Math Ed Res J (2019) 31:67–88
https://doi.org/10.1007/s13394-018-0242-x

ORIGINAL ARTICLE

Prospective elementary teachers' conceptions of multidigit number: exemplifying a replication framework for mathematics education

Erik Jacobson[1] · **Amber Simpson**[2]

Received: 15 September 2017 / Revised: 23 March 2018 / Accepted: 28 March 2018 /
Published online: 27 April 2018

Abstract Replication studies play a critical role in scientific accumulation of knowledge, yet replication studies in mathematics education are rare. In this study, the authors replicated Thanheiser's (Educational Studies in Mathematics 75:241–251, 2010) study of prospective elementary teachers' conceptions of multidigit number and examined the main claim that most elementary pre-service teachers think about digits incorrectly at least some of the time. Results indicated no statistically significant difference in the distribution of conceptions between the original and replication samples and, moreover, no statistically significant differences in the distribution of sub-conceptions among prospective teachers with the most common conception. These results suggest confidence is warranted both in the generality of the main claim and in the utility of the conceptions framework for describing prospective elementary teachers' conceptions of multidigit number. The report further contributes a framework for replication of mathematics education research adapted from the field of psychology.

Keywords Mathematics education · Multidigit number · Prospective elementary teacher · Initial teacher education

Introduction

Number and operations—arithmetic—have been the core strand in the elementary (ages 5 to 11 years old) mathematics curricula for more than a century (Stanic and Kilpatrick 1992), and therefore, a critical area in which teachers need to have deep conceptual understanding. Unfortunately, decades of research reveal that many teachers

✉ Erik Jacobson
erdajaco@indiana.edu

[1] Department of Curriculum and Instruction, Indiana University, 201 N Rose Avenue, Bloomington, IN 47405, USA

[2] Binghamton University, Binghamton, USA

find it challenging to explain the algorithms and procedures they use to solve arithmetic problems (e.g., Borko et al. 1992; Leinhardt 1989; Ma 1999; da Ponte and Chapman 2015). Without explicit knowledge of why algorithms work, teachers are ill equipped to teach number and operations in a meaningful way.

Initial elementary teacher education is a natural site to improve elementary teachers' understanding of mathematics in general and their conceptions of multidigit numbers in particular. Because elementary teachers are often proficient with computational algorithms, limitations in their understanding of these algorithms can be difficult for teacher educators to identify and address. Thanheiser (2009, 2010), building on Fuson et al. (1997), introduced a framework designed to help mathematics teacher educators better understand and support prospective teachers (PTs)' conceptions of multidigit numbers.

Thanheiser (2009) used interviews (N = 15) of PTs enrolled in an elementary teacher preparation program, but who had not yet taken mathematics coursework to identify PTs' conceptions. The resulting framework describes four different ways PTs conceptualize multidigit whole numbers: the concatenated-digits and concatenated-digits-plus conceptions (both incorrect) and the reference-unit and groups-of-ones conceptions (both correct). PTs with the least sophisticated conception—*concatenated-digits*—treat each digit as a symbol or character; under this conception, the threes in 383 are in different locations but mean the same thing—three. PTs with the *concatenated-digits-plus* have partially correct conceptions but regularly conceive of at least one digit incorrectly. Using the example above, the 8 in the tens position is viewed as 8 tens, while the 3 in the hundreds position is viewed as 3 tens. The less sophisticated correct conception is the *groups-of-ones* conception, and PTs with this conception interpret each digit accurately in terms of ones. Thus, the number 383 is understood inflexibly as the sum of 300 ones, 80 ones, and 3 ones. The more sophisticated conception is also more flexible. PTs with the *reference-units* conception understand and apply the 10-to-1 and 1-to-10 relationships between digits and are able to conceive of digits accurately in terms of more than one set of appropriate units. For example, a PT with reference-units conception might understand the first 3 in 383 either as 3 hundreds, 30 tens, or 300 ones, depending on what was strategic for her or his purpose. The four major conceptions are illustrated in Fig. 1.

In a follow-up study (2010), Thanheiser replicated the initial findings by coding open-response survey questions (N = 33). Importantly, the sample for the 2010 study was of PTs who had already completed the mathematics coursework required by their teacher preparation program. Interviews of the PTs in the follow-up study revealed strong evidence that the survey questions could accurately distinguish between correct and incorrect conceptions of number. Moreover, Thanheiser (2010) used the survey results to refine the 2009 framework, subdividing the most prevalent PT conception—*concatenated-digits-plus*—into three sub-conceptions: "(a) digits consistently explained as 10, (b) digits explained consistently depending on context (i.e., 10 in subtraction, 1 in addition), and (c) changed interpretations of the digit depending on the question posed" (p. 249). For ease of reporting, we refer to these conceptions hereafter in this paper as *consistently 10*, *consistent in addition*, and *inconsistent*,

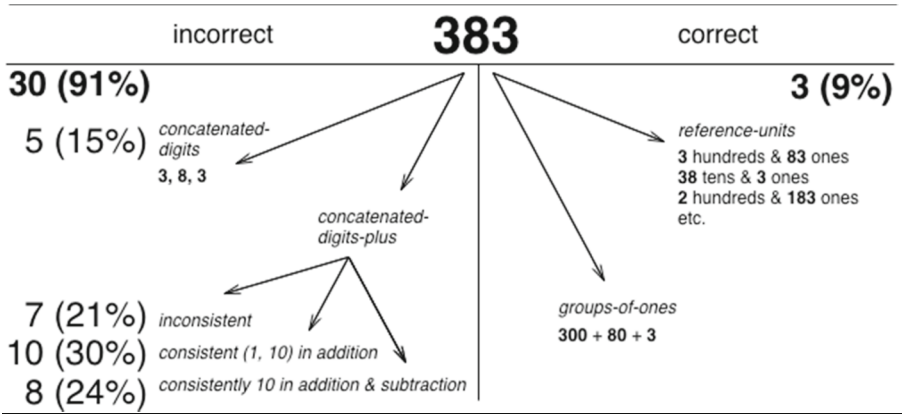


Fig. 1 The structure and prevalence of conceptions of multidigit whole numbers and sub-conceptions of the concatenated-digits-plus conception (see Thanheiser 2009, 2010). *Note*: correct conceptions were not distinguished in the survey data

respectively. The structure and prevalence of the conceptions and sub-conceptions reported in the original survey study (n = 33, Thanheiser 2010) are described in Fig. 1. The primary claim of the article was that "most [PTs] do see the digits incorrectly in terms of ones at least some of the time" (2010, p. 246). If the findings from Thanheiser's articles (2009, 2010) generalize to elementary teachers more broadly, these results have profound implications. In short, teacher educators face the challenge that most of their students misunderstand multidigit numbers and that this misunderstanding persists in spite of mathematical coursework thought to be sufficient.

In considering the survey study of Thanheiser (2010), however, a major limitation is the small sample size, which reduces confidence that generalization is appropriate. Thanheiser readily acknowledges this limitation, and in the time since the original studies were published, she has conducted two replications of the survey study at different US institutions (E. Thanheiser, personal communication, June 5, 2017). These replication studies had sample sizes of 25 and 23 participants. The estimated population mean of participants with correct proportions from the study with the largest sample (N = 33) is 9%, and the 95% confidence interval is (2.3%, 29.5%). Confidence intervals are generally larger with smaller samples, but increasing the sample size provides more accurate population estimates by narrowing the confidence interval.

As a set, the original survey study and the two replications suffer an additional limitation: all three were conducted by the researcher who introduced the framework. While the findings across these replication studies have strengthened the empirical foundation for the framework, a researcher who replicates her or his own work has a conflict of interest that fundamentally limits the additional confidence such studies can provide the field. Findings from replication studies conducted by an independent research team possess a unique element of credibility that cannot be otherwise obtained.

In this article, we report on a replication study of Thanheiser's (2010) survey study of PTs' conceptions of multidigit number. The main contribution of the study is

reporting a replication of Thanheiser's 2010 study that is independent and has a substantially larger sample size than the original. In particular, we answer the following research question:

> To what extent is the reported proportion of prospective teachers with correct conceptions and the distribution of sub-conceptions of the concatenated-digits-plus conception replicated in an independent sample?

A second contribution of the study is methodological. We adapted a replication framework from the field of psychology in order to conduct the study, and we anticipate this adapted framework may be of use to other mathematics education researchers who are interested in conducting a replication study.

Replication framework

We situate this study in the field as an example of a replication study that may be useful for other mathematics education researchers. Because replication studies are so rare in the social sciences (Schmidt 2009) and mathematics education researchers may not have had training to conduct such studies, this article contributes to the field of education reseach in part by describing how we utilized and adapted a framework for replication studies from psychology (Brandt et al. 2014) to conduct and report the present study. This is referred to by some educational researchers as a *closely aligned* replication study as direct replications in education are nearly impossible to conduct, given the unavoidable variation in factors such as educational context, teachers' instructional methods, and students' cultural and historical backgrounds (Coyne et al. 2016).

Adapting a replication recipe from psychology

Although differences in context cannot be controlled, the replication framework (*replication recipe*) discussed by Brandt et al. (2014) allowed us to conduct and evaluate a "convincing close replication" as we modified the following five ingredients to align more closely with our study within mathematics education (p. 218):

1. Carefully defining the effects and methods that the researcher intends to replicate;
2. Following as exactly as possible the methods of the original study (including participant recruitment, instructions, stimuli, measures, procedures, and analyses);
3. Having high statistical power;
4. Making complete details about the replication available, so that interested experts can fully evaluate the replication attempt (or attempt another replication themselves);
5. Evaluating replication results and comparing them critically to the results of the original study.

Each ingredient was further clarified with a set of questions researchers could address to satisfy the replication recipe (see Table 1 for selected questions).

Table 1 Replication recipe questions for mathematics education

Replication recipe ingredient*	Selected clarifying questions**	Analogous questions for mathematics education (revisions in bold)
1. Carefully defining the effects and methods that the researcher intends to replicate	A. Why is it important to replicate the effect I am trying to replicate? B. Where was the original study conducted? (e.g., lab, in the field, online) C. What kind of sample did the original study use?	Finding and method to be replicated A. Why is it important to replicate the original findings? B. What was the context of the original study? C. Who were the participants in the original study?
2. Following as exactly as possible the methods of the original study (including participant recruitment, instructions, stimuli, measures, procedures, and analyses)	A. Rank the replication as exact, close, or different in terms of similarities/differences a. … in the measures b. … in the procedure c. … between participant populations d. … in the location e. … of the analysis plan B. What differences between the original study and your study might be expected to influence the size and/or direction of the effect? C. What steps have been taken to test whether the differences listed will influence the replication outcome?	Alignment of replication A. Describe the similarities/differences between the original and replication a. … in the measures? b. … in the procedure? c. … between participant populations? d. … in the location? e. … of the analysis plan? B. What differences between the original study and your study might be expected to influence the trustworthiness and comparability of the replication findings? C. What steps have been taken to evaluate how the differences listed will influence the replication findings?
3. Having high statistical power	A. What is the target sample size? B. What is the rationale for the sample size (e.g., power of the design)?	Sample size and confidence in results A. How many participants will be recruited? B. Is the planned sample size large enough to warrant confidence in the results if the original findings are not replicated?
4. Making complete details about the replication available, so that interested experts can fully evaluate the replication attempt (or attempt another replication themselves)	A. Where can interested experts obtain the data and analysis syntax? B. Where can the reported analyses be obtained?	Transparency A. Where can interested experts obtain the data and analysis tools (e.g., coding rubrics)? B. Where can the reported analyses be obtained (e.g., coded data, resolution of coding disagreements, thematic summaries, analytic memos, etc.)?
5. Evaluating replication results and comparing them critically to the results of the original study	A. What is the effect size of the replication? B. What is the confidence interval of the replication effect size?	Evaluation of the significance and trustworthiness of replication findings in light of original study A. Describe the conclusions and the practical significance of the replication findings.

Table 1 (continued)

Replication recipe ingredient*	Selected clarifying questions**	Analogous questions for mathematics education (revisions in bold)
	C. Is the replication effect size significantly different from the original effect size? D. Is the replication a *success* (different from the null, and similar to or larger than the original and in the same direction), an *informative failure to replicate* (either not different from null or in the opposite direction from the original, and significantly different from original), a *practical failure to replicate* (both significantly different from the null and from the original), or *inconclusive* (neither significantly different from null nor the original)? E. What are the limitations of the replication study?	B. How trustworthy are the findings? C. Are the conclusions and the practical significance of the findings from the replication study decidedly different than those of the original study? D. Is the replication a *success* (clear results that strengthen the original conclusions), an *informative failure to replicate* (clear results that raise questions about the trustworthiness of the original conclusions), a *practical failure to replicate* (results that are clearly weaker than the original results and call for modified conclusions), or *inconclusive* (unclear results that neither strengthen nor raise questions about the original conclusions)? E. What are the limitations of the replication study?

*Brandt et al. (2014)

**Paraphrased from Brandt et al. (2014)

In applying the replication recipe to this study, we reflected upon how these ingredients and questions may not align with all current mathematics education scholarship. For instance, one of the questions clarifying ingredient 2 asked about "experimenter knowledge of participant experimental condition" (p. 219). The use of experimenter and experimental condition seems to imply a laboratory setting in which variables and conditions can be better controlled as opposed to a formal or informal learning environment in which it is difficult to control for participant differences or affective variables for example. Additionally, the replication recipe seems to be grounded in the positivist or postpositivist theoretical paradigm, in which the

> intent is to reduce the ideas into a small, discrete set of ideas to test, such as the variables that comprise hypotheses and research questions. The knowledge that develops…is based on careful observation and measurement of the objective reality that exists 'out there' in the world (Creswell 2008, p. 7).

We acknowledge that such a viewpoint contrasts with other theoretical viewpoints in mathematics education including social constructivism (e.g., Ernest

1998), postmodernism, (e.g., Walshaw 2004), and critical theories including critical race theory (e.g., Ladson-Billings and Tate 1995) and LatCrit theory (e.g., Gutiérrez 2013).

In spite of these limitations, we contend that the replication recipe, although geared to experimental, quantitative research methods of laboratory psychology, can be adapted for replication studies in qualitative and mixed methods research as well, and thus be useful for many (but certainly not all) of the studies conducted in mathematics education. Table 1 lists sample questions paraphrased from Brandt et al. (2014) alongside modified questions that are likely to be more broadly useful for mathematics education researchers. We next discuss each ingredient in detail, using our replication study of Thanheiser (2010) to illustrate how we adapted the original framework to be more useful for mathematics education research.

Applying the adapted replication recipe to the present study

We applied the replication recipe to design our own study and to write up the results by considering how we had addressed each ingredient. The first ingredient has to do with the finding and method to be replicated. We discuss the significance of the finding in the introduction of this paper and discuss the context of the original study, including a description of the participants, in the "Method" section. The second ingredient concerns the alignment of the replication, that is, how closely the replication study follows the original study. In the "Method" section, we also detail the ways in which we followed the original study methods as much as possible, and describe the places where the methods between the original and replication study differed. A main difference is the use of new codes in the replication study that were not used in the original study. We evaluate how these differences influence the trustworthiness of the findings in the "Discussion" section.

The third ingredient deals with sample size and confidence in the results. The sample size of this study (N = 79) is roughly 2.4 times more than the original study (N = 33). This strengthens our study in that "replications need 2.5 times as many observations as the original study to obtain about 80% power to reject a detectable effect" (Simonsohn 2014, p. 14). In other words, in designing the replication study, we planned for a large enough sample size to guard against incorrectly agreeing with the outcome of the original study because of too little information (Brandt et al. 2014). In the "Discussion" section, we describe the implications that the larger sample size has for confidence in the replication study findings.

The fourth ingredient describes transparency. To this end, we have included our code book in the "Method" section and provided detailed examples of how we coded and reconciled initial disagreements in coding. The fifth and final ingredient deals with the comparison of the original and replication findings and an evaluation of their significance and trustworthiness. In the "Results" section, we provided the results of several statistical comparisons between the descriptive findings of these two studies, and in the "Discussion" section, we address the significance of the similarities and differences. The replication

framework helped us clarify our goals for the replication study and our report of what we have found.

Method

Participants and context

This study was conducted at a large research university in the midwestern region of the USA. Participants for this study were 79 prospective elementary teachers (73 females, 6 males) enrolled in an elementary mathematics methods course during the 2017 Spring semester of their junior year. The course was designed to address two major goals: (1) know how elementary students think about and learn mathematics and (2) design instruction and classroom environments to help students understand key mathematical ideas of the curriculum. Prior to this course, prospective elementary teachers were required to pass two mathematics content courses—Teaching and Learning Elementary School Mathematics I and II. The first course focused on numbers and operations, while the second focused on geometry.

Similar to the original study, participants were prospective elementary teachers enrolled in an elementary mathematics methods course at a research institute located in the USA, albeit in the northwestern region of the USA. Additionally, PTs were required to pass two mathematics content courses before enrolling in the mathematics methods courses. It should be acknowledged that the content courses in the two studies more than likely varied in instructional practices, content focus, assignments, and so forth. Moreover, the PT participants in the original study enrolled in the methods course their fourth year of a 5-year program, while participants in the replication study were enrolled in the course the third year of a 4-year program.

Data source

Similar to the original study, the survey was administered to participants at the beginning of the semester prior to instruction of place-value concepts. Participants were asked to complete two tasks (see Fig. 2), which are exact replications of the tasks posed by Thanheiser (2010). The Addition Task was developed by Thanheiser (2009) and the Ones Task was developed by Philipp et al. (2008). In the original study, the tasks were administered via a paper and pencil instrument during class time. In this study, the tasks were part of a longer online survey administered outside of class time during the first week of class. The other survey items pertained to participants' beliefs about teaching and learning mathematics.

The intent of the Addition Task (Fig. 2) was to examine participants' explanation regarding the value of the regrouped digits. The intent of the Ones Task (Fig. 2) was to gauge if participants were able to articulate how the value of the regrouped digits in the addition algorithm and the subtraction algorithm were similar to and/or different from one another.

Data analysis

The first author attempted to apply the description of the conceptions framework and analysis available in published reports to the survey data but found a large number of cases that did not seem to clearly fit the available categories. One challenge was inferring the conceptions of participants who, for example, wrote about the digit 3 as "in the hundreds spot" while elsewhere writing about the same digit as if it was 3 ones. The first author contacted Thanheiser who shared an original coding rubric and discussed the coding process as well as describing the original process for reconciling ambiguous or edge cases (May 1, 2017, personal communication). A key insight from this conversation was the inference of a concatenated-digits conception with which participants spoke about the location of digits rather than the value of digits, as in the example described above. After this conversation, the first author again coded the survey data. To capture the insights and clarifications from the conversation with Thanheiser, the first author expanded the original coding rubric by adding more detail to each response option by selecting recurring words and phrases from participant's responses. The final coding rubric for the Addition Task is presented in Table 2. In providing example responses, we include numbers to align with the survey questions for each task. For instance, in the example below, (1) is the response from the first question within the Addition Task in Fig. 2 (i.e., What does the 1 above the 8 represent?), (2) is the response from the second question (i.e., What does the 1 above the 3 represent), and (3) is the response from the third question (i.e., Compare the two ones. Are they the same or are they different? Please be as specific as you can.)

As an example of the coding process, consider the following response for the Addition Task:

> (1) The ten from 14 when 9 and 5 were added. (2) The ten from 16 when 8 and 7 were added together. (3) No, when [sic] represents hundreds place while the other represents the tens place.

Addition Task

Please consider the regrouped ones in the problem below:

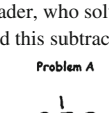
1 1
389
+ 475
864

- What does the 1 above the 8 represent?
- What does the 1 above the 3 represent?
- Compare the two 1s. Are they the same or are they different? Please be as specific as you can.

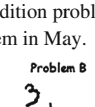
Ones Task

Below is the work of Terry, a second grader, who solved this addition problem and this subtraction problem in May.

Problem A



Problem B



- Does the 1 in each of these problems represent the same amount? Please explain your answer.
- Explain why in addition (as in Problem A) the 1 is added to the 5, but in subtraction (as in Problem B) 10 is added to the 2.

Fig. 2 Tasks and survey questions (Philipp et al. 2008; Thanheiser 2009)

This response was coded as *Both Ten* because the one is noted as representing a regrouped 10 in both addition problems. Although participant 11 did mention in the latter part that one represents the hundreds place and the other the tens place, the response implies place value is seen as a spot or location rather than referring to the value of the numbers. As another example, the following from participant 16 was coded as *Both One*:

> (1) It represents the one from the fourteen, when the 9 and five were added. They carried the one to the next place value. (2) The one over the three represents the one from the 16 after adding 8, 7, and the 1 over the eight. They carried it to the next place value. (3) They are the same because they represent the tens place from a two-digit number.

Although this participant mentions the tens place as a location in the third survey question, the emphasis is on one in the first two survey questions. Here, we gave more weight to the first two survey questions as a dominate way of thinking of place value.

The coding rubric for the Ones Task is presented in Table 3. As an example of the coding process, consider the following response, which was coded as *Ten, Hundred* because participant 57 emphasized the value of the numbers appropriately as tens in the addition problem and as hundreds within the subtraction problem.

> (1) No, the 1 in problem A represents 10 that should be added to the 5 and 3. The 1 in problem B is added to the 2 to make it 12. This represents 12 tens, or 120. (2) The one is added to the 5 because it represents the 10 that is taken from adding 9 and 8 to get 17. 10 is added to 2 because you would have to borrow from the hundreds in order to be able to subtract 3 from 2. The 1 represents 100, which would make 120 − 30.

As another example, consider the following response from participant 2:

> (1) The ones in these two problems represent different amounts because one is an addition problem and the other is a subtraction problem. In the addition problem the 1 represents just 1 being added to the 5 and 3. In the subtraction problem the 1 makes the 2 a 12 because it was borrowed from the 4. (2) In the addition problem 1 is added to 5 because the 1 is coming from the ones place value and therefore it loses the zero when it goes to the tens place. In comparison, in the subtraction problem 10 is added to 2 because the 1 is coming from the hundreds place value and just loses one zero when it moves over to the tens place value.

This response was coded as *Ones, Tens* as the participant distinguished the one in the addition problem as 1 and the one in the subtraction problem as 10.

As a means to establish initial reliability of the coding scheme, the second author coded ten participants' responses for both task. Agreement was established at 90%. Both members coded the remaining responses. Cohen's weighted kappa (Cohen 1968) for the remaining responses was established at

Table 2 Codebook for Addition Task

Codes for Addition Task	1 above 8 represents/shows/is…	1 above 3 represents/show/is…	The 2 ones are…
Ten, Hundred	• 10 carried from the ones place OR • The number 10 • A group of 10 • Adds 10 • The tens place	• 100 carried from the tens place • The number 100 • 80 + 70 = 160 • Regrouped 100 (from 160) • A group of 100 • Adds 100 • The hundreds place*	• Different because they represent different values • Different because they represent different numbers: 10 vs. 100 • Could use location/position language if value understanding is clear from other responses
Both Ten	• 10 (from the 14) • Tens place (from the 14) • The tens place*	• 10 (from the 16) • Ten carried over • A group of 10 ones • A regrouped 10 • The tens place*	• Same • Different because they are in a different position/location/place/column
Both One	• 1 (from the 14) • 10 carried over as 1 • 1 the tens place	• 1 (from the 16) • 1 carried from the tens place • 1 carried into the hundreds place • 10 carried over as 1 • 1 from a tens column • The tens or hundreds place	• Different because they represent different numbers: 14 vs. 16 • Same/different for non-relevant reason
No code	*No explicit naming as* a value (e.g., "hundred," "ten," or "one") within any response		

* Some PTs used place or location language in the first part(s), then used value language in the last part; we coded these as *Both Ten* or *Ten Hundred* rather than *One*

.82 for the Addition Task and .80 for the Ones Task; thus, inter-rater for both tasks is satisfactory or viewed as *almost perfect* by Landis and Koch (1977) as kappa was between 0.8 and 1.0.

Disagreements were discussed and agreed upon before proceeding with the analysis. To illustrate this process, consider the following response to the Ones Task by participant 32:

> (1) They both represent the same amount but since one is addition and one is subtraction the ones got there differently. (2) In the addition problem, the 1 is added to the 5 because 9 plus 8 is 17 which makes you have to carry the one. In the subtraction problem, since you are subtracting 3 from 2 you have to borrow from the 4 to make the 2 a 12.

One member of the research team coded as *Both Ones*, while the other *Ones, Tens*. We agreed to code as *Both Ones* because the participant claimed in the first response that "they both represent the same amount," while, in the second response, noted "you have to carry the one."

As another example, for the Addition Task, participant 46 stated that the ones are different because "they are over different placements. For example, the 1 over the eight is adding to the tens spot and the 1 over the 3 is adding to the hundreds place." One member coded as *Both Ones*, while the other member of the research team coded as *Ten, Hundred*. Our discussion was around the difference between place as a position and representing different values; therefore, we agreed to code as *Both Ones* as the emphasis is on 1 being added to a particular column in the standard algorithm.

Additionally, through this process, an additional code for the Ones Task was added to the coded scheme, *One, Not Explicit* (refer to Table 3). Moreover, we noticed a few instances in which participants were not consistent in their response within a task. In the Addition Task, for example, participant 17 stated the 1 above the 8 represents "carrying the ten's place from 9 + 5 = 14," while the 1 above the 3 represents "carrying the ten's place from 8 + 7 = 16." Yet, when comparing the two ones, this participant claimed, "The one above the 8 adds ten and the one above the 3 adds one hundred." We discussed how, although participants are both correct and incorrect in their response, they seem to exhibit flexibility in their thinking regarding reference units of multidigit numbers, and as such, acknowledged this by coding the higher of the two possible codes. In this case, this response was coded as *Ten, Hundred* as opposed to *Both Tens*.

After agreement was reached, we looked across all the responses of each participant, one at a time. Following Thanheiser (2010), we were, in this way, able to identify participants with correct responses (*Ten, Hundred*) on both tasks. These were attributed the groups-of-ones or the reference-units conceptions, but we did not specify which of these two conceptions because the survey questions could not be used to make this distinction. The rest of participants were attributed the concatenated-digits-only conception if they had responses coded as *Both Ones* for both survey tasks, and otherwise attributed the concatenated-digits-plus conception.

Again, following Thanheiser (2010), we split participants in the concatenated-digits-plus conception into three sub-conceptions by determining whether participants' conceptions of the digit 1 were consistent across the two contexts in the

Table 3 Codebook for Ones Task

Codes for Ones Task	1 in problem A represents/shows/is…	1 in problem B represents/shows/is…	Explanation
Ten, Hundred	• 10 (from the 17) • 1 carried over from the tens place so it is 10 • 10 added to the 50 and 30 • The tens place*	• 100 (from the 400) • 1 taken from the hundreds place, so it is 100 • 120 that the 30 could be taken away • The hundreds place*	• Value of numbers emphasized throughout, e.g., 10 + 50 − 30 (not 1 − F5 − F3) and 120 − 30 (not 12 − 3)
Both Ten	• 10 extra ones (in the tens place) • The extra group often ones (a ten) • 1 group of 10 • 15 tens the tens place*	• 10 extra ones (in the tens place) • Make the 2 into 12 • Regrouping of the base 10 column • 10 from 15 • 12 tens • The tens place*	• Carry 1 borrow tens • May say the ones represent different amounts but then call both by the same name (ones, tens) • May include place-value language as location (e.g., hundreds place/column)
Both One	• 1 represents 1 • Carried 1 • The tens place	• 1 borrowed • Carried 1 • The hundreds place	
One, Ten**	• 1 is carried over into the tens place • Adds a value of 1 to the column • 1 is added to 5	• Makes the 2 a 12 • 10 is added to the 2 • 1 as a tens place • Adds a value of "10" to the column	
One, Not Explicit	• 1 is carried over into the tens place • Adds a value of 1 to the column • 1 is added to 5	• Unclear: no explicit naming as a value	
Not Explicit	*No explicit naming as a value* (e.g., "hundred," "ten," or "one"); can be used for either value *or both values with any other code*		

* Some PTs used place or location language in the first part(s), then used value language in the last part; we coded these as *Both Ten* or *Ten Hundred* rather than *One*

** Both authors independently found a small number of cases in which the digit 1 was valued 10 in the first part and 1 in the second part of the Ones Task; we coded these *Ten One*

Addition Task and the two contexts in the Ones Task. We coded participants as *Consistently 10* if their responses to both tasks were *Both Tens*. In these cases, participants consistently conceived of the digit 1 as having the value 10 in all four contexts. We coded participants as *Consistent in addition* if they conceived of the digit 1 as the same value (either 1 or 10) in the addition contexts (i.e., both contexts of the Addition Task and the first context of the Ones Task), but held a different conception of the digit 1 in the subtraction context of the Ones Task. We coded all other participants as *Inconsistent*. For example, a participant who received a code of *Both Ones* on the Addition Task and a code of *One, Ten* on the Ones Task was coded as *Inconsistent*.

The last stage in our analysis was comparing the results of the replication study with those of the original study to determine to what extent the underlying distribution of conceptions and sub-conceptions was plausibly the same. We used the chi-square test of independence to compare the distribution of correct versus incorrect responses and to compare the distribution of coded responses for each task. We used the two-sample test of proportional equivalence to test whether the proportions of each conception and sub-conception were equivalent in the samples from the replication and original study.

Results

The results are organized to facilitate comparison with the original survey study (Thanheiser 2010). We have included each of the three tables from the original study with an extra column to report the analogous values we found in the replication study. We additionally report distributions of conceptions and sub-conceptions and compare these between the original and replication studies.

In Table 4, the number of participants with correct and incorrect conceptions of multidigit number is reported for the interview study (Thanheiser 2009), the original survey study (Thanheiser 2010), and this replication study. The percentages of participants with correct and incorrect conceptions fall between the corresponding percentages reported in the original interview and survey studies.

In Table 5, we report the distribution of codes for each task found in the replication study alongside analogous information from the original study. We used the chi-square test of independence to examine the distribution of the Addition Task and the Ones Task codes by study (original or replication). The relation between the study and the distribution of codes for the Addition Task was significant (χ^2 (2, N = 111) = 6.645, p = 0.036). Participants in the replication study were more likely to state appropriate values for the Addition Task and less likely to incorrect values than were the participants in the original study.

Next, we examined the relation between the study (original or replication) and the distribution of codes for the Ones Task. There were two challenges we faced in this analysis. First, our analysis of the Ones Task in the replication study produced several new codes that were not reported in the original study. As described in greater detail in the "Method" section, these codes captured variability in participant's responses that was not evident from the original study and more accurately reflected our uncertainty about some of the participants' conceptions. Second, many codes—including some of

the new codes—had only a small number of people, making the chi-square test methodologically inappropriate.

Our pragmatic solution to these problems was to group together all codes in which either digit or both digits were said to have a value of one (see noted codes in Table 5). This choice is theoretically defensible because such statements are evidence that the participant has the lowest level of conception, concatenated-digits-only (Thanheiser 2009), thus providing a meaningful comparison even though the codes differed somewhat between the original and replication study. It also allowed us to include many participants even though they were not explicit in their responses for all parts of the task. Moreover, this grouping makes the chi-square test more rigorous by ignoring some of the potential differences between code distributions by grouping codes together. Instead, we focused on differences in distribution among two codes and a group of related codes: appropriate values, both digits are ten, and codes in which one or both digits are valued as one. Importantly, this grouping strategy guarantees that observed differences in code distributions between studies cannot be attributed to the use of new codes in the replication study.

We found that the relation between study (original or replication) and the distribution of codes for the Ones Task was significant (χ^2 (2, N = 111) = 7.805, p = 0.020). Participants in the replication study were more likely to state that both values were ten for the Ones Task and less likely to state that one (or both) values were one than were participants in the original study. Thus, we found evidence that the distributions of codes for both tasks differed between the original and replication studies.

In Table 6, we report the distribution of code combinations across the two tasks. We have labeled each combination with a conception and sub-conception (if applicable). Because the sample in the replication was about 2.4 times the size of the original sample, table values for the replication study column are expected to be about 2.5 times as large as corresponding values for the original study.

We conclude this section by comparing the population estimate supporting the main claim from Thanheiser (2010) that most prospective teachers think of digits incorrectly at least some of the time. Recall that the results of the original study produce an estimate that only 9% of prospective teachers hold correct conceptions (95% CI (0.007, 0.189)). The estimate from the replication study is greater than that from the original and has a somewhat narrower confidence interval; we estimate that 18% of prospective teachers hold correct conceptions (95% CI (0.093, 0.261)).

Table 4 Number and percentage of participants by category in the original and replication studies

Category	Interview study[a] (N = 15), n (%)	Original survey study[b] (N = 33), n (%)	Replication of survey study (N = 79), n (%)
Correct conceptions (reference units or groups of ones)	5 (33)	3 (9)	14 (18)
Incorrect conceptions (concatenated-digits-plus or concatenated-digits-only)	10 (64)	30 (91)	65 (82)

[a] Thanheiser (2009)

[b] Thanheiser (2010)

Table 5 Number and percentage of participants by task code in the original and replication studies

	Original study[a] ($N = 33$), n (%)		Replication study ($N = 79$), n (%)	
	Addition Task	Ones Task	Addition Task	Ones Task
Original codes				
Appropriate values for both digits	8 (24)	4 (12)	39 (49)	15 (18)
Both digits are 10	13 (39)	12 (36	23 (29)	44 (56)
Both digits are 1[b]	12 (26)	8 (24)	16 (20)	3 (4)
Digit is 1 in addition; 10 in subtraction[b]		8 (24)		11 (14)
Digit is 10 in addition; 1 in subtraction[b]		1		1
New codes				
Digit is 1 in addition; 100 in subtraction[b]				1
Digit is 1 in addition; not explicit in subtraction[b]				3 (4)
Digit not explicit in addition; 10 in subtraction[c]				1
Digit value is not explicit in either context[c]			1	

[a] Thanheiser (2010)

[b] Codes with either digit or both digits valued as 1, grouped for chi-square analysis of Ones Task

[c] Excluded from chi-square analysis

The counts in Table 6 are too sparse to use the chi-square test of homogeneity, so we created a new table of the distribution of conceptions and sub-conceptions in each study (Table 7). This table does not have an analogue in Thanheiser (2010). The last column reports a two-sample test of proportion equivalence that compares the proportion of the conception or sub-conception in the replication and original samples. Importantly, none of these tests are statistically significant, which means the observed differences in proportions are not large enough to rule out the possibility that the underlying proportions are equal in the population. In particular, the difference between the estimates of prospective teachers with correct conceptions from the original and replication studies is not statistically significant ($\chi^2 = 0.760$, $p = 0.383$; see the first row of Table 7). This finding was particularly striking to us, because it seemed to contradict the earlier finding that the distribution of codes for each task was not the same across studies.

Discussion

We pursue two goals in the discussion of this study. First, we return to the research question and evaluate the results of replication study and their comparison with the original study. Second, we discuss limitations of the replication study and implications for other researchers interested in conducting replication work in mathematics education.

The replication results revealed that the proportion of correct conceptions and the distribution of sub-conceptions of the most prevalent incorrect conception (concatenated-digits-plus) did not have differences when compared with the original study results

to a greater extent than could be attributed to chance. On its face, this finding suggests a successful replication, that is, clear results that strengthen the original conclusions (see Table 1). However, the distribution of codes for each task was different between the two studies, suggesting that there may be underlying differences between the two populations. For example, almost half of the replication sample was coded as providing an appropriate response on Addition Task whereas only a quarter of the original sample was coded in this way. Because the concatenated-digits conceptions are defined to involve PTs who "see the digits incorrectly in terms of ones at least some of the time" (Thanheiser 2010, p. 246), these differences at the task level may be irrelevant for inferences about underlying conceptions of digits.

On the other hand, there is a plausible reason for these differences that has implications for teacher education. One factor that might explain the incongruity is the possibility that there was a clearer focus on the value of digits in addition contexts in mathematics content classes taken by the replication sample than that taken by the original sample. Although both programs required two college-level math courses,

Table 6 Distribution of participants by code pattern

Addition Task	Ones Task	Inferred conception and sub-conception	Original[a] ($N = 33$)	Replication ($N = 79$)
Appropriate values	Appropriate values	RU[b] or GO	3	14
	Both digits are 10	CDP, consistent in addition	3	23
	Both digits are 1	CDP, inconsistent	2	0
	Digit 1 in addition, 100 in subtraction*	CDP, inconsistent	0	1
	Digit not explicit in addition, 10 in subtraction*	CDP, inconsistent	0	1
Both as 10	Appropriate values	CDP, consistent in addition	1	1
	Both digits are 10	CDP, consistently 10	7	16
	Both digits are 1	CDP, inconsistent	1	1
	Digit 1 in addition, 10 in subtraction	CDP, inconsistent	3	4
	Digit 10 in addition, 1 in subtraction	CDP, consistent in addition	1	1
Both as 1	Both digits are 10	CDP, inconsistent	2	4
	Both digits are 1	CDO	5	2
	Digit 1 in addition, 10 in subtraction	CDP, consistent in addition	5	7
	Digit 1 in addition, not explicit in subtraction*	CDP, consistent in addition	0	3
Not explicit	Both as 10*	CDP, inconsistent	0	1

*New codes introduced in the replication study

[a] Thanheiser (2010)

[b] The four conceptions are reference unit (RU), groups of ones (GO), concatenated-digits-plus (CDP), and concatenated-digits-only (CDO)

Table 7 Distribution of participants by conception and sub-conception

	Original study[a] ($N = 33$), n (%)	Replication study ($N = 79$), n (%)	Test of equivalent proportions
RU* or GO	3 (9)	14 (18)	$\chi^2 = 0.760$; $p = 0.383$
CDP, consistently 10	7 (21)	16 (20)	$\chi^2 = 0.021$; $p = 0.885$
CDP, consistent in addition	10 (30)	32 (41)	$\chi^2 = 0.644$; $p = 0.422$
CDP, inconsistent	8 (24)	10 (13)	$\chi^2 = 1.537$; $p = 0.215$
CDO	5 (15)	2 (3)	Insufficient data for test

*The four conceptions are reference unit (RU), groups of ones (GO), concatenated-digits-plus (CDP), and concatenated-digits-only (CDO)

[a] Thanheiser (2010)

participants in the original study took "courses of their choice (not necessarily designed for elementary school teachers)" (Thanheiser 2010, p. 244), whereas participants in the replication study took two mathematics courses designed for elementary teachers.

The implication for teacher education is that mathematical preparation focused on the meaning of digits in addition contexts may not generalize to subtraction contexts, because the replication study participants performed much better on the Addition Task but did not perform substantially better on the Ones Task. Stated more precisely, the replication study estimates with high confidence that less than 30% of PTs who have taken the required number of mathematics courses in their teacher preparation program correctly conceive of the value of digits consistently across addition and subtraction contexts. Such a conclusion is similar to research of young children (Fuson et al. 1997; Selter 2001). As noted by Fuson et al. (1997), "multidigit subtraction seems to be more difficult for children than multidigit addition" (p. 151), and in the case of this study, multidigit subtraction seemed more difficult for PTs than multidigit addition. We contend that both addition and subtraction contexts be intermixed in mathematical courses for prospective elementary teachers (Fuson et al. 1997).

The apparent difference at the level of tasks codes, but agreement at the level of conception codes, led us to consider two issues related to the conceptions framework. The first has to do with how differences in a PT's responses are sometimes taken as evidence of flexibility and sometimes as evidence of a misconception in the multidigit conceptions framework. The hallmark of the most sophisticated conception in the framework—referent units—is the ability to think about digits in terms of different units (ones, tens, hundreds). Many PTs with the most prevalent conception—concatenated-digits-plus—also interpret digits in different ways depending on the context. We agree that to state that the value of 3 in the hundreds place is 3 tens is inaccurate. However, from the perspective of the standard algorithm, we also believe that it is perfectly reasonable to conceive of the digit 3 in the hundreds place as 3 tens *in the course of* using a subtraction algorithm to regroup. For the limited purpose of accurately regrouping within the algorithm, it is not necessary to recall the actual value of a digit. In fact, this is precisely the power of the standard algorithms: all digits can be thought of as tens and ones so that single-digit arithmetic can be leveraged to solve all multidigit problems. It seems possible that some PTs may simply not understand

questions that ask about the value of a digit, and they interpret the question in terms of the standard algorithm. Such PTs may be capable of thinking flexibly but not realize that such flexibility is what the task requires. Under this possibility, what seems like a limitation of PTs' conceptions may ultimately concern communication. Interview data is likely required to clarify these issues.

The second issue related to the conceptions framework has to do with the importance of differentiating the ideas of position versus value. In many of our coding decisions, we gave credit for a conception if PTs explicitly stated the value of a digit, but did not give credit for a sophisticated conception if PTs only referred to the position of the digit without writing directly in terms of its value. We believe based on the scholarship conducted by Thanheiser (e.g., 2009) and McClain (2003) that PTs with less sophisticated conceptions tend to rely on a calculational or algorithmic approach to multidigit addition and subtraction problems and often speak in terms of position rather than value. However, we wondered if PTs who neglected to speak in terms of value were incapable of such language or if they were simply using imprecise language. The question behind this pondering is about the nature of the conceptions described by the framework: are the links between conceptions and language stable and resistant to change or do some PTs speak and write in an ad hoc manner that does not always indicate clearly how they are thinking? More research, and in particular interview studies, would be useful to address both of the issues we have raised related to the conceptions framework.

We end this section by considering a limitation of our work that others who conduct replication studies in mathematics education are likely to face. To conduct the study, we had to operationalize the conceptions framework in order to code PT responses. We initially believed that this would be possible based on the published record alone. However, after we had collected the survey data we encountered edge cases that we could not decide based on the available description of the conceptions framework. In our discussion with the original author, we learned the importance of the distinction between position and value in the original work. Although we incorporated this distinction in our operationalization of the framework (see "Method"), we are not confident we did so in a way that is completely aligned with the original researchers. It is also possible that in spite of our best efforts, other factors of which we are unaware may have influenced how codes were determined. For example, the original team coded the survey responses after extensive experience in conducting interviews with the same questions, experience that we lacked, and this may have shaped their interpretation of participants' statements in ways that were not made explicit in our conversations with the original author. Thus, it is possible that the replication study was different than the original study simply because the researchers conducting the study brought different perspectives to the coding process.

A possible way for addressing this limitation would be to compare the inter-rater reliability between the two teams and to engage in discussion to reconcile differences. In general, however, the sustained contact that such an endeavor would require presents the potential to undermine the goal of conducting an independent replication study. If the new team of researchers is in effect trained by the original team, in what sense can the resulting research said to be independent? In the case of the present study, we mitigated this limitation by contacting the author of the original study to discuss a small

number of edge cases and by subsequently adapting and expanding an analysis document that was made available to us. However, we recognize arguments in favor of both for sustained training until a certain level of agreement has been reached between teams and for limiting contact between teams. Researchers who replicate mathematics education research that involves the coding of qualitative data are likely to face this dilemma.

The study opens several avenues for future research. The most interesting open questions, in our view, address the role of conceptions and sub-conceptions in teacher knowledge and learning. Thanheiser (2018) has recently argued that interviews based on the conceptions framework are one tool that could help PTs develop such meanings. We agree and further argue that the conceptions framework could be even more useful for teacher education if it was adapted or expanded to address the two issues we have raised above: clarifying how evidence warrants inferences of flexibility versus misconception and clarifying how the choice of language (e.g., position language versus value language) is construed as a conception (i.e., possible ways of thinking). Moreover, interview research is needed to understand whether conceptions of multidigit numbers are possible for PTs to learn within the context of teacher preparation or whether the conceptions framework describes fundamental differences in cognition that require extensive experience, as is likely the case with multiplicative and geometric concepts (e.g., Livy and Herbert 2013; Luneta 2014).

Further research on PTs' conceptions of multidigit numbers could also address theoretical questions about how the conceptions framework fits into the larger theoretical landscape including teacher learning and teacher knowledge. When considering teacher learning, we would like to know how the sub-conceptions in the framework are ordered either in terms of sophistication or in terms of a learning trajectory. For example, does the inconsistent sub-conception mark a conducive state for learning because it means PTs are likely to transition to a more sophisticated level, or does it suggest learning will be difficult and that special efforts are required on the part of the teacher educator? Another question we have is whether the consistent in addition sub-conception is homogenous when considering learning trajectories. This sub-conception includes both those who conceived of the digit 1 in the Addition Task as 10 and those that conceived of it as 1. Perhaps both groups within this sub-conception are not at the same place along a learning trajectory from less to more sophisticated conceptions of multidigit numbers.

Taking a broader view, we argue that more replication studies are needed in mathematics education, and more needs to be done to increase recognition of the value of this kind of work. Replication studies play a critical role in the scientific accumulation of knowledge. However, replication studies of education studies are rare; only about 1 in 1000 studies published in major education journals is a replication study (Makel and Plucker 2014).

Similarly, replication and replication-like studies of mathematics education studies published between 1990 and 2017 are rare. In a search of prominent mathematics education journals (Nivens and Otten 2017) including *Journal of Research in Mathematics Education*, *Journal of Mathematics Teacher Education*, *Educational Studies in Mathematics*, *Research in Mathematics Education*, *Mathematics Education Research Journal*, *Journal of Mathematical Behavior*, and *International Journal of Science and Mathematics Education*, we found a

total of 18 replication studies, 10 of which were replications of their own studies (e.g., Mejía-Ramos and Inglis 2011; Thornton 1990). Therefore, the field of mathematics education has published only eight independent replication studies in these journals over the last two and a half decades, or about one every 3 years, on average.

The rarity of replication extends even further in the past as documented by Eastman in an article published in the *Journal for Mathematics Education Research* in 1975. He listed three reasons he believed replications were not common then; we believe two are pertinent today: "(a) Researchers are not conducting replication studies. (b) Researchers are conducting replication studies but are not submitting them because they believe the studies are not 'worthy' of publication" (p. 67). Increasing capacity in the field to conduct replication studies, recognition of their value and outlets for publication are all critical to increasing the replication studies in mathematics education.

Conclusion

Replication is not a straightforward or simple endeavor in mathematics education. As the present study illustrates, even replication of quantitative studies can require researchers to apply coding schemes which inherently require interpretation. There is great need for improved methodology for replication that is tailored to the specific needs of the field of mathematics education and educational studies more broadly. In addition, standards of the field should be in place that guarantee published work supports the possibility of future replication, for example by establishing norms for reporting reproducible processes for reconciling disagreements when qualitative data is coded. Journals could require that analysis documents are archived on a public platform such as the Open Science Framework (OSF, osf.io) to increase transparency. We hope this study contributes to the field by improving motivation and capacity to conduct replication studies in mathematics education.

References

Borko, H., Eisenhart, M., Brown, C. A., Underhill, R. G., Jones, D., & Agard, P. (1992). Learning to teach hard mathematics: do novice teachers and their instructors give up too easily? *Journal for Research in Mathematics Education, 23*, 194–222.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., et al. (2014). The replication recipe: what makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217–224. https://doi.org/10.1016/j.jesp.2013.10.005.

Coyne, M. D., Cook, B. G., & Therrien, W. J. (2016). Recommendations for replication research in special education: a framework of systematic, conceptual replications. *Remedial and Special Education, 37*, 244–253. https://doi.org/10.1177/41932516648463.

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220.

Creswell, J. W. (2008). *Research design: qualitative, quantitative, and mixed methods approaches* (3rd ed.). Thousand Oaks: Sage.

da Ponte, J. P., & Chapman, O. (2015). Prospective mathematics teachers' learning and knowledge for teaching. In L. D. English & D. Kirshner (Eds.), *Handbook of international research in mathematics education* (pp. 275–296).

Eastman, P. M. (1975). Replication studies: why so few? *Journal for Research in Mathematics Education, 6*(2), 67–68.

Ernest, P. (1998). *Social constructivism as a philosophy of mathematics*. Albany: SUNY.

Fuson, K. C., Wearne, D., Hiebert, J. C., Murray, H. G., Human, P. G., Olivier, A. I., Carpenter, T. P., & Fennema, E. (1997). Children's conceptual structures for multidigit numbers and methods of multidigit addition and subtraction. *Journal for Research in Mathematics Education, 28*, 130–162.

Gutiérrez, R. (2013). The sociopolitical turn in mathematics education. *Journal for Research in Mathematics Education, 44*(1), 37–68.

Ladson-Billings, G., & Tate, W. F. (1995). Toward a critical race theory of education. *Teachers College Record, 97*, 47–68.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Leinhardt, G. (1989). Math lessons: a contrast of novice and expert competence. *Journal for Research in Mathematics Education, 20*, 52–75.

Livy, S., & Herbert, S. (2013). Second-year pre-service teachers' responses to proportional reasoning test items. *Australian Journal of Teacher Education (Online), 38*(11), 17–32. https://doi.org/10.14221/ajte.2013v38n11.7.

Luneta, K. (2014). Foundation phase teachers' (limited) knowledge of geometry. *South African Journal of Childhood Education, 4*(3), 71–86.

Ma, L. (1999). *Knowing and teaching elementary mathematics: teachers' understanding of fundamental mathematics in China and the United States*. Mahwah: Erlbaum.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: replication in the education sciences. *Educational Researcher, 43*(6), 304–316. https://doi.org/10.3102/0013189X14545513.

McClain, K. (2003). Supporting preservice teachers' understanding of place value and multidigit arithmetic. *Mathematical Thinking and Learning, 5*(4), 281–306.

Mejía-Ramos, J. P., & Inglis, M. (2011). Semantic contamination and mathematical proof: can a non-proof prove? *The Journal of Mathematical Behavior, 30*, 10–29. https://doi.org/10.1016/j.jmathb.2010.11.005.

Nivens, R. A., & Otten, S. (2017). Assessing journal quality in mathematics education. *Journal for Research in Mathematics Education, 48*(4), 348–368.

Philipp, R., Schappelle, B., Siegfried, J., Jacobs, V., & Lamb, L. (2008). *The effects of professional development on the mathematical content of K-3 teachers*. New York: Paper presented at the American Education Research Association.

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. https://doi.org/10.1037/a0015108.

Selter, C. (2001). Addition and subtraction of three-digit numbers: German elementary children's success, methods and strategies. *Educational Studies in Mathematics, 47*(2), 145–173.

Simonsohn, U. (2014). Small telescopes: detectability and the evaluating of replication results. *Psychological Science*. Retrieved at. https://doi.org/10.2139/ssrn.2259879.

Stanic, G. M., & Kilpatrick, J. (1992). Mathematics curriculum reform in the United States: a historical perspective. *International Journal of Educational Research, 17*(5), 407–417.

Thanheiser, E. (2018). The effects of preservice elementary school teachers' accurate self-assessments in the context of whole number. *Journal for Research in Mathematics Education, 49*(1), 39–56.

Thanheiser, E. (2010). Investigating further preservice teachers' conceptions of multidigit whole numbers: refining a framework. *Educational Studies in Mathematics, 75*(3), 241–251.

Thanheiser, E. (2009). Preservice elementary school teachers' conceptions of multidigit whole numbers. *Journal for Research in Mathematics Education, 40*(3), 251–281.

Thornton, C. A. (1990). Solution strategies: subtraction number facts. *Educational Studies in Mathematics, 21*(3), 241–263.

Walshaw, M. (Ed.). (2004). *Mathematics education within the postmodern*. Greenwich: Information Age.