Mathematics and Science Teachers’ Use of and Confidence in Empirical Reasoning: Implications for STEM Teacher Preparation

Nicholas H. Wasserman
Teachers College, Columbia University

Dara Rossi
Southern Methodist University

*The recent trend to unite mathematically related disciplines (science, technology, engineering, and mathematics) under the broader umbrella of STEM education has advantages. In this new educational context of integration, however, STEM teachers need to be able to distinguish between sufficient proof and reasoning across different disciplines, particularly between the status of inductive and deductive modes of reasoning in mathematics. Through a specific set of mathematical conjectures, researchers explored differences between mathematics (*n *= 24) and science (*n *= 23) teachers’ reasoning schemes, as well as the confidence they had in their justifications. Results from the study indicate differences between the two groups in terms of their levels of mathematical proof, as well as correlational trends that inform their confidence across these levels. Implications particularly for teacher training and preparation within the context of an integrated STEM education model are discussed.*

In recent years, the national discourse about the teaching of mathematically related disciplines (science, technology, engineering, and mathematics), and ways to improve it, have been unified under the broader umbrella of STEM education. While there are various interpretations of what STEM education should look like in practice, many argue that genuine integration of the various disciplines is a key premise for STEM education (e.g., Johnson, 2012). In many ways, this blending is useful, as these four draw on and apply mathematical ideas to solve problems in ways that other disciplines do not. Such educational integration is both practical because of the common foundations but also potentially impractical because of their differences. Each of the four disciplines has philosophical and epistemological distinctions: the ways in which disciplinary truths are established and validated are different. In particular, considering two of the more traditional STEM disciplines—science and mathematics—the notion and status of deductive mathematical proof, as opposed to empirical arguments, is one such distinction.

Teachers are the primary vehicles for students’ learning about each of these disciplines. Consequently, meaningful integration of STEM content (to take advantage of their connections) and meaningful separation (to provide disciplinary integrity) is a challenge that teachers will be responsible for navigating in the classroom. As middle and secondary teachers are often prepared as experts in a specific STEM field, their disciplinary training may pose some difficulties with regard to the knowledge and practices for managing the tension between integration and disciplinary integrity—especially in the broader realm of reasoning and justification, where mathematics and science in particular have epistemological distinctions. In this mixed methods study, we compare mathematics and science teachers’ approaches for and confidence in validating a set of mathematical conjectures in order to explore broader implications for STEM preparation with regard to distinguishing disciplinary reasoning and justification.

Background

STEM Education

In the past decades, publicized rankings for American students in mathematics and science on a number of international tests (e.g., Program for International Student Assessment [PISA] 2003 [Lemke et al., 2004]; Trends in International Mathematics and Science Study [TIMMS] 2003 [Gonzales et al., 2004]) have led to calls for improving instruction in these areas. Partly stemming from these results, “STEM education” has come to embody the necessity to improve education in science, technology, engineering, and mathematics, and, generally, to prepare more students for careers in these burgeoning fields. Yet the interpretation and implementation of what STEM education means in theory and practice varies widely (e.g., Breiner, Harkness, Johnson, & Koehler, 2012). The California Department of Education (2013) echoes this discord, stating on their Web site that STEM education could be “. . . a stand alone course, a sequence of courses, activities involving any of the four areas, a STEM-related course, or an interconnected or integrated program of study.”

While it is the case that STEM education for some may only mean enhancing the teaching of these individual content areas, the acronym itself perhaps implies some genuine integration. In a recent editorial for a special issue in the *School Science and Mathematics* journal, Johnson (2012) argues that meaningful integration between the disciplines, a "paradigm shift for most" (p. 1), is a key premise for STEM education. Others echo this emphasis: the California STEM Learning Network (2012), for example, states that STEM education is more than just individual disciplines, and indicates an interdisciplinary and applied approach to teaching these subjects. For them, "STEM education removes the traditional barriers erected between the four disciplines by integrating them into one cohesive teaching and learning paradigm." The Dayton Regional STEM center, an institute for professional practice in STEM fields, coordinates a network of institutions and professionals to help provide resources and professional development on STEM instruction. They, too, see STEM as an integration of these disciplines. As a part of their work, they created a framework and rubric that described ten components that help characterize the quality of STEM learning (Dayton Regional STEM Center, 2013). Two of these components are particularly relevant to the discussion at hand: the degree of STEM integration and the integrity of the academic content. Regarding integration, "quality STEM learning experiences are carefully designed to help students integrate knowledge and skills from Science, Technology, Engineering, and Mathematics"; regarding integrity of content, "quality STEM learning experiences are content-accurate, anchored to the relevant content standards, and focused on the big ideas and foundational skills critical to future learning in the targeted discipline(s)." Evident from this framework, STEM education requires both meaningful integration and also disciplinary integrity.

Discipline-based education research, which has developed predominantly in undergraduate sciences and engineering, aims to understand the ways people learn and develop expertise in specific STEM disciplines. One of the premises is that across different domains just within science, such as physics, biology, geoscience, and chemistry (let alone across the entirety of STEM), the teaching and learning in these fields require teachers to have deep, discipline-specific knowledge that is distinct even from the other science-related disciplines (National Research Council, 2012). Shulman (1986) similarly maintained the necessity for teachers to have deep and profound subject-specific knowledge, understanding the structure, composition, and principles of inquiry in the discipline. Acquiring such depth of knowledge across multiple STEM subject areas, let alone in one content area, poses potential challenges as well as significant demands on those charged with instruction.

Philosophical Distinctions

The sciences use and apply a significant amount of mathematics. Results in mathematics have often yielded new and interesting conclusions about the world, informing scientific insights. Thus, there is a significant amount of mathematics and mathematical ideas that are part of the learning and teaching of science. Indeed, the road goes both ways: many of the applications discussed in mathematics classes draw on various science-related settings. Yet while mathematics is often classified as a science, and science frequently utilizes mathematics, the two have philosophical differences. In an introduction to the philosophy of mathematics, Horsten (2012) touches on three such distinctions between mathematics and other sciences: the entities of interest, the acquisition of knowledge, and the status of knowledge. These shape some of the ontological and epistemological differences between the two disciplines.

While there are varying definitions for science, according to the Next Generation Science Standards (2013), for example, the sciences are devoted to studying and explaining the structure and behavior of the natural and material world—from gravity to the composition of matter, planetary motion to the neurological workings of organisms. Thus, the objects of interest in science exist in the real world. As Devlin (2003) points out, however, the entities that form the substance of mathematics (e.g., points, lines, numbers, and functions) do not exist in the physical world. They are representations and abstractions of ideas that exist in the mind. This disciplinary difference between the objects of interest has implications for how knowledge is acquired and validated. Given the desire to study phenomenon in the world, the criterion for whether or not an idea or theory in science is valid is based on empiricism or observation (e.g., Rosenberg, 2000). The scientific method of forming a hypothesis, isolating the phenomenon under study, and conducting experiments—in particular those that attempt to falsify a claim (e.g., Popper, 1963)—is critical to developing new knowledge in the sciences, as it provides observable instances that support or refute a claim. As such, one of the primary modes of reasoning in science is induction—forming general conclusions based on specific observations. (Although, while perhaps most common, this is not the only form of reasoning; scientists frequently use observational generalizations as the premise for deducing [i.e., deductive reasoning] other results [Losee, 1972].) On the other hand, mathematics is not about observations, but abstractions, making the veracity of statements or claims predicated on logical consequences from definitions and axiomatic systems. As such, one of the primary modes of reasoning in mathematics is deduction—deducing specific conclusions that must be true based on a set of more general and agreed-upon statements. (Although, again, while perhaps most common, this is not the only form of reasoning; mathematicians frequently use examples and cases [i.e., inductive reasoning] as a way to understand and approach a problem.) Lastly, the differences in acquiring knowledge lead to different standings of knowledge: mathematical claims are certain and absolute whereas scientific ones are not. In particular, deductive reasoning from a set of agreed upon axioms and statements leads to results that will always be true, whereas inductive conclusions, piecing together explanations based on observing results, will not always hold. Epistemological differences between the means of investigating and validating claims, as well as the certainty of claims, capture some of the differences between mathematics and science.

Mathematics and science are connected and intertwined, mutually helpful for advancing knowledge in each one. Indeed, the natural interplay and exchange among the disciplines requires a degree of familiarity with and understanding of the common ground between them. Yet some of the disciplinary distinctions—particularly the status of empirical evidence for validating claims—may influence the approaches each would use for justifying specific claims, as well as the degree of confidence in those approaches. For integrated STEM curricula, teachers will be responsible for navigating these differences in ways that remain true to each discipline.

Taxonomies for Reasoning in Mathematics

In mathematics education, there has been a renewed focus on student reasoning and sense making in the classroom, as advocated for by prominent organizations such as the National Council for Teachers of Mathematics (e.g., National Council of Teachers of Mathematics, 2000). While deductive proof based on axioms is the standard for validating mathematical ideas in the discipline, the emphasis on reasoning in mathematics education has expanded beyond this single focus of justification. Indeed, many have studied and debated the role that proof should play in mathematics education (e.g., Chazan, 1993; Knuth, 2002; Stylianides, 2007). As a part of some of the work on proof, different taxonomies describing the sophistication of proof schemes have developed.

Balacheff (1988) articulated four levels of increasing sophistication of proof based on observations of secondary school students. The first three levels of Balacheff's proof scheme are all based on empirical observations with examples. He described naïve empiricism (a) as arriving at a conclusion about a universal assertion based on someone citing a small number of examples. For example, referencing a regular square, hexagon, and octagon as evidence for the truth of a statement about all regular polygons. A crucial experiment (b) similarly draws on examples to make claims about a universal assertion, the difference being a test of a deliberately chosen example. For example, in addition to testing cases for $x = 1$, 2, and 3, someone may test the case $x = 100$ or $x = -50$ to verify that it also works with larger or negative values. The third level, a generic example (c), was characterized by Balacheff as justification through the use of a particular example, but one that is used as representative of a larger class of objects. In other words, someone may use a specific example as illustrative of more generic reasoning. The last level of this proof scheme was a thought experiment (d), where someone draws logical deductions based solely on the properties and relationships of the situation. This last level typifies deductive reasoning representative of the broader discipline and mathematical proof.

Harel and Sowder (1998) similarly used exploratory studies about students' proof schemes to inform the development of their taxonomy. They characterized three broad categories of increasing sophistication: external conviction proof schemes, empirical proof schemes, and analytical proof schemes. Within each of these categories, they provided subcategories to further distinguish the ways that students attempt to prove. External conviction schemes, the lowest level of proof scheme, are characterized by reliance on authority figures, ritualistic arguments, or symbolic form in determining the validity of mathematical ideas. For example, students may rely on a teacher or textbook as an authority figure for certain facts, or they may believe something to be convincing simply because it is written in a standard proof form. Empirical proof schemes are those that rely on examples to answer universal assertions. Harel and Sowder further distinguish between an inductive empirical and a perceptual empirical proof scheme. The inductive proof scheme uses specific cases to validate a universal assertion, whereas a perceptual scheme is based on a rudimentary mental image of the general case that limits complete relational understanding. Lastly, analytical proof schemes are those that utilize logical deductions to arrive at conclusions: both axiomatic proof and reasoning beginning from specific terms and axioms, as well as transformational proof schemes, based on full mental images and appropriate mental operations and transformations, fall under this category. These

Table 1
Participants

Mathematics Teachers (n = 24)		Science Teachers (n = 23)	
17 middle school 7 high school	4 male 20 female	7 middle school 16 high school	9 male 14 female
Average years of teaching experience: 7.17		Average years of teaching experience: 11.91	

works on proof schemes in mathematics education have helped inform a hierarchy of proof and reasoning in the discipline.

Methods

From the literature, we find that key premises for STEM education include both meaningful integration and disciplinary integrity, that mathematics and science are simultaneously inherently connected yet epistemologically distinct, and that deductive proof, as opposed to inductive arguments, is a necessary but frequently difficult part of mathematical learning. There seems to be significant merit and potential to an integrated approach to STEM learning; however, broad disciplinary distinctions, in particular the status of empirical arguments in mathematics and science (i.e., empirical arguments are less sophisticated in mathematics than deductive approaches), pose subsequent demands on teachers for navigating integrated STEM curricula that also maintains the integrity of each discipline.

This study explores some of this tension, in particular, whether mathematics and science teachers approach justifying mathematical conjectures in similar or different ways. For current mathematics and science teachers trained primarily in their individual disciplines, would the strong mathematical basis and frequent use of mathematics in science result in similar approaches to mathematical conjectures, or would the broad epistemological differences between mathematics and science lead to differing approaches? Also, because teachers are responsible for choosing how to justify the truth of statements in class, do they report similar or different degrees of confidence in empirical arguments as convincing justification in mathematics? Such questions, while not indicative of pedagogical style, do provide insight about the potential likelihood of using empirical or deductive approaches for explaining mathematical ideas. Using a set of three mathematical conjectures, two researchers compared how mathematics and science teachers approached validating these claims, with participants' self-reported confidence scores determining the degree to which they found empirical arguments to be mathematically convincing. Results have implications for STEM teacher preparation.

In particular, the study uses a mixed methodology to address two research questions: (a) Are there differences in the reasoning schemes mathematics and science teachers use to validate mathematical ideas—in particular, how much do they rely on empirical evidence? (b) How confident are mathematics and science teachers in their use of empirical evidence versus deductive arguments as being mathematically convincing?

Participants

Practicing middle and secondary mathematics (n = 24) and science teachers (n = 23) were recruited to participate in the study. The teachers, who volunteered to participate, were primarily recruited from two graduate programs in mathematics education and science education at a mid-sized private university in a large urban metroplex. This sample of teachers, connected to graduate study in mathematics or science education, was utilized because each program's entry requirements would be similar and such teachers frequently have strong disciplinary training (in mathematics or science), which was of particular interest to the study. Overall, the participants in both groups were predominantly female, with comparable years of experience teaching; the proportion of middle and high school teachers were approximately flipped between the groups. Table 1 presents a breakdown of the participants.

Framework

The aim of the study was to understand if whether being a mathematics or science teacher influenced the type of reasoning engaged in when justifying a mathematical conjecture (particularly differences between inductive and deductive arguments), as well as the teachers' confidence in such reasoning. Based on the taxonomies of proof in mathematics, we would anticipate higher taxonomical classifications to align with higher degrees of confidence, and lower taxonomical classifications to align with lower degrees of confidence. Additionally, based on epistemological distinctions, we hypothesized that science teachers may be more prone to inductive means of justifying a statement than the mathematics teachers, and also would have more confidence in this form of reasoning. Figure 1 depicts the research framework.

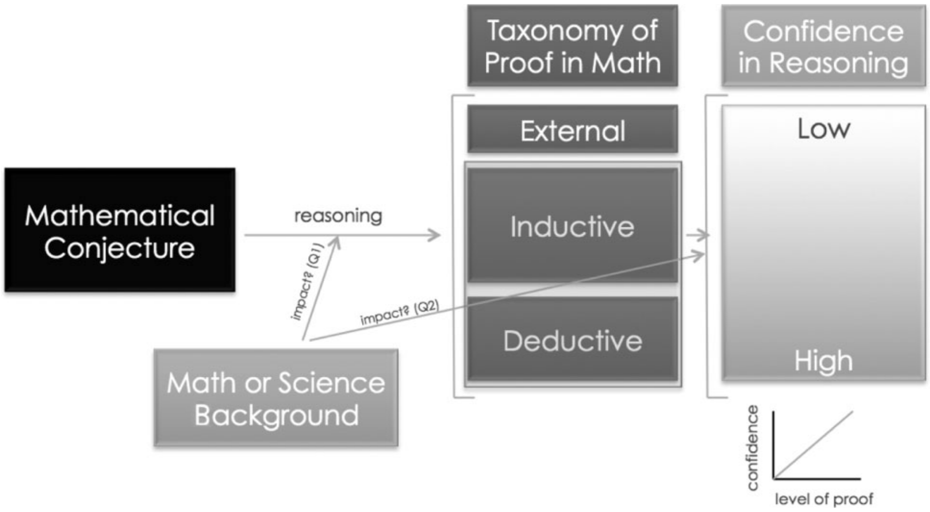
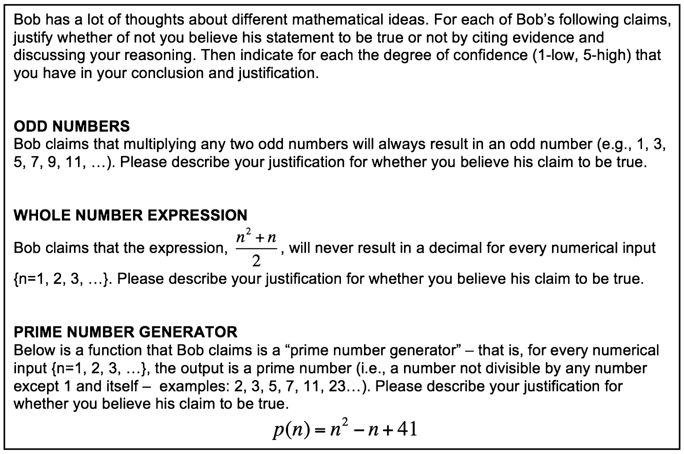


Figure 1. Research framework.

Task Design

To study if having a background as a mathematics or science teacher influenced the reasoning used to validate a mathematical idea, the researchers designed a set of three conjectures, pilot testing them with a small sample of teachers and making minor revisions accordingly. The three conjectures were all universally quantified statements (not existential statements) to examine the reliance on empirical versus deductive reasoning schemes. In addition, each of the mathematical conjectures was designed so that empirical evidence from testing specific cases was a particularly appealing approach for explanation in order to study the confidence participants may lend to such reasoning as well as whether participants moved beyond this initial approach. They dealt with familiar concepts, but the claims themselves likely were increasingly unfamiliar. The three tasks (odd numbers, whole number expression, and prime number generator) are presented in Figure 2. Two conjectures were true and one was false; however, the false claim (prime number generator) was true for many cases before a counterexample emerged. The design elements also aimed to help isolate potential differences in the levels of confidence that the participants gave to empirical evidence; they were asked to indicate the degree of confidence (on a scale from 1 to 5) in their justification for each task.

Participants were given instructions about the tasks and then asked to complete them individually. While they were not given a time limit, participants spent approximately 25–30 minutes responding to the conjectures.


Bob has a lot of thoughts about different mathematical ideas. For each of Bob's following claims, justify whether of not you believe his statement to be true or not by citing evidence and discussing your reasoning. Then indicate for each the degree of confidence (1-low, 5-high) that you have in your conclusion and justification.

ODD NUMBERS
Bob claims that multiplying any two odd numbers will always result in an odd number (e.g., 1, 3, 5, 7, 9, 11, …). Please describe your justification for whether you believe his claim to be true.

WHOLE NUMBER EXPRESSION
Bob claims that the expression, $\frac{n^2+n}{2}$, will never result in a decimal for every numerical input {n=1, 2, 3, …}. Please describe your justification for whether you believe his claim to be true.

PRIME NUMBER GENERATOR
Below is a function that Bob claims is a "prime number generator" – that is, for every numerical input {n=1, 2, 3, …}, the output is a prime number (i.e., a number not divisible by any number except 1 and itself – examples: 2, 3, 5, 7, 11, 23…). Please describe your justification for whether you believe his claim to be true.

$$p(n) = n^2 - n + 41$$

Figure 2. Set of three mathematical conjectures.

Data Analysis

The 47 participants yielded a total of 141 responses over the set of three conjectures. The researchers adapted Balacheff's (1988) and Harel and Sowder's (1998) proof taxonomies, synthesizing and modifying their ideas into a framework to analyze the data (Table 2). With the exception of two categories (flaw and partial analytical), the code names and descriptions reference and combine the previous discussion of the two taxonomies (e.g., Crucial inductive empiricism blends Balacheff's crucial experiment and Harel and Sowder's inductive empirical scheme). The two additional modifications were included specifically for coding responses from this study: in particular, some participants demonstrated false reasoning based on a flawed understanding of the mathematical statement (flaw), and others made deductions based on

Table 2
Coding Scheme

Code		Score	Description
Flaw (F)			Flawed understanding of the mathematical statement; results in incorrect reasoning and conclusions about the problem
External	External conviction (E)	0	Reasoning is linked to external authoritative statements
Example-based evidence (inductive)	Naïve inductive empiricism (N)	1	Arriving at a conclusion based only on a small number of particular examples. May demonstrate reasoning based on limitations in examples chosen.
	Crucial inductive empiricism (C)	2	After looking at particular examples, justifies claim by examining a case that is nonparticular (a *deliberate* choice is made in the selection of the example)
	Perceptual empirical example (G)	3	Uses a particular example as representative of the general situation and performs operations/transformations on the example to arrive at a justification; the generic example is based on a rudimentary image, limited in some way that makes the argument incomplete
Deductive reasoning	Partial analytical (P)	3	Analyzes and makes deductions based on partially complete reasoning; relies on familiar knowledge as "axioms"
	Thought experiment (T)	4	Logical deductions based on awareness of the properties and relationships of the situation
	Axiomatic proof (A)	4	Logical deductions based on rigorous axioms and definitions; correct use of a counterexample

partial generalizations and familiarity with the content (partial analytical). The numerical scores are consistent with the differing degrees of sophisticated arguments based on these frameworks, with perceptual empirical example and partial analytical both making arguments that rely on properties of the situation, although limited in some way, and both thought experiment and axiomatic proof being considered complete deductive arguments.

Coding

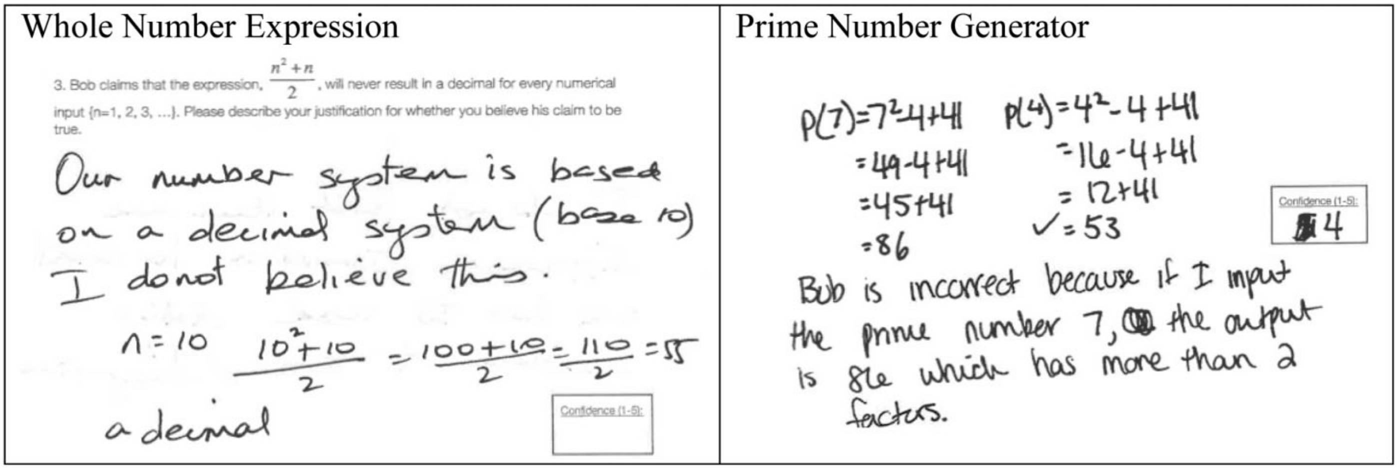
For this study, the two researchers adopted a collaborative coding method (Harry, Sturges, & Klingner, 2005) in order to ensure that scores were consistent and that proper attention and consideration were paid to all aspects of participants' written work. The goal became consensus, not simply comparable independent coding, while debating and determining appropriate proof codes for each of the responses. Compared with individual coding, this process was particularly beneficial given the difficult nature of determining the scope and validity of arguments with the evidence presented, as well as with mutually identifying some of the minute, although important, mathematical errors. In order to give a better sense of the types of responses that fell within each code, we discuss some representative examples.

Flaw (F). While working on a conjecture, some participants made a mistake interpreting the problem that led to incorrect reasoning about the conjecture. These came primarily in two forms: incorrect interpretation and incorrect calculation. Figure 3 documents one participant's incorrect interpretation of a "decimal" output, showing 55 (likely because it can be expressed as 55.0) as a counterexample to the conjecture, and an incorrect calculation, as one participant accidentally evaluated $n^2 - 4 + 41$ instead of $n^2 - n + 41$.

External conviction (E). While there were only three cases coded as external conviction, they were all declarations with no justification besides some statement of fact. For example, on the odd numbers conjecture, one participant simply wrote: "I believe this is true . . . because it is mathematical theory/law." The response indicates that the participant believed this fact, without further need for verification.

Naïve (N) and crucial empiricism (C). The primary difference between naïve and crucial empiricism was evidence of intentional choices being made in the selection of example cases. A response was labeled crucial empiricism when, after testing a simple case or two, another example was used as a "test case" to verify the conjecture. Sometimes this looked like trying a large number or a random number; in other situations, this looked like trying a qualitatively different example (e.g., "even if non-prime numbers, like 6, are used it still works"). In contrast, when only a small set of examples, with no evidence of deliberate choices in the selection of test cases was presented, it was labeled naïve empiricism (Figure 4).

Perceptual empirical example (G). A perceptual empirical example was evident when participants made general claims by using a specific example to illustrate. In these responses, a participant operated on the specific example in his/her attempt to convey the more general



Whole Number Expression

3. Bob claims that the expression, $\frac{n^2+n}{2}$, will never result in a decimal for every numerical input {n=1, 2, 3, ...}. Please describe your justification for whether you believe his claim to be true.

Our number system is based on a decimal system (base 10) I donot believe this.

$n=10 \quad \frac{10^2+10}{2}=\frac{100+10}{2}=\frac{110}{2}=55$

a decimal

Confidence (1-5):

Prime Number Generator

$P(7)=7^2 4+41$ $\quad P(4)=4^2-4+41$

$=49-4+41$ $\quad =16-4+41$

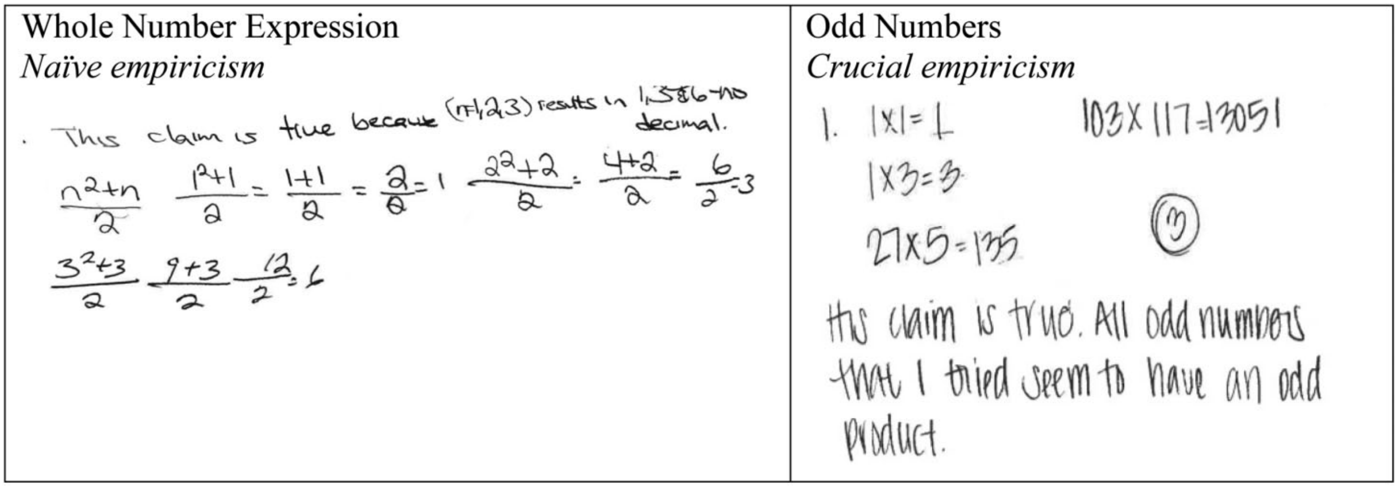
$=45+41$ $\quad =12+41$

$=86$ $\quad ✓=53$

Confidence (1-5): 4

Bob is incorrect because if I input the prime number 7, the output is 86 which has more than 2 factors.

Figure 3. Flaw examples.



Whole Number Expression

Naïve empiricism

. This claim is true because (n=1,2,3) results in 1,3&6-no decimal.

$\frac{n^2+n}{2} \quad \frac{1^2+1}{2}=\frac{1+1}{2}=\frac{2}{2}=1 \quad \frac{2^2+2}{2}=\frac{4+2}{2}=\frac{6}{2}=3$

$\frac{3^2+3}{2} \; \frac{9+3}{2} \; \frac{12}{2}=6$

Odd Numbers

Crucial empiricism

1. $1\times1=1$ $\quad 103\times117=13051$

$1\times3=3$

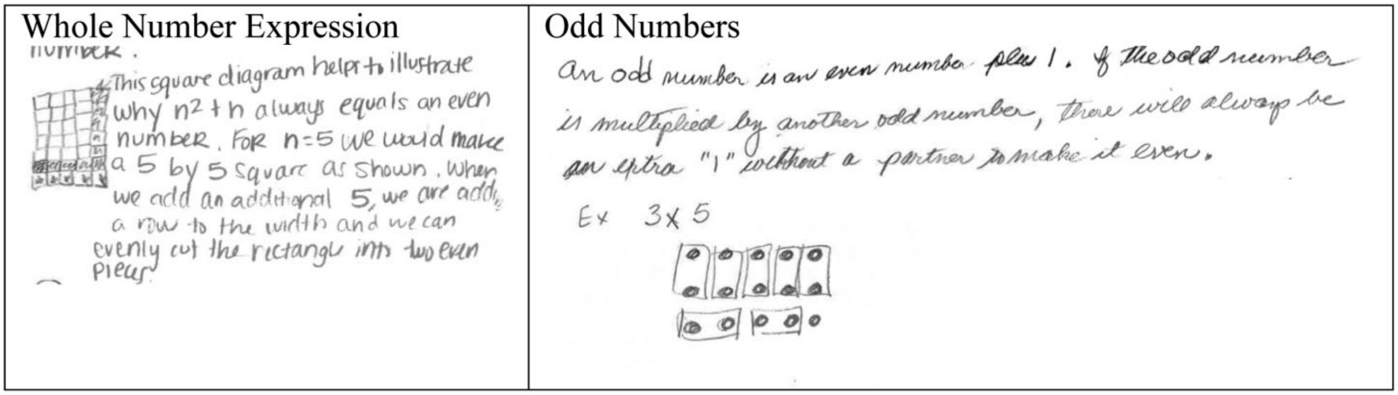
$27\times5=135$

this claim is true. All odd numbers that I tried seem to have an odd product.

Figure 4. Naïve and crucial empiricism examples.

statement (Figure 5); however, while the singular generic examples were illustrative of more general principles at work in the problem, in each case, the participant did not explicitly or completely reason about the more general argument. In this way, the generic examples all had some limitations for a more general argument. For example, the 3 × 5 example below (Figure 5—odd numbers) demonstrates that there is an additional unpaired dot, thus an odd product, but does not generalize why this would have to be the case for the product of every pair of odd numbers; the argument made using the $n = 5$ square is similarly limited to odd numbers, where using an even number would require a vertical (not horizontal) "cut" to produce two even pieces.

Partial analytical (P). The three categories of deductive reasoning have one primary commonality: participants' responses made assertions based on general statements, as opposed to specific examples. However, because of the familiarity of the content, some participants made general claims without providing sufficient evidence; they seemed to take for granted certain facts and treat them as axioms. Responses such as these were coded as partial analytical. Figure 6 depicts two such examples: the first participant recognizes that the numerator is the product of consecutive numbers and that the numerator needs to be even to generate a whole number, but claims that "adding one . . . makes your numerator even"—there is no discussion about why multiplying two consecutive numbers must produce an even number (i.e., one of them would have to be even); the second participant recognizes that the product of two odds is the sum of an odd number of odd groups, which is at the heart of the explanation, but



Whole Number Expression

number.

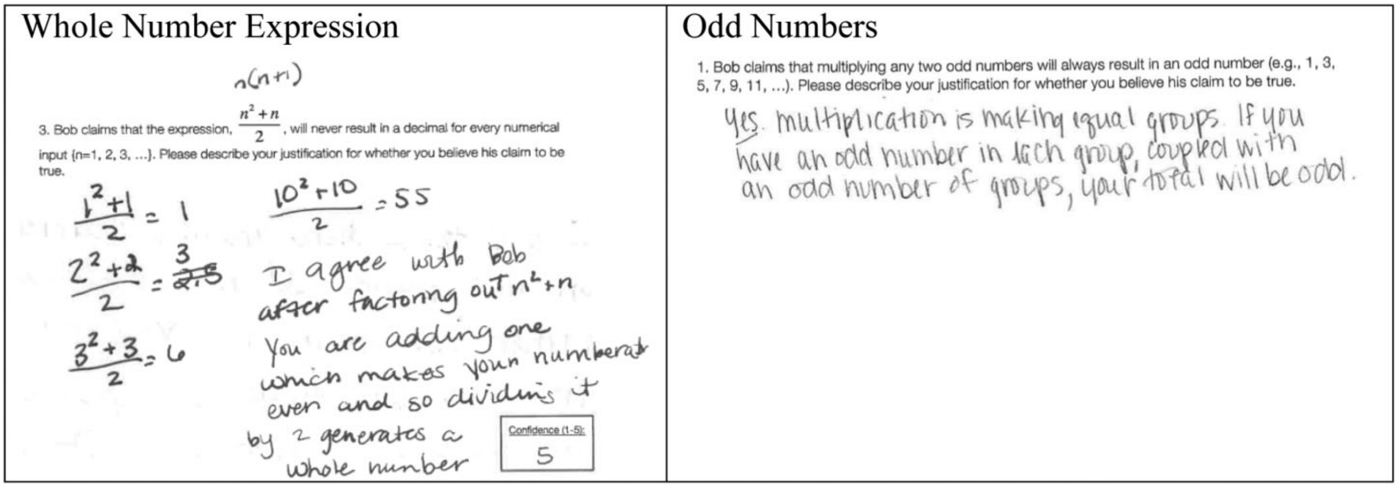
This square diagram helps to illustrate why $n^2 + n$ always equals an even number. For $n=5$ we would make a 5 by 5 square as shown. When we add an additional 5, we are adding a row to the width and we can evenly cut the rectangle into two even pieces.

Odd Numbers

An odd number is an even number plus 1. If the odd number is multiplied by another odd number, there will always be an extra "1" without a partner to make it even.

Ex 3x5

Figure 5. Generic examples.



Whole Number Expression

$n(n+1)$

3. Bob claims that the expression, $\frac{n^2+n}{2}$, will never result in a decimal for every numerical input {n=1, 2, 3, …}. Please describe your justification for whether you believe his claim to be true.

$\frac{1^2+1}{2} = 1$

$\frac{2^2+2}{2} = \frac{3}{\cancel{2.5}}$

$\frac{3^2+3}{2} = 6$

$\frac{10^2+10}{2} = 55$

I agree with Bob after factoring out n^2+n You are adding one which makes your numerat even and so dividing it by 2 generates a whole number

Confidence (1-5): 5

Odd Numbers

1. Bob claims that multiplying any two odd numbers will always result in an odd number (e.g., 1, 3, 5, 7, 9, 11, …). Please describe your justification for whether you believe his claim to be true.

Yes. multiplication is making equal groups. If you have an odd number in each group, coupled with an odd number of groups, your total will be odd.

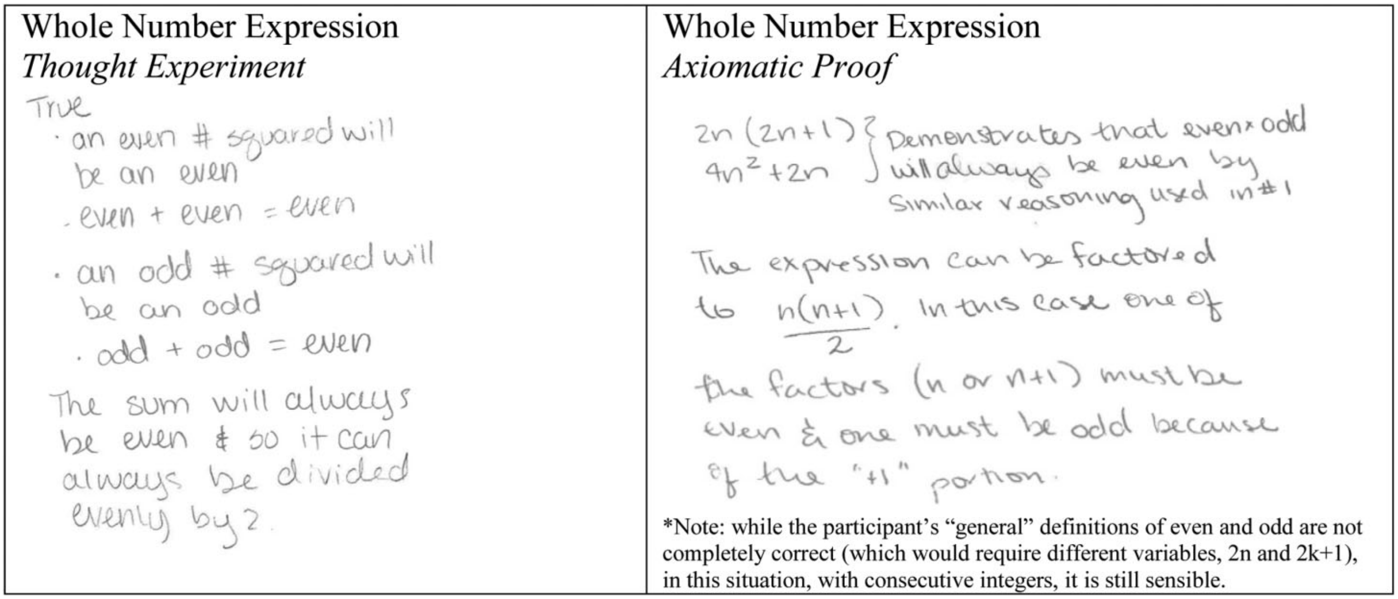
Figure 6. Partial analytical examples.

does not explain why the sum of an odd number of odd groups cannot be even.

Thought experiment (T) and axiomatic proof (A). In contrast, those labeled as either thought experiment or axiomatic proof made a more complete set of deductions about the situation, forming a more rigorous argument. The difference between these two was the degree of formality and reliance on precise definitions and statements. (There were very few examples of axiomatic proof; one was a correct counterexample to the prime number generator, $n = 41$). We also note that while most responses coded as thought experiments were correct, there were four responses to the prime number generator problem that made assertions based on incorrect generalizations (e.g., because the quadratic expression cannot be factored, the result of any input must be prime). Their reasoning was still coded as a thought experiment despite the inaccuracies; however, the relatively few such examples do not impact the overall findings. Figure 7 depicts an example of each coding category.

Analysis

After coding all of the responses ($n = 141$), we removed those coded as flawed ($n = 18$) because the participants' incorrect interpretations unduly altered their reasoning, justification, and reported confidence levels. For example, participants who identified a counterexample to the conjecture, but through a mathematical mistake (e.g., evaluating $n^2 - 4 + 41$, see Figure 3), reported uncharacteristically high degrees of confidence despite the inaccuracy; additionally, because such reasoning was about a fundamentally "different" problem than what the other participants considered, these few cases were removed from further analysis. This left 123 responses across all of the three tasks, 68 from mathematics teachers, and 55 from science teachers. The research questions were addressed primarily through quantitative analysis based on the proof coding, with the two populations being comparable given the relatively strong mathematical background required of both secondary mathematics and science teachers, as well as the similar-



Whole Number Expression
Thought Experiment

True
· an even # squared will be an even
. even + even = even
· an odd # squared will be an odd
· odd + odd = even
The sum will always be even & so it can always be divided evenly by 2.

Whole Number Expression
Axiomatic Proof

$2n(2n+1)$ } Demonstrates that even×odd
$4n^2+2n$ } will always be even by similar reasoning used in #1

The expression can be factored to $\frac{n(n+1)}{2}$. In this case one of the factors (n or n+1) must be even & one must be odd because of the "+1" portion.

*Note: while the participant's "general" definitions of even and odd are not completely correct (which would require different variables, 2n and 2k+1), in this situation, with consecutive integers, it is still sensible.

Figure 7. Thought experiment and axiomatic proof examples.

ity of the program entry requirements and their background teaching demographics.

To answer the first research question about whether there were differences between the mathematics and science teachers' responses, in particular their reliance on empirical evidence, we looked first at a *t*-test comparison of the means of their proof scores, which would be approximately normally distributed for both populations. However, to make sure that some of the other assumptions were reasonable, we looked at a nonparametric Mann–Whitney *U*-test to make sure the 0–4 ordinal proof scale was not that far from an interval scale. We also used a multilevel regression model, grouping by individual teachers, to verify that the individual data points were independent enough, despite three coming from each teacher. Lastly, we used a *t*-test to compare the proportion percentage of correct deductive responses (scores of 4) between the two groups, and the proportion percentage of inductive responses (scores of 1 or 2) between the two groups.

To answer the second research question about the degrees of confidence that mathematics and science teachers lend to their use of empirical evidence versus deductive arguments, we looked at linear regression models associating proof score with reported confidence level. In particular, for those only able to provide empirical evidence, not deductive proof, the self-reported confidence score gave an indication about how mathematically convincing they found such reasoning to be. First, however, there were eight conjectures in which the participant forgot to provide a "confidence" rating. These cases were removed from this second analysis, leaving 115 responses, 66 mathematics teachers, and 49 science teachers. (For mathematics teachers, this removed two responses, given ratings of 2 and 3 on their proof scale; for science teachers, this removed six responses, given ratings of 0, 1, 1, 1, 2, and 2). After this, we looked at whether the slope coefficients of the linear regression models were significantly different from 0 because proof scores and levels of confidence should be positively associated (see Figure 1). In addition, isolating just the completely inductive responses (scores of 1 or 2), we computed whether there was a different distribution of numerical confidence values across the two populations.

Results

After coding each of the participants' responses, the overall distribution of codes and scores for the two groups of teachers' reasoning across all three tasks is listed in Table 3.

To answer the first research question, the *t*-test comparison of mean proof scores across all three tasks resulted in a statistically significant difference between the mathematics and science teachers (p = .01), with an approximate medium effect size (d = .47). While the hierarchy of proof schemes substantiates the value of the scores being an ordinal variable, it is unclear whether the scale is also interval. To make sure that the assumptions of the *t*-test, both in being normally distributed and having an interval variable, were not unreasonable and did not provide erroneous significance results, we verified the difference between the two groups' distribution of responses using

Table 3
Distribution of Codes and Scores

	Mathematics Teachers (n = 68)							Science Teachers (n = 55)						
Code	E	N	C	G	P	T	A	E	N	C	G	P	T	A
Quantity	0	17	21	3	9	16	2	3	19	19	1	6	6	1
Score	0	1	2	3		4		0	1	2	3		4	
Quantity	0	17	21	12		18		3	19	19	7		7	
	Mean score: 2.46							Mean score: 1.93						

the nonparametric Mann–Whitney U-test. This test also resulted in a statistically significant difference between the two groups' distribution of scores (p = .014). In addition, to make sure that using each teacher's responses across all three items as three individual data points were also reasonable, we ran a multilevel regression model, grouped by teacher, to verify that the responses were independent enough. The result of the multilevel regression model demonstrated a statistically significant fixed size effect on scores between the two groups (p = .045). From the multilevel regression model, we see the increase in p-value as an indication that not all the data points should be considered completely independent, but not so much so that using all the data points skewed the significance of the comparison between the two groups. These results indicate that there were statistically significant differences between the reasoning schemes that mathematics and science teachers used to validate the mathematical conjectures.

Across the entire distribution of scored responses, there is a difference between the two groups. However, looking also at two specific categories of interest, we compared the proportion percentage of correct deductive proof schemes (scores of 4, less the four incorrect thought experiment responses) and the proportion percentage of completely empirically based proof schemes (scores of 1 or 2). For the mathematics teachers, 15 of the 68 responses (22%) were correct deductive proof schemes compared with 6 of the 55 science teacher responses (11%). A t-test comparison of proportions resulted in the probability of there being a difference between the two groups of p = .051, which, although not statistically significant (p < .05), indicates a potentially reasonable distinction between these two proportions. As for the completely empirical proof schemes, 38 of 68 responses (56%) for the mathematics teachers' and 38 of 55 (69%) of the science teachers' responses were coded as such. The proportions t-test comparison resulted in the probability of there being a difference between the two groups' responses of p = .067. Similarly, although the result is not statistically significant (p < .05), there seems to be some potential for the two groups using only empirically based arguments more and less frequently.

To answer the second research question, each of the 115 responses that included a confidence score were used to determine linear regression models for the mathematics teachers (n = 66) and science teachers (n = 49). In particular, based on the research framework, higher levels of proof should correspond with higher confidence scores, and lower levels of proof should correspond with lower confidence scores. Therefore, we fit linear regression models for both groups, particularly interested in the values of the slopes. Figure 8 depicts the data for each group and the linear regression lines.

The two linear regression lines, for the mathematics teachers of $\hat{y} = 0.317\hat{x} + 3.18$ and for the science teachers of $\hat{y} = 0.045\hat{x} + 4.26$, both indicate a positive slope, which is desirable from the research framework. The r values for the linear models indicate their relative effect size, with the model for the mathematics teachers having a medium effect size (r = .30) and for the science teachers having a very small effect size (r = .05). As for the numerical values of the two slopes, the mathematics teachers not only had a higher slope value, but upon testing the significance of the slope (against the null hypothesis that the slope is equal to 0) the results indicated a statistically significant slope for the mathematics teachers (p = .015) but not for the science teachers (p = .71). Thus, there seems to be evidence that higher proof scores indeed were correlated with higher confidence scores and lower proof scores with lower confidence for the mathematics teachers, but not necessarily for the science teachers. In fact, the confidence scores for the science teachers were relatively constant across all hierarchical levels of proof, indicating very little difference between their confidence in inductive and in deductive reasoning schemes for verifying mathematical conjectures.

Interestingly, analyzing only the one conjecture that was untrue—the prime number generator task—the linear regression models for the two groups both indicate a slightly negative slope value (for mathematics teachers of $m = -.0316$, and for science teachers of $m = -.167$), neither of which were statistically different from a slope of zero. Some of this may be attributed to the comparably few data points that received a high proof score on this task for both groups; however, this task also was likely the most unfamiliar to both groups of students and difficult because only at the 41st value does the conjecture become clearly untrue.

Across the entire distribution of proof scores and confidence values, there seems to be an indication that there was a difference in the degree of confidence the two

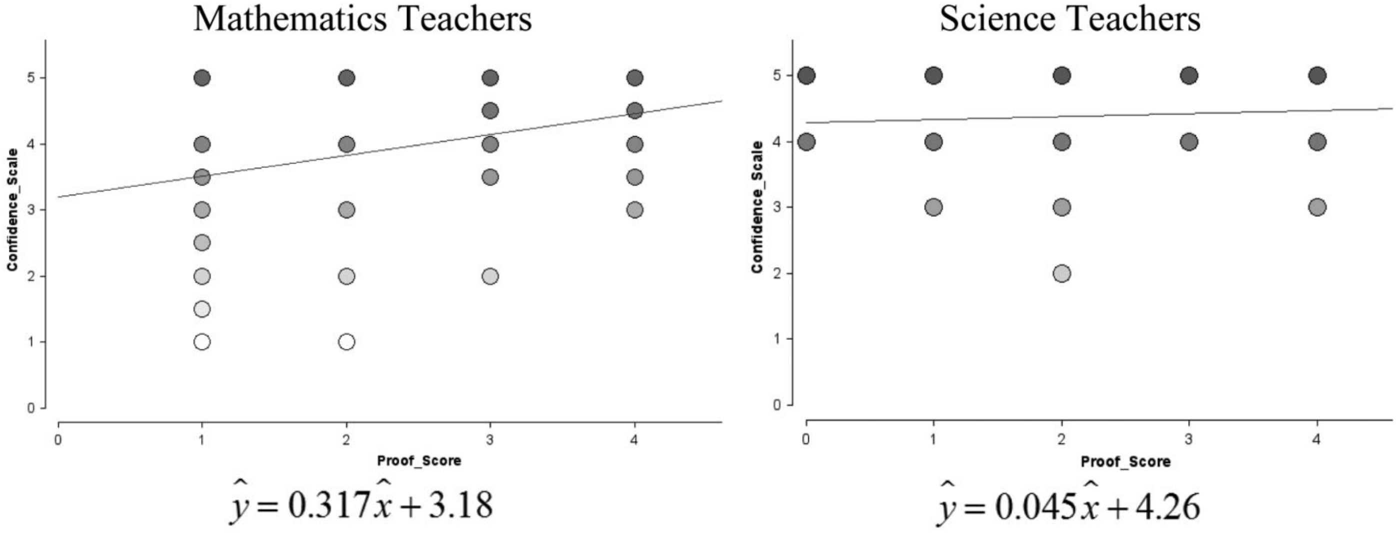


Figure 8. Linear regression lines and data for both groups of teachers.

groups placed across the hierarchy of proof and reasoning schemes. In addition, looking just at the confidence placed in completely empirical arguments (scores of 1 or 2), we analyzed the distribution of confidence scores for only those data. Of the 38 scores of 1 or 2 from the mathematics teachers, 37 also had confidence scores; of the 38 from the science teachers, 33 had confidence scores. Because the confidence scores were not necessarily normally distributed, likely skewed left because of there being a maximum confidence score, the nonparametric Mann–Whitney *U*-test was considered, as there seemed to be no overarching reason to question that one group was more or less likely to provide higher or lower confidence scores (symmetry). The result of the test was not statistically significant (p = .058), which, again, although not statistically significant indicates some potential for difference between the two groups' confidence in completely empirical arguments.

Discussion

Evident from this study is that the mathematics and science teachers employed somewhat different forms of reasoning to justify the truth or falsehood of the three mathematical conjectures. Although the numerical differences are relatively small, they categorically point toward a shift in reasoning from more inductive approaches to more conceptual and deductive ones. In addition, the two groups expressed varying degrees of confidence across the hierarchy of mathematical reasoning and proof schemes. While teacher education in individual STEM disciplines has been studied more extensively, these findings have implications for preparing teachers for more integrated STEM approaches, for which less is known.

In particular, the science teachers in this study were more prone to use inductive methods of reasoning to validate mathematical ideas. They also seemed to express equal confidence across both inductive and deductive modes of reasoning in their responses. Likely, the epistemological approach in science, and their background in this discipline, influenced their reasoning on the mathematical conjectures by establishing the validity of a statement through empirical testing; additional qualitative studies may help provide further support of and insight into this outcome. While this result and the expressed confidence in this type of reasoning is not unreasonable, the science teachers did not seem to demonstrate awareness that verification in mathematics requires a substantively different approach. Although using repeatability as a means for establishing the truth of a statement can be a useful pedagogical approach, even for teaching mathematics, this inductive type of reasoning has certain limitations in terms of mathematical integrity. In particular, to make sure STEM experiences are "content-accurate" and "focused on foundational ideas" (Dayton Regional STEM Center, 2013), the use of examples should move a classroom mathematics discussion toward more complete explanations and arguments based on the properties of the situation, not simply repeated empirical validation.

The responses from the mathematics teachers, on the other hand, indicated being slightly more attuned to deductive reasoning and proof schemes, as well as the limitations of inductive reasoning. However, their responses are far from reassuring. A large majority of the teachers used less rigorous means (i.e., inductive reasoning) to establish the truth of the conjectures. And while they simultaneously indicated a slightly lower degree of confidence in this reasoning, this finding also suggests that these practicing mathematics teachers were unable to determine meaningful reasons, based on the mathematical properties of the conjecture at hand, for why certain phenomenon occurred. Also, while the slope of the regression line for the mathematics teachers was positive and statistically different from zero, which indicates slightly more desirable confidence levels across the different levels of proof, in the ideal model—where teachers would have zero confidence in external arguments, etc.—the slope value of a linear model would have been closer to $m = 1.25$ (rather than $m = .37$). This suggests that while the mathematics teachers had more understanding of the limitations of empirical reasoning in mathematics than the science teachers, they still had a relatively high degree of confidence in this form of justification.

Considering STEM education, it would be prudent for schools and colleges to focus on meaningful enhancement and integration. This study has found that despite some of the connections between the STEM disciplines and the mathematical applications present in them, the overall understanding and use of deductive reasoning as the preferred form of mathematical validation is not always grasped or employed. In particular, the epistemological foundations in science that more often rely on inductive arguments for validating claims may promote similar strategies for reasoning about mathematical conjectures. This result poses some potential tension in terms of teachers' preparation and knowledge for integrating the teaching of STEM fields, particularly in the realm of mathematical integrity; understanding the core ideas and the nature of a discipline—particularly how it may differ from others—is part of deep disciplinary knowledge that is important for teaching.

Despite the wide variety of interpretations and implementations of STEM education (e.g., Breiner et al., 2012), there seems to be some coalescence suggesting integration as a meaningful component. The results of this study find that if integration between the disciplines is a key premise for STEM education, as Johnson (2012) suggests, then preparation and training that attend to both the disciplinary similarities and the disciplinary differences, particularly epistemological ones, must be a key principle for preparing STEM teachers. Part of developing teachers content knowledge for teaching (e.g., Shulman, 1986) compels STEM teacher education to be attuned to broad disciplinary distinctions, particularly in the realm of valid disciplinary approaches to reasoning and justification. The knowledge for meaningfully integrated STEM curricula, which simultaneously maintains disciplinary integrity, poses relatively high demands on teachers. This facet of conceptualizing STEM as an integrated educational domain has real implications for preparing teachers. Teachers need to know each discipline deeply and be able to reason appropriately—and sometimes differently—depending on the subject at hand; yet they should also be knowledgeable about meaningful interdisciplinary connections. And although students, too, are responsible for learning and making meaning out of the integrative process, the challenge of integrating STEM education initially will rest on teachers, for which the findings of this study add to the dialogue about STEM education and indicate a need for additional attention to their disciplinary and interdisciplinary training.

This study contributes to the conversation about STEM education by documenting some of the differences in how mathematics and science teachers approach problems in one of the disciplines. Particularly in this study, researchers were looking at how mathematics and science teachers approach reasoning about mathematical problems. Given the mathematical connection across the STEM fields, as well as the use of empirical reasoning in science and its status in mathematics, this study was a sensible first step; it documents a significant difference in the approaches that science teachers took and the confidence that they had when looking at mathematical conjectures (compared with mathematics teachers). While the results do not inform the best ways to approach STEM education, it does point to some potential challenges in preparing teachers that need to be considered for teaching the STEM disciplines as an integrated educational domain. In particular, for both mathematics and science teachers, in understanding the different philosophical approaches to reasoning, verification, and proof within the disciplines, and the challenge to simultaneously understand, differentiate, and integrate these in their teaching. Yet the findings from this study also contain limitations. While the three different tasks increased the overall quantity of responses, having completely independent data points, and thus a larger sample of teachers, could improve the findings. However, for this study, the grouping by teachers did not impact the significance level. And while it is plausible that one group or the other had stronger disciplinary (mathematics or science) training, or more inherent reasoning ability, we find such a difference at the group level unlikely given that the participants were primarily selected from graduate programs in mathematics and science education having similar program requirements. In addition, there were some differences in gender and grade level assignment present between the two groups.

Yet, because high school teachers often have had even more disciplinary teaching, we anticipate that having had more high school mathematics teachers to be more on par with the number of high school science teachers in the study would only have further differentiated the groups. Lastly, further corresponding studies, such as those incorporating responses to scientific conjectures, would be useful to help further clarify understanding of how differing backgrounds in STEM disciplines influence approaches to discipline-specific questions.

Conclusion

The trend for incorporating and enhancing STEM education has increased over the past decade. Precisely what is intended by STEM education is still unclear, although there seems to be a more recent shift toward meaningful integration. This study informs differences for how mathematics and science teachers, two of the STEM disciplines, approached reasoning about mathematical conjectures. While the strong use of mathematics in the sciences—and across the STEM disciplines for that matter—could suggest that both mathematics and science teachers have similar understandings about reasoning in mathematics, evident from this study was a difference between the two groups regarding their levels of reasoning and confidence. This difference is potentially informed by epistemological distinctions between the two disciplines. Regarding STEM education, this study suggests that integrating the teaching of STEM disciplines will require particular attention to the preparation and training of STEM teachers as a key premise for navigating the tension between meaningful integration and disciplinary integrity. In this study, this was especially the case for mathematics and the use and status of deductive modes of reasoning, relying less frequently on and being less confident in inductive arguments. While formal deductive reasoning is not necessarily the end goal of learning mathematics, this mode of reasoning is still fundamentally important (indeed, to both mathematics and the sciences) and cannot be left out of integrated STEM education endeavors. Understanding how epistemological distinctions between mathematics and other STEM disciplines, such as science, inform valid modes of reasoning and justification will be important for meaningfully integrated STEM education, and for informing teachers' disciplinary and interdisciplinary preparation.

References

Balacheff, N. (1988). Aspects of proof in pupils' practice of school mathematics. In D. Pimm's (Ed.), *Mathematics, teachers and children* (pp. 216–235). London: Hodder and Stoughton.

Breiner, J. M., Harkness, S. S., Johnson, C. C., & Koehler, C. M. (2012). What is STEM? A discussion about conceptions of STEM in education and partnerships. *School Science and Mathematics*, *112*(1), 3–11.

California Department of Education. (2013). *Science, technology, engineering, and mathematics*. Retrieved from http://www.cde.ca.gov/pd/ca/sc/stemintrod.asp

California STEM Learning Network. (2012). *What is STEM?* Retrieved from http://cslnet.org/what-is-stem/

Chazan, D. (1993). High school geometry students' justification for their views of empirical evidence and mathematical proof. *Educational Studies in Mathematics*, *24*(4), 359–387.

Dayton Regional STEM Center. (2013). *STEM education quality framework*. Retrieved from http://www.washingtonstem.org/STEM/media/Media/Resources/STEM-Ed-Quality-Framework.pdf?ext=.pdf

Devlin, K. (2003). *Mathematics: The science of patterns*. New York: Henry Holt and Co.

Gonzales, P., Guzmán, J. C., Partelow, L., Pahlke, E., Jocelyn, L., Kastberg, D., & Williams, T. (2004). *Highlights from the Trends in International Mathematics and Science Study (TIMSS) 2003* (NCES 2005–005). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Harel, G., & Sowder, L. (1998). Students' proof schemes: Results from exploratory studies. In A. Schoenfeld, J. Kaput, & E. Dubinsky (Eds.), *Research in collegiate mathematics education III* (pp. 234–283). Providence, RI: American Mathematical Society.

Harry, B., Sturges, K., & Klingner, J. K. (2005). Mapping the process: An exemplar of process and challenge in grounded theory analysis. *Educational Researcher*, *34*(2), 3–13.

Horsten, L. (2012). Philosophy of mathematics. In E. N. Zalta's (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2012 Edition). Retrieved from http://plato.stanford.edu/archives/sum2012/entries/philosophy-mathematics/

Johnson, C. C. (2012). Editorial: Four key premises of STEM. *School Science and Mathematics*, *112*(1), 1–2.

Knuth, E. (2002). Secondary school mathematics teachers' conceptions of proof. *Journal for Research in Mathematics Education*, *33*(5), 379–405.

Lemke, M., Sen, A., Pahlke, E., Partelow, L., Miller, D., Williams, T., Kastberg, D., & Jocelyn, L. (2004). *International outcomes of learning in mathematics literacy and problem solving: PISA 2003, results from the U.S. perspective* (NCES 2005–003). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Losee, J. (1972). *A historical introduction to the philosophy of science*. New York: Oxford University Press.

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

National Research Council. (2012). S. R. Singer, N. R. Nielsen, & H. A. Schweingruber (Eds.), *Discipline-based education research: Understanding and improving learning in undergraduate science and engineering*. Washington, D.C.: National Academies Press.

Next Generation Science Standards (NGSS). (2013). *Next generation science standards: For states, by states (Appendix H: Nature of science)*. Retrieved from http://www.nextgenscience.org/next-generation-science-standards

Popper, K. R. (1963). *Conjectures and refutations: The growth of scientific knowledge*. London: Routledge and Keagan Paul.

Rosenberg, A. (2000). *Philosophy of science: A contemporary introduction*. Florence, KY: Routledge.

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*(2), 4–14.

Stylianides, A. J. (2007). Proof and proving in school mathematics. *Journal for Research in Mathematics Education*, *38*(3), 289–321.