanding, grades 4-10*. Stenhouse Publishers.

Kazemi, E., & Hintz, A. (2014). *Intentional talk: How to structure and lead productive mathematical discussions*. Stenhouse Publishers.

Kersting, N. B. (2008). Using video clips of mathematics classroom instruction as item prompts to measure teachers' knowledge of teaching mathematics. *Educational and Psychological Measurement*, *68*(5), 845–861. http://doi.org/10.1177/0013164407313369

Kersting, N. B., Givvin, K. B., Sotelo, F. L., & Stigler, J. W. (2010). Teachers' analyses of classroom video predict student learning of mathematics: Further explorations of a novel measure of teacher knowledge. *Journal of Teacher Education*, 61(1–2), 172–181.

Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. W. (2012). Measuring usable knowledge: Teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal*, *49*(3), 568–589. http://doi.org/10.3102/0002831212437853

Levin, A. (2006). *Educating school teachers*. Washing, DC: The Education Schools Project. Retrieved from http://files.eric.ed.gov/fulltext/ED504144.pdf

Lewis, C. C., Perry, R. R., & Hurd, J. (2009). Improving mathematics instruction through lesson study: A theoretical model and North American case. *Journal of Mathematics Teacher Education, 12*(4), 285–304. http://doi.org/10.1007/s10857-009-9102-7

Morris, A. K. (2012). Using "lack of fidelity" to improve teaching. *Mathematics Teacher Educator*, *1*(1), 71–101.

National Research Council. (2010). *Preparing teachers: Building evidence for sound policy*. Washington, DC: National Academies Press.

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*(2), 90–100. http://doi.org/10.1037/a0015108

Smith, N. (1970). Replication studies: A neglected aspect of psychological research. *American Psychologist*, *25*(10), 970–975. http://doi.org/10.1037/h0029774

von Glasersfeld, E. (1995). Piaget's constructivist theory of knowing. In *Radical constructivism: A way of knowing and learning* (pp. 53–75). Bristol, PA: Falmer Press.

Yong, E. (2012, May). In the wake of high-profile controversies, psychologists are facing up to problems with replication. *Nature*, 298–300.

Zeichner, K. (2014). The struggle for the soul of teaching and teacher education in the USA. *Journal of Education for Teaching*, *40*(5), 551-568.

Authors

Siobahn Suppa
Stockton University
Galloway, New Jersey
email: Siobahn.Suppa@stockton.edu

Joseph DiNapoli
Montclair State University
Montclair, New Jersey
email: dinapolij@montclair.edu

Robert Mixell
203 Willard Hall,
University of Delaware,
Newark, DE 19716
email: mixellr@udel.edu

Appendix A: Coding Rubric

We describe our coding rubric with more detail and specific examples of each of our four coding dimensions: Pedagogical-descriptive, Pedagogical-critique, Mathematical-descriptive, and Mathematical-critique. We discuss each one in turn. As a reminder, participants could earn a score ranging from 0 to 2 for each dimension.

Dimension One: Pedagogical-Descriptive Dimension

The **Pedagogical-Descriptive (PD)** dimension assesses participants' ability to determine what pedagogical moves by the teacher (if any) led to students having a deeper understanding of the content. This means that the teacher contributed to students having a deeper understanding of a key idea, students having greater knowledge of why something is true, students having greater sense-making, or students being able to more effectively conceptualize the key idea. In addition, the participant must be able to determine if such moves involved the teacher explaining key relationships or allowing students to wrestle, or productively struggle, with key ideas as a means to deepen students' understanding.

Explaining key relationships requires a participant to be clear about what items or ideas are being related to one another (e.g., relating other vocabulary or the results of an activity to the current content, discussing how one idea is similar to or has different characteristics than another idea). However, the teacher *does not* have to explain what these key relationships are. Such a description or critique of relationships, as is mathematical in nature, would thus be captured by either the Mathematical-Descriptive (MD) or Mathematical-Critique (MC) dimensions. *Wrestling* involves a teacher requiring students to productively struggle or practice exploration as a means for students to make sense of the material themselves.

Table 4
Pedagogical-Descriptive Coding Scheme

Code	Description	Example(s)	Notes
0	A PD score of 0 points is given when a participant writes that a teacher in the video helped students understand something about the mathematics to be true by explaining the key relationships underlying the mathematics and/or allowed students time to wrestle with key ideas. Both of these moves however must be linked to an observation of the occurrence of student understanding of key ideas or why something is true as a result (not merely	"She has students think through their mistakes on their own, rather than tell them if they are correct or incorrect. It helps students to better make sense of their ideas and develop that conceptual understanding that the teacher is striving for in her students."	The instructor is the subject of the statement. The students "think[ing] through their mistakes on their own" clearly implies student wrestling, and helping students to "better make sense of ideas" and fostering "conceptual understanding" clearly shows that the participant did in fact view conceptual growth in the students as a result of the teacher move.

	addressing misconceptions, "checking" for understanding, or providing a better understanding of how to solve a problem). Note: No videos had any evidence of either of these two teacher moves supporting students' conceptual understandings.		
1	A PD score of 1 point is given when a participant writes that the teacher helped students understand something about the mathematics to be true, but did not write anything about the teacher explaining the relationships underlying the mathematics, and/or allowing students time to wrestle with key ideas to develop such an understanding. This code is slightly better than the PD score of 0 because participants are not noticing specific important pedagogical moves that did not occur in any of the videos. Instead, they are just mentioning some other teacher move that lead to evidence of student understanding.	"As the students made sense of how to use manipulatives, the instructor did a great job of incorporating the various vocabulary words to increase knowledge and solidify understanding."	The instructor is the subject of the statement. The incorporation of various vocabulary words does not explicitly refer to key relationships or student wrestling. Yet, student understanding is talked about as a result of the teacher move.
		"The teacher is fantastic at prompting the students to think about the mathematical concept in its simplest form. The students were able to see this concept, relate it to their lives, and therefore understand the lesson and meet the objective."	The teacher is the subject of the statement. It is not that students understand the lesson or objective that allows the assignment of 1 point. It is rather the participant's perception that students were able to "see the concept," which implies that student understanding of the key idea resulted from the teacher prompt.

2	A PD score of 2 points is given when a participant does not explicitly mention students' understanding of key ideas or knowing why something is true.	The participant states that the teacher using a particular pedagogical move allows them to "look at things differently, help them elaborate or expand on thoughts, or use manipulatives to prove a point."	Even though a teacher move may allow for students to do various things, such as "to prove a point", there is no explicit mentioning of student understanding of the key idea resulting from the teacher move.
		"She asks a lot of good questions too. She pauses from time to time and makes sure the students really understand the language."	Student understanding of language does not imply student understanding of the key idea resulted from the teacher move.

Additional Notes:

For any pedagogical activity, the teacher must be the subject of the statement. That is, the teacher's enactment of the activity (not the activity itself) must be the reason for deeper understanding. Furthermore, it must be clear that the teacher-provided opportunity did in fact lead to student understanding. Any terminology that does not suggest definitive understanding on the part of the students (e.g., "checking for understanding," "addressed misconceptions," students "seemed" to understand) or any mere discussion of how a pedagogical move "could help" or "does typically help" does not count toward a score of 0.

Dimension Two: Pedagogical-Critique Dimension

The **Pedagogical-Critique (PC)** dimension assesses to what degree a participant effectively critiques pedagogical moves performed by the teacher in the video. This includes their ability to notice areas in which the teachers could improve his/her pedagogical practice, as well as their ability to potentially offer suggestions about such practices. This dimension also accounts for participants' suggestions regarding the use of teacher moves that particularly allow for the *identification and explanation of key relationships*, as well as more *productive struggle through student wrestling*.

Table 5
Pedagogical-Critique Coding Scheme

Code	Description	Example(s)	Notes
0	A PC score of 0 points is given when a participant makes no critiques or suggestions. Each video has ample evidence of pedagogy that can be improved.	Any response not containing any negative statements or suggestions about teacher pedagogy.	
1	A PC score of 1 point is given under one of two circumstances. It is given when a participant makes any negative observations about teacher moves; or when a participant makes suggestions about pedagogical changes the teacher should enact, but does not explicitly state either of the two pedagogical moves: explicitly explaining relationships or allowing students time to wrestle. A PC score of 1 is better than a PC score of 0 because the participants are noticing improvements can be made and suggesting how this could be done. When a participant suggests the use of a pedagogical move as a means to change the way the mathematics is discussed in the lesson (e.g., "they should better explain the place holder"), this suggestion is captured under the mathematical-critique (MC) score.	"I don't believe that the student fully understood the steps. The teacher basically took the clipboard from the student and wrote down the answer."	The participant provides a negative response of a pedagogical move, which counts as a pedagogical critique.
		"I would have provided more scaffolding depending on student level and how familiar they were with the content at this point."	There is no negative observation. However, there is a suggestion about a pedagogical move, yet not with respect to students identifying key relationships or wrestling.
		"I think it would have been most helpful to the students if at the end he had asked them to decide which the right answer, instead of stating it was."	A suggestion is provided here regarding a pedagogical move, but a student deciding if an answer is correct or not does not imply exploration or productive struggle with understanding an idea. Thus, this response was coded as a PC score of 1.
2	A PC score of 2 points is given when the participant suggests that the teacher could improve the lesson by enacting the two pedagogical moves: explicitly explaining key relationships and/or allowing for student wrestling. This score is also applied	"The teacher gives students a chance to show and explain their work, but does not allow them to continue exploring with the tiles once an answer was given."	Exploration with the tiles suggests student wrestling with the key idea of sense-making for division of fractions.

	if participants mention suggestions such as allowing for more inquiry for understanding an idea or allowing for students to do more exploratory work with mathematics, both of which imply more student wrestling with ideas. A PC score of 2 is the best PC score because participants are noticing that improvements can be made and are suggesting moves that better support students' conceptual learning.	"The entire class had a manipulative, which is extremely helpful for the students to visualise the problem. However, the students were never given the opportunity to share their ideas, explore, connect their previous knowledge, or draw their own conclusions."	The participant suggests that students should explore more and come up with their own conclusions, thus this response was coded as a PC score of 2.

Additional Notes:

Once a participant suggests the use of a pedagogical move as a means to change the way the mathematics is discussed in the lesson (e.g., "they should better explain the place holder", then this is coded under the mathematical-critique (MC) dimension and not the pedagogical-critique (PC) dimension.

A participant wanting students to merely "talk more" or that is, to talk more in groups or ask more questions with no tie back to understanding ideas does not earn a PC score of 2. However, if a participant suggests that "the teacher should allow for more inquiry," this implies student exploration of an idea and thus a PC score of 2. Furthermore, a participant's suggestion of allowing students to "do more work," such as do more mathematics problems to understand how to solve a certain type of problem, but void of students doing the leg work to building an initial understanding of an idea does not imply that exploratory work or wrestling is occurring. In both cases, a PC score of 1 is assigned.

Dimension Three: Mathematical-Descriptive Dimension

The **Mathematical-Descriptive (MD)** dimension assesses a participant's ability to identify the mathematical content in each video. This includes the ability to notice mathematical operations, procedures, concepts and the relationships between these items. In addition, this dimension also allows for determining whether participants recognize some or all of the components pertaining to the key mathematical ideas within each of the four topics (division of fractions, subtraction of fractions, multiplication of whole numbers, and finding the mean). For our coding purposes, we identified the following components as the key ideas within each of the four topics:

- **Division of Fractions (target topic)**
  1. Two thirds fits into one-half less than one time; and
  2. two thirds (the whole) must be partitioned (divided into equal-sized parts) to make part of it fit into one-half.
- **Subtraction of Fractions (target topic)**
  1. What counts as one is made explicit (or what counts as 1/9 or 1/12 or 1/3).
- **Multiplication of Whole Numbers (target topic)**
  1. Multiplying by a factor of 10 means the product will be ten times as big as multiplying by a factor of 1; and

2. the answer will end in zero (zero is part of the answer).

- **Finding the Mean (control topic)**
  1. The total for a set is the number of items times the average size of each item; and
  2. each item has more or less than the average.

Table 6
Mathematical-Descriptive Coding Scheme

Code	Description	Example(s)	Notes
0	A MD score of 0 points is given when a participant does not mention any mathematics, or only mentions, at most, a single word or label about the mathematics (ex: says "place value," but doesn't describe what it is at all). This is the worst code because in the videos there is a lot of mathematics, thus a great potential to unpack that mathematics.	Division of Fractions	
		"In the beginning, it was a good strategy to have the one boy student who decided his answer was one and one-sixth to come up to the board and show his thinking."	The mere mention of a student coming up with the answer to a problem (1 1/6) is not enough to be considered as a description of the mathematics.
		Subtracting Fractions	
		"She asked them to take pieces away from the whole and had students think critically about what exactly was going on and what they had left."	The idea of taking pieces away is mentioned, but no actual numbers are mentioned to illustrate the occurrence of this operation in the video.
		Multiplying Whole Numbers	
		"When asked the question about the reason for the zero, I thought it was great that the teacher emphasized the importance of knowing why you're doing something as well as how to do it and was able to go back and explain step-by-step why the zero is there."	The word "zero" is mentioned, but not in relation to any other procedures or concepts.
		Finding the Mean	
		"The students were able to find the average number of pets, which was 4."	The mere mention of the goal of the problem and no other mathematics is not enough to constitute a description of the mathematics.
1	A MD score of 1 point	Division of Fractions	

	is given when a participant describes the mathematics with at least a phrase about a procedure, concept, definition, misconception, task, etcetera that was present during the video. However, the participant does not discuss all of the key mathematical ideas at stake (they may have discussed some of them, but not all). This is better than the previous code because the participant is starting to notice and describe some of the mathematics in the video.	"When he calls the first boy up to show how he got the answer 1 1/6 the student shows how 1/3 fits into 1/2 instead of how 2/3 fits into 1/2. The student is clearly confused and the teacher redirects him by showing the previous example (1/2 divided by 1/4) then redoing the problem they are working on. I found it extremely helpful when the teacher puts the pink 1/2 circle and the orange 3/4 circle together then breaks it apart into pieces."	The participant discusses a lot of mathematics from the video. However, none of the key concepts (2/3 fitting in less than once or equal-sized pieces being necessary) are discussed.
		Subtracting Fractions	
		"She then instructs them to subtract 1/3 from what remains. I found this strategy very interesting because she did not stop to discuss what the students started with, but rather expected students to know that they were taking 1/3 from 9/12."	The participant mentions the task 9/12 – 1/3. However, there is no mention of the key mathematical component, that is, what counts as one (12 squares), 1/12 (1 square), or 1/3 (4 squares).
		Multiplying Whole Numbers	
		"She did not take the time to explain that the '3' in '36' really means '30.' We are using a '0' as a place holder so that we don't have to multiply 30x50 and 30x6."	The participant describes some of the mathematics, but does not mention all of the key components at stake (10 times as big or ending in zero).
		"She allows students to ask and answer questions while she is explaining clearly why there is a 0 in the ones place. The teacher outlines for students the separate ones and tens place so they can clearly see where the tens place is and then you hear children responding with, 'ahh,' as they begin to see why the 0 is in the ones place. She then explains that the - is a place holder and its job is to hold the ones place."	The participant discusses different place values and zero being in the ones place as a place holder, but does not mention all of the key components at stake (10 times as big or ending in zero).
		Finding the Mean	

		"The students showed that by continually adding the average number of pets, which was 4, seven times, which corresponded to the number of pets, they were able to find the total number of pets (28)."	The first component for finding the total number of pets is addressed, but not both components.
		"The students were able to find the total number of pets (7 x 4 = 28). However, they were still confused about the difference between the mean and the median."	Even though none of the components for finding the mean are mentioned, some of the mathematics described in the video (e.g., multiplication operation, mean versus median) is discussed.
2	A MD score of 2 points is given when a participant describes all components of the key mathematical idea that was at stake in the video. These components were identified ahead of time by the research team. This is the best code because not only are participants noticing some mathematics, but they are choosing to describe the most important aspects of the mathematics: the key mathematical ideas underlying each topic. Keep in mind, however, that the identification of these components is different than a participant mentioning suggestions for how the mathematics should be changed. Such responses would	Division of Fractions	
		"He has one piece representing 1/2 and allows the students to come up and show how many 2/3 fit into the 1/2…Then, he draws lines to show how it can be split into four equal pieces and how only 3 of those 4 fit into the 1/2, and that is why the answer to the equation is 3/4."	The participant discusses the partitioning idea and the idea that 2/3 fits into 1/2 less than one time.
		Subtraction of Fractions	
		"The teacher then re-defined what one-third was 'Remember we said 1/3 was this many (and gestured to the 4 blocks)'."	The fraction of 1/3 equaling 4 blocks is the one key component for understanding subtraction of fractions.
		Multiplying Whole Numbers	
		"The teacher showed that when performing the standard algorithm and multiplying 3 times 50, we are actually multiplying 30 times 50, which is ten times bigger than what we are actually used to saying for this step…Our final answer will end in zero."	All key components are mentioned (10 times as big and ending in zero).
		Finding the Mean	
		"The students discover that they can work backwards from the standard formula for the mean by multiplying	Both components are mentioned (number of families multiplied

	be coded under the mathematical-critique (MC) dimension.	the number of families by the average number of pets, which was given to them to get the total number of pets, which was 28…One girl was able to show that it is okay for different numbers of families to have different number of pets, because when the numbers of pets are added up, there are still 28."	by average equals the total and the unchanging total).

Dimension Four: Mathematical-Critique Dimension

The **Mathematical-Critique (MC)** dimension is about the changes participants suggest in the mathematics that was present in the video. This includes a focus on either the explanations presented by a teacher in a video, a problem that is presented in the video, or the way in which the mathematics was discussed. This dimension accounts for both participants' *general* suggestions about changes in the mathematics, as well as *specific* ways in which such suggestions could be carried out in the classroom. Note that any participant observation regarding a pedagogical move that contains mathematics but suggests a change in the mathematics would be considered a mathematical critique, not a pedagogical critique.

Table 7
Mathematical-Critique Coding Scheme

Code	Description	Example(s)	Notes
0	A MC score of 0 points is given when a participant makes make no suggestions about mathematical changes that should occur. This is the worst code because these videos have many opportunities for improvement regarding how mathematical concepts, procedures, and relationships could be taught and learned. Keep in mind that any negative statement regarding the mathematics that does not include a suggestion earns a MC score of 0.	"The explanation about place value with the number 30 was very confusing." "When the teacher talked about subtracting 1/3 from 9/12, she wasn't clear with the students about whether 1/3 of 9/12 was to be subtracted or 1/3 of a whole."	These are only negative critiques, not suggestions.
1	A MC score of 1 point is given when a participant makes a general suggestion about something that could have been done differently with the mathematics (suggesting a different kind of task, suggesting explanations should include specific mathematical ideas,	"The teacher then re-defined what one-third was. 'Remember we said 1/3 was this many (and gestured to the 4 blocks)'. I think that this part was very confusing to the students, so the teacher should have explored it more. We're taking one third away from twelfths, not ninths and done more with that."	The participant alludes to the teacher needing to discuss more deeply about the representation for 1/3. However, the participant does not fully suggest a particular way in which this should have occurred.

	suggesting that mathematics could have been done better, etc.). However, the participant does not identify any of the components of the key mathematical idea at stake in their suggestion. This is better than the previous code because the participant is suggesting specific improvements about the mathematics that they think might improve the lesson.	"They should also be using more academic language while they are describing their answers to their peers and while the teacher is describing information to his students."	The participant suggests the mathematics could have been discussed differently, yet without a particular example for how it should be done. This is a mathematical and not pedagogical suggestion because the use of more academic language is with regard to students describing answers and the teacher describing information to the students.
2	A MC score of 2 points is given when a participant identifies at least one component of the key mathematical idea at stake in their suggestion. This is the best code because participants are suggesting specific improvements about at least one of the most important aspects of the mathematics.	"What I would do different, especially if the students were focusing on place value, is rather than stating what six times five was as the teacher did, I would ask this but also ask what six times fifty was and then add ten to that so that the students really understood the concept of place value when multiplying and why the positions of the numbers within the multi-digit multiplication algorithm is significant."	A clear and specific mathematical suggestion of adding 10 to the partial product is made for the goal of deepening understanding of the main idea of how place value plays a part in the value of digits in the algorithm.
		"In my opinion, these students still seem very confused by this concept and could benefit from calculating that 1/3 equals 4/12 so they needed to take away 4 twelfths."	A clear and specific mathematical suggestion of finding an equivalent fraction is made for the goal of understanding the proportion of squares necessary to perform the subtraction from 12 squares.