RESEARCH IN MATHEMATICS EDUCATION
2021, VOL. 23, NO. 1, 21–38
https://doi.org/10.1080/14794802.2020.1722960

Comparing middle grade students' oral and written arguments

Tye G. Campbell [a], Shande King [b] and Jeremy Zelkowski [a]

[a]University of Alabama, Tuscaloosa, AL, USA; [b]University of Tennessee, Knoxville, TN, USA

ABSTRACT
Previous research making claims about students' proving capabilities usually only explored one modality (e.g. oral or written arguments). This study compares the written and oral representations of eighth-grade students' mathematical arguments to determine whether the two modalities portray similar understandings. We replicate Stylianides' (2019) analysis using a different temporal sequencing of oral and written arguments and a larger sample size. Using an existing proof scoring scheme, we found a statistically significant difference between students' written and oral representations of proof. Students often struggled to write arguments they could otherwise represent orally, and we found no evidence that students' written arguments were more sophisticated than their oral arguments. We also share two illustrative cases representing the disparity between students' written and oral arguments. The findings of this study suggest that students' proving capabilities might not be as dire as previous research implies.

KEYWORDS
Mathematics; proof; oral arguments; written arguments

1. Introduction

Proof is an essential component of learning, doing and communicating mathematics (Schoenfeld, 1994). Not only do scholars in the field find proof to be an integral part of mathematics (Stylianides, Bieda, & Morselli, 2016), but policymakers around the world are increasing the emphasis of reasoning and proof in Kindergarten to grade 12 classrooms (Nardi & Knuth, 2017). For instance, the National Council of Teachers of Mathematics (NCTM, 2000) recommended reasoning and proof as a process standard for students of all ages in the USA, and policymakers recently named creating viable arguments and critiquing the reasoning of others as a mathematical practice standard for school mathematics (CCSSI, 2010). Further, the UK Department for Education (2014) recognised reasoning with an emphasis on justification as a priority for school mathematics. Proof is also a primary topic of research in mathematics education for scholars all over the world (e.g. Cáceres, Nussbaum, Marroquín, Gleisner, & Marquínez, 2018; Komatsu, 2016; Tsamir & Sheffer, 2000) signifying its relevance across many cultures and geographic regions.

CONTACT Shande King nking5vols.utk.edu

Despite the prominence of proof in various curricula, the literature suggests students of all ages struggle to create viable arguments (e.g. Healy & Hoyles, 2000; Lannin, 2005; Lin, Yang, & Chen, 2004; Sen & Guler, 2015). However, most literature only considers one modality (e.g. written or oral) when making claims about students' proving capabilities. Further, large-scale studies (e.g. Healy & Hoyles, 2000), which are widely cited in the literature on proof and argumentation, almost exclusively use students' written arguments to make claims about their proving abilities. Some scholars have raised doubt into whether students' written arguments are analogous to their actual understandings of proof (Evens & Houssart, 2004; Soto-Johnson & Fuller, 2012; Stylianides, 2019). In this study, we seek to further examine differences in students' oral and written representations of proof.

The purpose of this study is to compare middle grade students' collaboratively constructed oral and written arguments. Collaboratively constructed arguments refer to arguments created in collaboration with others. We found it important to analyse students' collaboratively constructed proofs as opposed to other methods for eliciting oral arguments (e.g. interviews) for two reasons. First, in our estimation, students are more likely to express their written and oral understandings of proof in contexts where they do not feel pressure from authority figures and other external sources. Second, allowing students to create collaborative arguments ensures a different temporal sequencing of the written and oral modality than was considered in previous research. Since we chose to analyse collaboratively constructed rather than individual arguments, there are other factors we must consider when making claims about the differences between written and oral modalities. We discuss these factors in detail within the body of this paper.

To serve the purpose for this study, we analysed audio recordings and written artefacts to compare eighth-grade students' oral and written arguments. We used statistical measures, largely replicating Stylianides (2019) analytic procedures, to show the discrepancies in written and oral modalities exhibited amongst each group. Then, we share two case studies illustrating differences in written and oral representations of proof. The research question that guides this study is as follows: How do eighth-grade students' collaboratively constructed oral and written arguments compare?

In the next section, we share literature about students' proof schemes followed by a section on students' proving capabilities. We address limitations in previous literature and identify methodological limitations of previous research. Then, we share the limited literature related to the differences between students' written and oral understandings to frame our study.

2. Proof schemes

Though others have developed taxonomies for understanding students' proof schemes (e.g. Balacheff, 1988), we find Harel and Sowder's (1998) to be the most comprehensive categorisation. Harel and Sowder (1998) developed a holistic view of students' proof schemes by observing the proving practices of mostly college students along with a case study of one junior high school student. In their analysis, they concluded that students' proof schemes fell within three categories: external conviction, empirical or analytical. The external conviction proof scheme occurs when a student relies on authority such as a teacher, parent or textbook to create a mathematical argument. For instance, Harel and Sowder (1998) found that students often believed proofs should contain complicated language or symbols because they saw sophisticated arguments in textbooks or other authoritative domains. They also found that students used authority figures as warrants for claims by using language such as "because my teacher said so". Other scholars similarly found authority to be a proving scheme exhibited by students of all ages (Flores, 2006; Fried & Amit, 2008; Sen & Guler, 2015). The empirical proof scheme refers to using examples, experiences or perceptions to justify a claim (Harel & Sowder, 1998). Students use this proof scheme when they attempt to show that a conjecture is true by verifying a finite number of cases. Lastly, the analytical proof scheme refers to arriving at a conclusion by using a sequence of logical deductions (Harel & Sowder, 1998). Students often prove that a statement is true in mathematics by using axioms, definitions and previously proved theorems to create a logical argument.

Though Harel and Sowder (1998) warned against treating their taxonomy as an hierarchical structure, they acknowledge that "there is often at least a partially hierarchical nature implicit in the categories" (p. 277). In fact, several scholars have used this taxonomy as a hierarchical rubric for determining how students' arguments change over time (e.g. Ellis, 2007; Flores, 2006; Lee, Chen, & Chang, 2014; Liu & Manouchehri, 2013; Stylianou, 2013). In the next section, we explore claims of previous literature on the categorisation and frequency of students' proof schemes.

3. Previous claims on students' proving capabilities

A vast amount of literature is devoted to understanding students' proving knowledge and capabilities (e.g. Flores, 2006; Harel & Sowder, 1998; Healy & Hoyles, 2000; Lin et al., 2004; Liu & Manouchehri, 2013; Martin & Harel, 1989; Weber, 2001). Scholars claim that students often rely on authoritative or empirical proof schemes (Flores, 2006; Healy & Hoyles, 2000; Lin et al., 2004; Martin & Harel, 1989). However, there are a variety of ways to measure students' proving abilities, and students' understanding might not be uniform across each measurement. For instance, Healy and Hoyles (2000) used a survey instrument to determine the proof conceptions of 14–15 year-old learners. In their study, students were given a questionnaire of conjectures with a list of possible arguments for which students were to rate which arguments would be similar to the argument they would create and which arguments would receive the best mark from the teacher. Students also had opportunities to construct written proofs within the questionnaire. They found that students rated empirical arguments highly for arguments they would create for themselves, and they often constructed written empirical arguments. However, students were aware that empirical arguments would not receive high marks from teachers.

Healy & Hoyles' (2000) study is widely cited for confirming deficiencies in students' proving practices along with other seminal pieces (e.g. Martin & Harel, 1989; Weber, 2001). However, it is possible that students' understanding of proof is constrained by the modality by which they construct an argument. For example, referring to Healy and Hoyles' (2000) study, only students' written proofs were considered for analysis along with a multiple choice format for choosing the best argument for a conjecture. Students' written understandings are not always reflective of their cognitive knowledge (Evens & Houssart, 2004), and research suggests students sometimes portray different understandings depending on the methodological instrument (Stylianides & Al-Murani, 2010). While other scholars have explored different modalities for constructing an argument, there has been scant literature comparing the different means by which students construct proofs. We claim that it is imperative to consider the different modalities for representing proving knowledge to better understand students' proving capabilities.

4. Comparing oral and written arguments

In our search, we found only two studies which explicitly compared different modalities by which students create mathematical arguments (Soto-Johnson & Fuller, 2012; Stylianides, 2019). However, recently scholars have noted the differences in students' written and oral socioscientific arguments (Berland & McNeill, 2010; Evagorou & Osborne, 2013; Knight & McNeill, 2015). Within socioscientific contexts, some empirical research suggests that students create more sophisticated written arguments (e.g. Knight & McNeill, 2015) while other research suggests that students' oral arguments are of higher quality (e.g. Berland & McNeill, 2010). In any case, the socioscientific literature together suggests that students' written and oral justifications are different.

Within mathematics argumentation literature, Soto-Johnson and Fuller (2012) compared the oral and written proofs of college students enrolled in an Abstract Algebra course by individually interviewing participants. They found that 10 out of 22 participants' oral proofs were more sophisticated than their written arguments. Some students were unable to make any progress in writing a proof, but they had success in verbally discussing some aspects of the proof. Soto-Johnson and Fuller (2012) argued that previous literature worked from a deficit model, but when considering alternate ways to assess proving knowledge, they suggested the reality of students' proving capacities might not be as dire as others have noted.

In a similar study, Stylianides (2019) found that students aged 14–15 years old often created more sophisticated oral arguments than written arguments. In his analysis, students created written arguments before orally presenting their arguments in front of the class. Stylianides (2019) recognised the temporal sequencing of having students present oral arguments after the creation of written arguments as a limitation of his study. He wrote,

> it is possible that a student's experience of producing a written proof helped the student build familiarity with the task and underlying concepts, thus placing the student in better position later on to orally present an argument that approximated the standard of proof. (p. 168)

Soto-Johnson and Fuller (2012) and Stylianides (2019) made important contributions by sharing examples of the differences in learners' written and oral arguments. While their research provides an existence proof of the phenomenon, there are important limitations of the previous research which we hope to address. Soto-Johnson and Fuller (2012) asserted that 10 out of 22 students created more sophisticated oral arguments compared to written arguments, but they did not use a systematic scheme to analyse their findings. This is not a criticism of their work; rather, the main purpose of their study was to share two case studies revealing the differences in oral and written modalities. Stylianides (2019) examined the general trends of 17 students' written and oral arguments, but a significant limitation in regard to the temporal sequencing of written arguments occurring before oral arguments was noted. In realising this limitation, Stylianides (2019) asserted that exploring a different temporal sequencing would "enhance the field's understanding of the relationship between students' oral and written arguments (perceived proofs) for the same claims, albeit under a different variation of temporal sequencing, with a different student population, and in a substantially different setting than in the study I reported in this article" (p. 176). The last limitation pertaining to both previous research studies (Soto-Johnson & Fuller, 2012; Stylianides, 2019) is that the relationship between written and oral arguments has not been quantified due to limited sample sizes. In the next section, we explain the methodology and how we complement the findings of previous research to advance the knowledge base.

5. Methods

5.1 *Setting and participants*

The study took place during the 2018–2019 academic year at a private school in the South-East United States. The participants were eighth-grade students enrolled in one of four different Pre-Algebra courses taught by the same teacher. The students were described by the classroom teacher as average in regards to their mathematical competencies. The classroom teacher often allowed students to work together in groups, so students were accustomed to communicating about mathematics with peers. There were about 70 students in total across each of the 4 classes, and 47 students agreed to participate with signed consent/assent forms. Students were mostly placed in groups of 3 (2 groups of 4; 15 groups) to work on mathematical proving tasks. The case studies that we share following our statistical analyses detail the oral and written arguments of two groups. Both groups consisted of one girl and two boys (Amber, Josh, and Aaron; Lisa, Tiger, and Nick).[1]

5.2 *Instructional design*

Author 1 and Author 2 co-taught each 55-min class period for 3 days of instruction. Because students had no prior experience with proving, on the first day, we introduced proving by facilitating a short discussion about the meaning of proving and the criteria for an argument to count as proof (e.g. a proof should show a claim is true for all cases). Within this discussion, we allowed students to share their ideas, and we guided this whole-class discussion to generate criteria we would use to validate arguments as a classroom community. Our discussion was largely guided by our conceptualisation of proof which aligns with Stylianides' (2007) definition of proof. Stylianides' (2007) definition requires that a proof use statements, forms of reasoning, and forms of expression accessible to and within the conceptual reach of the classroom community. After facilitating a discussion about proving, each of the 47 students was split into groups to work on one proving task per day (see Figure 1). For the purposes of this study, we analyse the written and oral arguments for Task 1 and Task 2. Due to its complexity, the third task was heavily scaffolded by the teachers (Authors 1 and 2), and, therefore, did not provide reliable data in regards to students' written and oral understandings.

For each of the tasks, students started by working individually to create a written argument for four minutes. Then, they worked for 20 min in groups to collaboratively construct an argument. During this time, they orally presented their ideas with one another

Task 1

The sum of any three consecutive integers is divisible by three. Is this conjecture true? Write an argument for why or why not.

Task 2

Conjecture: *The sum of two odd numbers is even.* Is this conjecture true? Write an argument for why or why not.

Task 3

How many triplet primes of the form p, $p+2$, $p+4$ are there? Write an argument for how you know.

Figure 1. Proving tasks.

followed by writing their collaboratively constructed argument on the task sheet. Finally, with the remaining class time, we selected groups to present their arguments to the class. Because we are interested in the differences between *collaboratively constructed* oral and written arguments, the 20 min students spent collaborating on an oral and written argument is the focal point of this analysis. Our temporal sequencing (orally present an argument → write an argument) follows the reverse pattern of Stylianides (2019) study (write an argument → orally present an argument), aligning with his suggestion for future research to explore a different temporal sequencing.

5.3 *Data and analysis*

We collected three data sources from each of the 3 days of instruction including audio recordings of each groups' oral collaborative argument, written collaborative arguments and field notes. All audio recordings were transcribed in full. Each of the 15 groups created an argument for Task 1 and Task 2 creating a total of 30 paired oral and written arguments. However, due to technical difficulties, three oral arguments were not captured via audio recording. Therefore, we considered 27 paired arguments in the analysis. As we seek to make claims about differences between students' oral and written arguments based on collaborative work, we must address four potential criticisms of this endeavour. The potential criticisms and our responses are listed below.

1. It might be argued that, though one group member presented a sophisticated oral argument, the argument was not transferred in written format on the task sheet because the group chose to use a different approach. We have addressed this criticism by only considering oral arguments which were agreed upon by the group. Therefore, orally presented arguments that were not taken up as the collaborative argument were not considered in this analysis.
2. It is possible that a strategy orally presented by one group member was recorded on the task sheet by another group member. Therefore, the written argument might portray an inability of one individual to synthesise another's ideas rather than a difference in oral and written modalities. In nearly every case, the transcripts and field notes reveal the generator of an idea took over responsibilities for writing the argument

making this criticism a minor limitation. Still, we find it important to report all instances regardless of who wrote the argument to determine differences in written and oral modalities in collaborative arguments.

3. It is possible that low-quality written arguments can be attributed to student laziness or an unwillingness to write their arguments on paper. We do not believe groups were susceptible to this in our study because we continually implored students to write their arguments in enough detail where their peers and teachers could understand their work. Additionally, we do not necessarily consider this a variant limitation since previous research making claims based on written arguments were also susceptible to students' lack of effort.
4. As will be illustrated in one of our case examples, the authors occasionally questioned students while they worked in groups. It could be argued that students' oral arguments might appear of higher quality because the authors probed them to clarify some aspects of their reasoning. We also do not consider this a variant limitation because groups had opportunities to provide additional information in their written argument after they were questioned by the authors. Even if it were the case that this seriously influenced the quality of a group's oral argument, the phenomenon at hand (students understand more than they write on paper) would still be illustrated.

Now that we have addressed potential criticisms, we detail our analytic procedures to answer the research question. The first part of our analysis is largely a replication of Stylianides' (2019) procedures. First, we analysed each group's oral and written collaborative arguments according to a coding system which measures the quality of an argument (Stylianides & Stylianides, 2009). The coding system includes the following types of arguments listed in hierarchical order: non-genuine argument (L1), empirical argument (L2), unsuccessful attempt at a valid general argument (L3), valid general argument but not a proof (L4) and proof (L5). The definition of each code, an exemplification from the data for each code, and our rationale for each exemplification are presented in Table 1. It should be noted that coding arguments is an interpretive activity and not all raters will agree on differences between, for instance, L4 arguments and L5 arguments. This is because scholars hold a variety of conceptualisations related to proving (Campbell, Boyle, & King, 2019), and even mathematicians sometimes do not agree on the validity of an argument (e.g. Weber & Czocher, 2019). For school mathematics, Arbaugh, Smith, Boyle, Stylianides, and Steele (2018) suggested the following criteria for an argument to count as proof: (a) the argument shows a claim is true for all cases, (b) the warrants used in the argument are true and accepted by the classroom community, (c) the argument follows a logical sequence and (d) the mathematics is correct. These criteria informed the delineation we made between L4 and L5 arguments. For instance, arguments that used correct mathematics and showed a claim is true for all cases, but did not meet one of the other criteria, were downgraded to L4 arguments.

Author 1 and Author 2 coded three paired oral and written arguments together, and subsequently coded 19 paired observations separately reaching agreement on 89% of the paired observations. Discrepancies were discussed until agreement was reached, and Author 1 coded the remainder of the paired observations. The interrater reliability indicated a high level of consistency amongst the raters. By providing exemplifications from the data, we increase the transparency of our use of the coding scheme. While the reader might operate from differing perspectives from our own, we invite the reader to

Table 1. Coding scheme and rationale for coding.

Definition	Example Oral Argument from Data	Rationale for Coding
Non-genuine argument (L1)—A non-attempt, or an argument that portrays little effort by the learner	No examples from data	
Empirical argument (L2)—An argument based on a finite number of examples.	*Task 1—After group agrees on a strategy, Author 1 questions the group's argument* **Lauren:** Um, so if I take the numbers like one, two, three or four, five, six or seven, eight, nine—you would add like one plus two and that would be three plus three which would be six divided by three is two. But, when you're trying—when you think something is correct, and you don't know it's correct, you have to have at least three examples of why it is. Because if you have two that could be right, but there could also be a wrong example. **Author 1:** So, you're saying three examples are good enough to show that it's true? **Lauren:** Mhm. Or more, but like if there's two then it's not enough, um to back it up.	This oral argument was coded as empirical because the group utilized three examples as warrants for the claim. Though Lauren acted as the representative, the group agreed on this strategy prior to Author 1's questioning.
Unsuccessful attempt at a valid general argument (L3)—An attempt to create an argument that applies to all cases within the domain that is invalid.	*Task 1—Brittany shares an idea, and the group agrees with her strategy* **Brittany:** Wait, I have an idea. We just need to figure out, you know, just add 1, 2, 3, and then 2, 3, 4, and keep doing that up until 9 because it's all going to be divisible by 3. **Mary:** Yeah! **Felicia:** Yeah! That's really smart. **Brittany:** And all the numbers end in 0 through 9. **Felicia:** Yeah, that's good.	This argument was coded as an unsuccessful attempt at a valid general argument because Brittany attempted to create an argument that showed the claim was true for all cases. First, she argued for checking all the single-digit cases, and then she argued all numbers end in 0-9, making it unnecessary to check other cases. This, of course, is an invalid approach.
Valid general argument but not a proof (L4)—A general argument that uses valid modes of reasoning which contains some warrants which are not accessible or made clear to the classroom community.	*Task 2—Group agrees on Carl and Ashley's argument presented below* **Ashley:** So, if you write like—so, let's take five and five because that's [odd numbers]—[Carl: Yeah]. If you take it and take away 1 to make it an even number, add that together and that's 8, and you take those two that you took away and make them [add them] … and then you get a number that equals ten. Ten is an even number. **Carl:** We'll do … it has to be odd. Seven and seven.	This argument was coded as a valid general argument but not a proof because the group shared a generic example. That is, they showed the claim was true for all cases by examining the structure of two specific examples ($5+5$ *and* $7+5$). They detailed a process wherein one could subtract one from the two odd numbers to create two even numbers. Then, they showed the sum of two even numbers was even and adding the two that were taken away originally would result in an even sum. Importantly, the group realized their rationale was true for every case. We did not count this argument as a proof

(*Continued*)

Table 1. Continued.

Definition	Example Oral Argument from Data	Rationale for Coding
	Ashley: Yeah, but they don't have to match though either, right? **Carl:** I know. No, they don't have to. **Ashley:** Seven and five. **Carl:** One minus one would be six and four. Ten plus two is twelve. **Ashley:** And that works with any number.	because they did not explicitly state some warrants. For instance, they did not explicitly state the warrant that two even numbers sum to an even number. Essentially, their argument was a valid rationale, but not a proof.
Proof (L5)—A valid general argument that uses modes of reasoning accessible to the class.	See section 6.2.1 (Amber, Josh, and Aaron's argument)	This oral argument was coded as a proof because it: (a) showed the claim was true for all cases, (b) used explicit warrants that were true and accepted by the classroom community, (c) followed a logical sequence, (d) and utilized correct mathematics.

Note: X represents missing audio data.

consider the phenomenon of interest in the case examples presented in the findings (i.e. do written arguments portray different understandings than oral arguments?).

After coding each oral and written argument, we explored general trends in the data, and sought to determine if there was a statistically significant difference between the oral and written scored arguments (H_0: *There is no significant difference between oral and written argument scores*). We utilised an exact sign test which is a non-parametric test that measures consistent differences between two pairs of observations. Our selection of the exact sign test is predicated on the assumptions of both non-parametric tests and the number of ties in the paired observations (Gibbons, 1993). The three assumptions that should be met to use an exact sign test are as follows: (1) the outcome variable should be continuous or ordinal, (2) the independent variable should consist of matched pairs and (3) the paired observations should be independent. All three assumptions were met making the test an appropriate model. This test was a more robust choice than the Wilcoxon-signed rank test due to the number of ties in the paired data.

Following the quantitative analysis, we used a revised version of an analytic model for understanding students' reasoning using video data (Powell, Francisco, & Maher, 2003) to share a case study of two groups which exhibited differences in the quality of their written and oral arguments. Powell et al. (2003) introduced seven non-linear phases for studying students' reasoning skills using video data: viewing the video data, describing the video data, identifying critical events, transcribing, coding, constructing a storyline and composing a narrative. Because our data consists of audio recordings, we replaced the first two phases with *listening to the audio data* and *describing the audio data*. Based on critical events, we coded the data for specific moments wherein students made a claim, argued the claim or critiqued their group members. We then constructed a storyline to provide narratives of students' oral arguments. Our field notes supplemented the audio recordings by highlighting key moments and physical attributes of the class (Emerson, Fretz, & Shaw, 2011). The two cases were chosen for different purposes. The first case strongly reveals the phenomenon that groups' oral arguments are of higher quality than their written arguments. The second case, while also revealing the phenomenon, was chosen to give insight into our coding procedures. Author 1 and Author 2 did not initially come to agreement on the coding of the second case, making it a valuable case to make our coding scheme transparent to the reader. In the next section, we share the details of our quantitative analysis before moving on to report on the two case studies.

6. Findings

6.1 *Quantitative measures*

Utilising a similar visual representation as Stylianides (2019), the aggregate data from Task 1 and Task 2 are summarised in Figure 2. The horizontal axis represents the scoring of the argument for the written modality while the vertical axis represents the scoring of the oral modality. The numbered boxes represent the number of collaborative arguments represented at each data point. As can be seen, there are no data points beneath the diagonal indicating no written arguments were scored higher than the corresponding oral argument. One interesting trend is that five groups wrote empirical arguments on their task sheet when their oral argument was indicative of an unsuccessful attempt at a valid general argument. This might suggest that students write empirical arguments when they are unsure how to proceed in creating a valid general argument. It was also interesting that four other groups' written arguments were indicative of an unsuccessful attempt at a valid general argument while their oral arguments were scored as a proof. This suggest even when students have the knowledge to successfully prove a statement, they struggle to write it on paper. Indeed, only one group created a written argument that was classified as a proof.

To statistically compare differences in groups' oral and written arguments, we ran separate exact sign tests for Task 1 and Task 2. We made this decision to ensure no

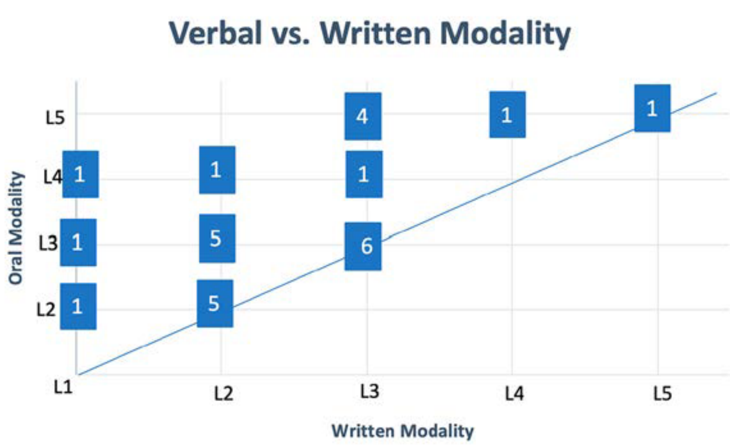

Figure 2. General trends.

Table 2. Differences in level of argument by modality.

Tasks	Groups														
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	0
Task 1 (Oral)	3	2	2	5	5	3	3	3	2	3	3	4	3	3	X
Task 1 (Written)	2	1	2	3	3	1	3	2	2	3	2	2	2	3	X
Difference	1	1	0	2	2	2	0	1	0	0	1	2	1	0	
Task 2 (Oral)	X	2	X	3	5	4	3	3	2	5	5	5	2	4	3
Task 2 (Written)	X	2	X	3	5	1	2	3	2	3	4	3	2	3	3
Difference		0		0	0	3	1	0	0	2	1	2	0	2	0

Note: X represents missing audio data.

dependency amongst the paired samples between the tasks. Both exact sign tests revealed a significant difference at the $\alpha = .05$ level (Task 1: $p = .004$; Task 2: $p = .031$), indicating oral arguments were scored significantly higher than written arguments. The differences between each group's oral and written arguments for both tasks are presented in Table 2.

6.2 *Case studies*

In this section, we first share the case group's written argument and how it was coded according to our scheme. Then, we share the group's interaction and how we coded their oral argument.

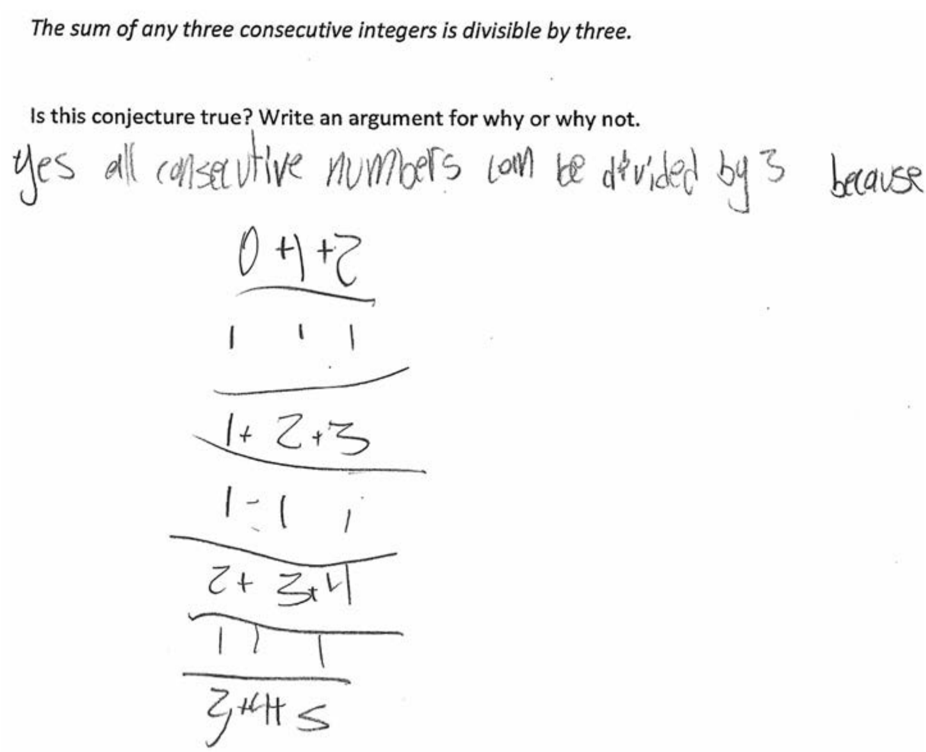

Figure 3. Amber, Josh and Aaron task 1 written argument.

6.2.1 *Amber, Josh and Aaron*

Amber, Josh and Aaron's collaborative written argument on Task 1 is presented in Figure 3. We conservatively coded their written argument as an unsuccessful attempt at a valid general argument, though other raters might conclude the argument is empirical. They stated, "all numbers can be divided by 3", and they provided a list of examples starting with "0+1+2" with what appears to be the numeral "1" written below each example. They may have recognised a pattern in three consecutive integers, but their written argument does not explicitly explain their ideas. Therefore, their written argument portrays weak capacities related to proving this conjecture. Their oral argument, however, portrays a much more sophisticated understanding.

In the 4 min, students were afforded to work on an individual argument, Josh constructed an idea that he immediately shared with his group when the collaboration period started (utterance 1).

1. Josh: So, you start with 0,1,2 as your base one. And then, what you do to bring it up to 1,2,3, is you add one to each of the integers. So, when you're adding one to each of the integers, you're really just adding 3. So, that's why it's divisible by 3. And then, that's why it's divisible by 3 every single time.

Josh's oral argument aligns with a proof by mathematical induction. He explains that the base case (0+1+2) is divisible by 3. Then, he explains adding one to each integer produces the next case (e.g. 1+2+3). He explains why this new case is divisible by three by stating "you're really just adding 3" appealing to the property that adding 3 to a number already divisible by 3 creates a new number that is also divisible by 3. The group asked for further clarification to which Josh further explained his reasoning (utterances 2–5).

1. Aaron: But, I do feel like in some instance it won't be right.
2. Josh: No, it will be right every time because you're just adding 3 every time.
3. Aaron: But, why? We need facts and evidence.
4. Josh: Because if you have three consecutive numbers, if you go up one for each of them, you just add 3. And if the base one does – is divisible by 3 – that means all the other ones will be divisible by 3. And in this instance the base one is divisible by 3.

After Josh's clarification, the group accepted his idea, and Josh became the scribe for the group. Later, Josh realised that he forgot to account for negative cases (e.g. (−3)+(−2)+(−1)). However, after realising this oversight, he stated "you could go into the negatives if you want" implying that one could subtract 1 from each integer in the base case for the other direction. Therefore, Josh constructed two inductive arguments from the base case (0+1+2), one accounting for positive integers and another accounting for negative integers. While Josh did not elaborate on the inductive process for negative integers, his utterance portrayed the understanding that induction would proceed similarly in the other direction.

Amber, Josh and Aaron's oral argument was clearly of higher quality than an unsuccessful attempt at a valid general argument. We coded their oral argument as a proof because Josh used valid modes of reasoning that were accessible to his

The sum of any three consecutive integers is divisible by three.

Is this conjecture true? Write an argument for why or why not.

Yes, any number three consecutive integers can be divisible by three because the integers add up and can be divided.

If you find the sum of the consecutive integers and divide it by three it will always be divisible because of the consecutive integers

Figure 4. Lisa, Tiger and Nick task 1 written argument.

classroom community. They created a proof by mathematical induction which is a sophisticated strategy especially for middle grade students. Later, in their work on Task 3, we learned that the group found it particularly challenging to write a proof as opposed to verbalising it. Amber stated "It's just hard to write an argument. Like, I know what I'm thinking, but I don't know how to word it". Lisa, Tiger and Nick had similar difficulties in Task 2.

6.2.2 *Lisa, Tiger and Nick*

Lisa, Tiger and Nick's collaborative written argument for Task 1 is presented in Figure 4. The group did not attempt to use any warrants in their argument. Instead, they essentially restated the claim, writing "If you find the sum of the consecutive integers and divide it by 3, it will always be divisible because of the consecutive integers". We (Author 1 and Author 2) agreed that the group's written argument should be coded as non-genuine. We came to this conclusion by interpreting their written statement as lacking effort and an elimination process of the other possibilities. For instance, this argument could not be coded as empirical because it does not use examples as warrants for the claim. Additionally, it could not be coded as an unsuccessful attempt at a valid general argument because it does not use any general warrants. Therefore, the written argument must be considered a non-genuine argument. We, however, did not initially agree on the classification of the group's oral argument.

Lisa, Tiger and Nick began their collaboration by empirically testing the truth of the claim. They eventually agreed that the claim was true based on empirical warrants (utterances 6–12).

1. Lisa: If you add them by 3, yeah it can.
2. Nick: Any three numbers?
3. Tiger: Like 42, 43, 44. If you add them together – I need to think. I need to do the math in my head. Hold on, I'm just going to do my calculator.
4. Lisa: OK, so … yes.
5. Tiger: So, each time …
6. Lisa: Why do we believe it? Because …
7. Tiger: I just did the math in my head, and each time I got – it was divided by 3.

Tiger tested a specific example (42+43+44) and verified that the claim was true for the example (utterance 8). Then, he claimed that he did math in his head and *each* time the sum was divisible by three (utterance 12). This strategy is clearly indicative of an empirical argument, and the group seemingly accepted Tiger's claim (e.g. utterance 9). However, later in the group's interaction it became clear that they were aware of the limitations of an empirical argument. In the following exchange, Author 2 questioned the group's strategy and Nick shared the group's empirical argument. When Author 2 asked for other thoughts related to the task, Tiger made an attempt to orally explain a general argument (utterances 13–14).

1. Author 2: OK, that's good thoughts, yeah. Anything else? Something different? Same thing?
2. Tiger: Well, I just said I, um, three consecutive integers until I got to number 30, and then I figured that they could – if 30 is able to be divided by 3, then you could add 30s and 30s …

Clearly Tiger's argument (utterance 14) is not valid, but it appears that he tried to construct an argument that meets the generality criteria. His argument was based on inductive logic, though it clearly was not successful. The group did not make any explicit comments revealing they accepted Tiger's argument, so at this point we could not credit the group argument as an unsuccessful attempt at a valid general argument. However, later in their collaboration, Lisa revealed frustration and further attempted to create an argument that met the generality criteria (utterances 15–20).

1. Lisa: I don't know – like, facts. Like, without examples, I don't know how to do it.
2. Tiger: Each time I add three consecutive numbers together, it's able to be divided by 3. And this can go up to the millions.
3. Lisa: I think it depends on like the integers that are given cause if you just use random ones, or if it just starts out with the first ones, I think it would.
4. Nick: 1, 2, yeah.
5. Lisa: And then, you just add 3 more, and then you divide it by 3, and you get the base.
6. Tiger: The, every – you add this, it would be divided by 3.

Lisa's comment "then, you just add 3 more, and then you divide it by 3, and you get the base" (utterance 19) is somewhat difficult to decipher. It is clear that the argument is invalid, but she made an attempt to eclipse the empirical argument previously generated by the group. Paired with her acknowledgment that examples do not constitute a valid argument (utterance 15), we viewed her idea as an attempt at a general argument utilising inductive reasoning. Tiger and Nick engaged with Lisa's ideas (utterances 18 and 20), but they did not explicitly acknowledge it as the grounds for their group argument.

Taking all of the events of this collaborative exchange into consideration, Author 1 and Author 2 had difficulty in coming to agreement on the coding of the group's oral argument. Author 2 coded this collaborative oral argument as an empirical argument because the group agreed early within the exchange that the claim was true based on testing examples. Author 1 coded the oral argument as an unsuccessful attempt at a valid general argument because the group seemingly understood the limitations of an empirical argument and made two attempts (utterance 14 and utterance 19) to create general warrants. The coding of this oral argument is complicated because the group did not explicitly agree on a single strategy. However, after discussion, Author 1 and Author 2 agreed that because two members in the group attempted to create a general argument and realised the limitations of an empirical argument, the argument should be coded as an unsuccessful attempt at a valid general argument. In either case, the group's oral argument was of higher quality than their written argument. While this case increases the transparency in our use of the coding scheme, it should be noted that most cases were much simpler to discern, as evidenced by our interrater reliability. These two episodes provide narrative accounts of the differences between oral and written understandings as it relates to proof. In the next section, we discuss the findings and implications for future research and practice.

7. Discussion

In this paper, we shared general trends and a quantitative measure revealing the differences in 27 oral and written arguments constructed by eighth-grade students working in collaborative groups. We followed this analysis by examining two case groups which exemplified discrepancies between oral and written representations of arguments. While previous research has examined differences in oral and written arguments, this study represents the first attempt to quantify these differences. Additionally, our study utilised a unique temporal sequencing which complements Stylianides' (2019) analysis. By utilising the temporal sequencing of "orally communicate an argument, write an argument", our analysis refuted the possibility suggested in previous research that an oral argument may appear to be more sophisticated because it occurs after a written argument (Stylianides, 2019). In short, the findings of this study combined with Stylianides' (2019) findings suggest the temporal sequencing does not have an impact on whether the written or oral argument is of higher quality.

The general trends in the data suggest when groups' oral arguments represent an unsuccessful attempt at a valid general argument, they often write an empirical argument on paper. This could be due to the perception that an empirical argument would be better received by a teacher or peers than a general argument which contains errors. This might be an interesting phenomenon to explore in future research. The general trends also revealed four out of six oral arguments coded as proofs were scored as an unsuccessful attempt at a valid general argument when considering the written modality. One reason for this might be that students are unwilling to exert effort to write each component of the argument on paper. We have addressed our attempts to mitigate this concern in the methods section. However, even if it were the case that students were unwilling to exert effort to write a clear argument, it is an important phenomenon to report. An extensive amount of research studies make claims about students' proving capacities based on the written modality. If effort is a major concern related to written arguments, then it might not be beneficial to assess the written modality to measure proving abilities. However, we believe there were other factors which caused students to struggle in writing an argument. Amber's comments about the difficulty of writing an argument on paper were corroborated in other groups. The cause of the strange phenomenon of students' inabilities to write what they otherwise produce orally might be the subject of future research endeavours.

The narrative case studies revealed specific instances of differences between oral and written modalities related to creating arguments. Both groups' written arguments were not indicative of their actual understanding. This suggests implications for aiding students in improving their written arguments. Soto-Johnson and Fuller (2012) suggested students might benefit from recording themselves discussing an argument with others, listening to the recording and modifying written arguments to align with the oral representation. This practice might be beneficial for learners to critically examine how they produce arguments in each modality. The findings also have implications for assessing proving knowledge in the classroom. Teachers might choose to allow students to orally represent their arguments when assessing proving practices. While this could give a better indication of students' strengths and weaknesses in proving, learning how to write a clear argument is an important skill in mathematics.

A potential limitation of this study is our use of *collaboratively constructed* arguments to challenge claims of previous literature. Healy & Hoyles' (2000) study, for instance, made claims about *individual* students' proving competencies. In the introduction, we discussed our reasoning for using collaboratively constructed arguments, but our methodological choice complicates the comparison between previous literature and our study. Still, we consider this a minor limitation because collaboratively constructed arguments are based off the ideas of individuals, and, in most cases, the generator of an idea took over writing responsibilities for the collaborative argument.

In closing, this study advances the current notions of proof in school contexts. Our findings agree with previous literature (Soto-Johnson & Fuller, 2012; Stylianides, 2019) that students often possess argumentative knowledge that they struggle to express in written formats. This brings into question the reality of students' proving competencies as described by previous literature. We agree with Soto-Johnson and Fuller's (2012) assertion that current literature often utilises deficit models when discussing learners' proving capacities. Researching students' abilities related to proving is complex and may require different methodologies than those utilised in past studies. Searching for new ways to explore what students know in regard to proof rather than what they do not know might aid practitioners in locating and building on students' current knowledge.

Note

1. All student names are psuedonyms.

Disclosure statement

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

ORCID

Tye G. Campbell http://orcid.org/0000-0001-7827-8430
Jeremy Zelkowski http://orcid.org/0000-0002-5635-4233

References

Arbaugh, F., Smith, M., Boyle, J. D., Stylianides, G. J., & Steele, M. (2018). *We reason and we prove for all mathematics: Building students' critical thinking, grades 6-12*. Thousand Oaks, CA: Corwin Press.

Campbell, T. G., Boyle, J. D., & King, S. (2019). Proof and argumentation in K-12 mathematics: A review of conceptions, content, and support. *International Journal of Mathematical Education in Science and Technology*, doi:10.1080/0020739X.2019.1626503

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

Berland, L. K., & McNeill, K. L. (2010). A learning progression for scientific argumentation: Understanding student work and designing supportive instructional contexts. *Science Education*, *94*(5), 765–793.

Cáceres, M., Nussbaum, M., Marroquín, M., Gleisner, S., & Marquínez, J. T. (2018). Building arguments: Key to collaborative scaffolding. *Interactive Learning Environments*, *26*(3), 355–371. Accessed on November 12, 2018

Common Core State Standards Initiative. (2010). *Common Core State Standards for mathematics*. Retrieved from http://www.corestandards.org/Math/. Accessed on November 12, 2018

Department for Education. (2014). Mathematics programmes of study: Key stages 1 and 2: National curriculum in England. Retrieved from https://www.gov.uk/government/publications/national-curriculum-in-england-mathematics-programmes-of-study. Accessed 19 December 2018

Ellis, A. B. (2007). Connections between generalizing and justifying: Students' reasoning with linear relationships. *Journal for Research in Mathematics Education*, *38*(3), 194–229.

Emerson, R. M., Fretz, R., & Shaw, L. (2011). *Writing ethnographic fieldnotes*. Chicago: University of Chicago Press.

Evagorou, M., & Osborne, J. (2013). Exploring young students' collaborative argumentation within a socioscientific issue. *Journal of Research in Science Teaching*, *50*(2), 209–237.

Evens, H., & Houssart, J. (2004). Categorizing pupils' written answers to a mathematics test question: 'I know but I can't explain'. *Educational Research*, *46*(3), 269–282.

Flores, A. (2006). How do students know what they learn in middle school mathematics is true? *School Science and Mathematics*, *106*(3), 124–132.

Fried, M. N., & Amit, M. (2008). The co-development and interrelation of proof and authority: The case of Yana and Ronit. *Mathematics Education Research Journal*, *20*(3), 54–77.

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). Washington, DC: Mathematical Association of America.

Healy, L., & Hoyles, C. (2000). A study of proof conceptions in algebra. *Journal for Research in Mathematics Education*, *31*(4), 396–428.

Gibbons, J. D. (1993). *Non-parametric statistics: An introduction (Sage University paper series on quantitative applications in the social sciences, series 07-090)*. Newbury Park, CA: Sage.

Knight, A. M., & McNeill, K. L. (2015). Comparing students' individual written and collaborative oral socioscientific arguments. *International Journal of Environmental and Science Education*, *10*(5), 623–647.

Komatsu, K. (2016). A framework for proofs and refutations in school mathematics: Increasing content by deductive guessing. *Educational Studies in Mathematics*, *92*(2), 147–162.

Lannin, J. K. (2005). Generalization and justification: The challenge of introducing algebraic reasoning through patterning activities. *Mathematical Thinking and Learning*, *7*(3), 231–258.

Lee, C. Y., Chen, M. J., & Chang, W. L. (2014). Effects of the multiple solutions and question prompts on generalization and justification for non-routine mathematical problem solving in a computer game context. *Eurasia Journal of Mathematics, Science & Technology Education*, *10*(2), 89–99.

Lin, F. L., Yang, K. L., & Chen, C. Y. (2004). The features and relationships of reasoning, proving and understanding proof in number patterns. *International Journal of Science and Mathematics Education*, *2*(2), 227–256.

Liu, Y., & Manouchehri, A. (2013). Middle school children's mathematical reasoning and proving schemes. *Investigations in Mathematics Learning*, *6*(1), 18–40.

Martin, W. G., & Harel, G. (1989). Proof frames of preservice elementary teachers. *Journal for Research in Mathematics Education*, *20*(1), 41–51.

Nardi, E., & Knuth, E. (2017). Changing classroom culture, curricula, and instruction for proof and proving: How amenable to scaling up, practicable for curricular integration, and capable of producing long-lasting effects are current interventions? *Educational Studies in Mathematics*, *96*(2), 267–274.

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

Powell, A., Francisco, J. M., & Maher, C. A. (2003). An analytical model for studying the development of learners' mathematical ideas and reasoning using videotape data. *Journal of Mathematical Behavior*, *22*, 405–435.

Schoenfeld, A. H. (1994). What do we know about mathematics curricula? *The Journal of Mathematical Behavior*, *13*(1), 55–80.

Sen, C., & Guler, G. (2015). Examination of Secondary School Seventh Graders' proof skills and proof schemes. *Universal Journal of Educational Research*, *3*(9), 617–631.

Soto-Johnson, H., & Fuller, E. (2012). Assessing proofs via oral interviews. *Investigations in Mathematics Learning*, *4*(3), 1–14.

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

Stylianides, A. J. (2019). Secondary students' proof constructions in mathematics: The role of written versus oral mode of argument representation. *Review of Education*, *7*(1), 156–182.

Stylianides, A. J., & Al-Murani, T. (2010). Can a proof and a counterexample coexist? Students' conceptions about the relationship between proof and refutation. *Research in Mathematics Education*, *12*(1), 21–36.

Stylianides, A. J., Bieda, K. N., & Morselli, F. (2016). Proof and argumentation in mathematics education research. In A. Gutiérrez, G. C. Leder, & P. Boero (Eds.), *The second handbook of research on the psychology of mathematics education* (pp. 315–351). Rotterdam: Sense Publishers.

Stylianides, A. J., & Stylianides, G. J. (2009). Proof constructions and evaluations. *Educational Studies in Mathematics*, *72*(2), 237–253.

Stylianou, D. A. (2013). An examination of connections in mathematical processes in students' problem solving: Connections between representing and justifying. *Journal of Education and Learning*, *2*(2), 23.

Tsamir, P., & Sheffer, R. (2000). Concrete and formal arguments: The case of division by zero. *Mathematics Education Research Journal*, *12*(2), 92–106.

Weber, K. (2001). Student difficulty in constructing proofs. The need for strategic knowledge. *Educational Studies in Mathematics*, *48*(1), 101–119.

Weber, K., & Czocher, J. (2019). On mathematicians' disagreements on what constitutes a proof. *Research in Mathematics Education*, doi:10.1080/14794802.2019.1585936