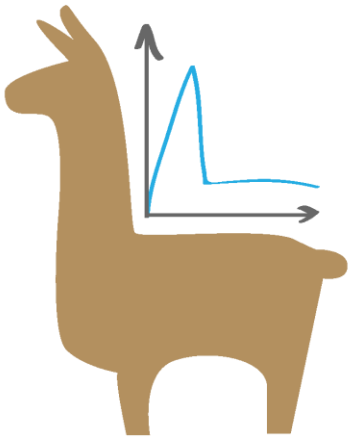
LONGITUDINAL LEARNING OF VIABLE ARGUMENT IN MATHEMATICS FOR ADOLESCENTS



LLAMA

August 2021

Year 5 Annual Report

Prepared for

National Science Foundation

2415 Eisenhower Avenue, Office E 11335

Alexandria, VA 22314

Longitudinal Learning of Viable Argument in Mathematics for Adolescents

Year 5 Annual Report

Prepared for
Michael Steele
National Science Foundation
2415 Eisenhower Avenue, Office E 11335
Alexandria, VA 22314

Prepared by

David Yopp, Principal Investigator
Anne E. Adams, Co-Principal Investigator
Rob Ely, Co-Principal Investigator
Veronica Blackham
University of Idaho
709 S Deakin Avenue
Moscow, ID 83843

Chandra Lewis, Co-Principal Investigator
Caroline Qureshi
RMC Research Corporation, Portland
111 SW Columbia Street, Suite 1030
Portland, OR 97201

Jean Hiebert Larson, Project Director
Larson Analytics
205 SW 2nd Avenue #2284
Estacada, OR 97023

Xin Wang, Co-Principal Investigator
Emma Espel
RMC Research Corporation, Denver
633 17th Street, Suite 2100
Denver, CO 80202

August 2021

Preferred Citation
Yopp, D., Hiebert Larson, J., Lewis, C., Wang, X., Qureshi, C., Adams, A., Ely, R., Blackham, V., (2021). Longitudinal Learning of Viable Argument in Mathematics for Adolescents Year 4 Annual Report. University of Idaho and RMC Research Corporation.

Award 1621438
Any opinions, findings, and conclusions or recommendations expressed in this brochure are those of the project Principal Investigator and Co-Principal Investigators and do not necessarily reflect the views of the National Science Foundation; NSF has not approved or endorsed its content.

Contents

Exhibits

Project Overview

The University of Idaho and RMC Research Corporation proposed a late stage design and development study to the National Science Foundation (NSF) Discovery Research K–12 (DRK–12) program that addressed the learning strand by studying the Longitudinal Learning of Viable Argument in Mathematics for Adolescents (LLAMA) intervention, an effort to improve Grade 8 students' mathematics learning through the construction of viable arguments, a national standard of mathematical practice. LLAMA was funded September 1, 2016 and will conclude at the end of a second no-cost extension year on August 31, 2022(NSF award 1621438). The project seeks to demonstrate the effectiveness of the LLAMA intervention and contribute to the knowledge base of student mathematical learning.

These goals are met by addressing 6 research questions:

1. To what extent did students in the treatment group demonstrate greater improvement on state assessments than students in the control group?
2. Does the implementation of the LLAMA intervention change **students'** ability to construct viable arguments and critique the arguments of others?
3. Does the implementation of the LLAMA intervention change **teachers'** ability to construct viable arguments and critique the arguments of others?
4. To what extent does treatment student learning align with that hypothesized in the LLAMA learning progression?
5. What pivotal intermediate conceptions are important for Grade 8 students in developing viable argumentation conceptions and practices?
6. What factors do teachers report as barriers to implementing the learning progression and the practice of teaching and learning through viable argumentation?

The LLAMA design is based on a review of current research and builds upon a DRK–12 exploratory study, Learning Algebra and Methods for Proving (LAMP), which developed a well-defined theory, intervention, and collection of materials. The LAMP pilot study showed promising results with a small sample (i.e., less than 50 students per condition) in which the treatment students outscored control students on Smarter Balanced Assessment Consortium (SBAC) state tests. Treatment students also made significant pre-post gains on the LAMP-developed argumentation protocol and control group did not.

The theory of action is:

- **If** teachers incorporate the LLAMA intervention into their curriculum and assessments,
- **Then** students will acquire the 12 conceptual pillars and increase their argumentation skills and mathematics achievement.

Treatment students experience the LLAMA intervention and the practice of teaching and learning with and through viable argumentation as features of daily instruction and regular assessment. To ensure implementation fidelity, LLAMA provided treatment teachers with school year and summer professional development workshops and regular coaching sessions in Year 1 and Year 2. In Year 3 and Year 4, the control teachers become a delayed treatment group and receive the professional development.

Intellectual Merit

A comprehensive understanding of how reasoning and proving skills develop alongside content learning in Grade 8 does not exist outside the LAMP pilot study. LLAMA addresses this gap in the research by extending the work of LAMP to all CCSS-M Grade 8 content domains and to larger and more diverse geographic settings to document students' learning trajectories and demonstrate that the LLAMA intervention is effective for all. Teaching of viable argument outside of high school geometry is meager despite calls over the past 2 decades from national organizations to place more attention on this standard at all grade levels. LLAMA will provide the resources teachers need to incorporate viable argument in their classroom by further developing and refining (a) a complete set of teacher materials that bring together the foundations for developing viable arguments and critiquing the arguments of others while targeting success with CCSS-M and the corresponding SBAC assessments and (b) an evidence-based learning progression that teachers can use to engage students in accessible proving tasks.

Broader Impacts

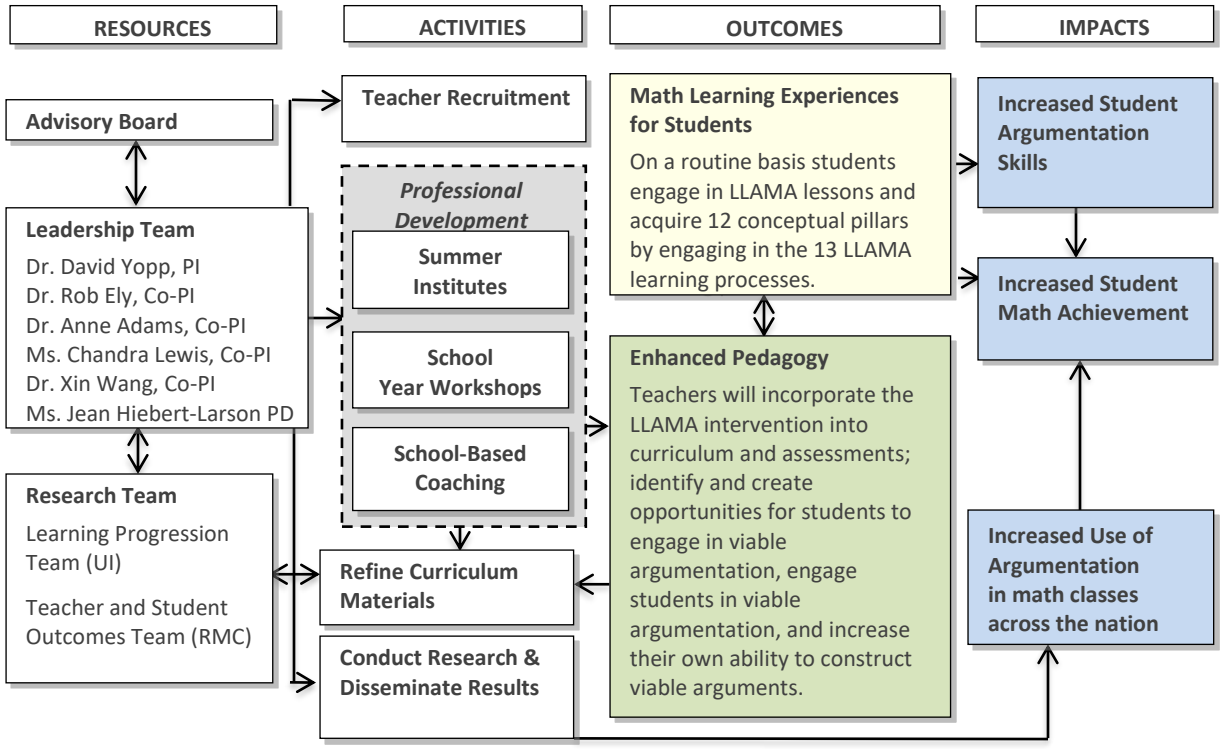
Beyond LLAMA's contribution to the research base on mathematics learning, LLAMA will (a) advance understanding of mathematics learning while promoting improved professional development of K–12 mathematics teachers by producing a detailed description of how to facilitate reasoning and argumentation learning in Grade 8 classrooms and meet the CCSS-M, (b) improve mathematics teaching and learning in the United States by developing curriculum materials and detailed instructions on facilitating viable argument in Grade 8 classrooms, and (c) improve students' viable argument skills, which are critical for a globally competitive STEM workforce.

Overview of Research Designs and Methods

Four study designs address the 6 research questions. RMC Research leads the research on the first 3 questions pertaining to the effectiveness of LLAMA on teacher and student outcomes. University of Idaho (UI) leads the research focusing on Research Questions 4–6, which promise greater understanding of how students learn and how teachers implement the intervention. This chapter provides an overview of the four original research studies. Over time, RMC Research and UI developed additional studies and modified the original study designs. The modifications and new studies are described in subsequent chapters. The LLAMA logic model is shown in Exhibit 1.

> *This section presents the research designs described in the proposal. Any major modifications to the designs are described with the report chapters for each study.*

Exhibit 1: LLAMA Logic Model



Research Questions

1. To what extent did students in the treatment group demonstrate greater improvement on state assessments than students in the control group?
2. Does the implementation of the LLAMA intervention change students' ability to construct viable arguments and critique the arguments of others?
3. Does the implementation of the LLAMA intervention change teachers' ability to construct viable arguments and critique the arguments of others?
4. To what extent does treatment student learning align with that hypothesized in the LLAMA learning progression?
5. What pivotal intermediate conceptions are important for Grade 8 students in developing viable argumentation conceptions and practices?
6. What factors do teachers report as barriers to implementing the learning progression and the practice of teaching and learning through viable argumentation?

Study 1: Student Achievement Study Design

RMC Research will conduct an experimental research study of the LLAMA intervention to address Research Question 1, "To what extent did students in the treatment group demonstrate greater improvement on state assessments than students in the control group?" The treatment group will consist of students whose teachers were randomly assigned to start participating in the LLAMA intervention in Year 1 and the control group consists of students whose teachers were randomly assigned to start participation in the LLAMA intervention in Year 3. In this design the independent variable is the LLAMA intervention and the dependent variable is state mathematics assessment scores (i.e., Smarter Balanced Assessment Consortium [SBAC] scores). The primary hypothesis is that students in the treatment group will improve significantly more in mathematics content learning measured by SBAC than students in the control group.

A hierarchical linear model (HLM) will be used as the primary analytic method. The study recognizes that that both mediating and moderating variables might have an impact on student achievement. Moderating variables are variables that exist at the time of the baseline and that may have an effect on outcomes (e.g., student gender, baseline achievement). Mediating variables are those that occur during the treatment time period and that may have an effect on the outcomes (e.g., number of coaching visits, hours of PD their teacher attended). At the time of the proposal the team identified 3 hypotheses to examine the moderating effects in secondary analyses. The first is that treatment teachers will be most effective in their third year of project participation; therefore, participation year is included as a moderator of the effect of the intervention on student outcomes. The effect of LLAMA on student outcomes is expected to be strongest for students with a treatment teacher in Year 3, who will have had 2 prior years of practice implementing the intervention. The second hypothesis is that treatment teachers who implement the LLAMA intervention with high fidelity will have a greater impact on student achievement than teachers who implement the LLAMA intervention with lower fidelity. A fidelity measure will be incorporated in the model as a moderating variable to assess the effect of the interaction between implementation fidelity and the intervention on student outcomes. To assess possible intervention mechanisms, the third secondary analysis hypothesis is that teacher content knowledge and practice mediate the relationship between the LLAMA intervention and outcomes. The Mathematical Knowledge for Teaching (MKT) assessment will be used to measure treatment and control teachers' baseline mathematical content knowledge for this moderating variable.

Study 2: Student Argumentation Study Design

RMC Research will conduct an experimental research study using a pre-post design and post-only design to address Research Question 2, "Does the implementation of the LLAMA intervention change the treatment students' ability to construct viable arguments and critique the arguments of others?" In the student achievement study design and this student argumentation study design, the treatment and control groups remain the same. The independent variable is the LLAMA intervention and the dependent variable is student argumentation and reasoning skills. In the pre-post design, treatment and control students in Years 1, 2, and 3 will complete the Student Argument and Reasoning Assessment at the beginning (pre) and end (post) of each school year. The pretest has 5 items: 4 that measure the ability to construct viable arguments, and 1 that assesses the ability to critique others' arguments. These items address mathematical content at the Grade 7 level to ensure the Grade 8 students have the mathematical knowledge necessary to adequately complete the assessment as a pretest at the beginning of their Grade 8 year (i.e., this approach ensures the assessment is measuring argumentation skills and not mathematical content knowledge). The posttest includes the same 5 items as the pretest and 4 additional items that address mathematical content that is taught to Grade 8 students during the school year. In the pre-post design, the hypothesis is that students in the treatment group will improve significantly more in argumentation skills than students in the control group (using the 5 items that are on both the pre and post). In the post-only design, the hypothesis is that students in the treatment group will score significantly higher on the posttest than students in the control group for the 4 items that are only included on the posttest.

Study 3: Teacher Argumentation Study Design

RMC Research will conduct an experimental research study to address Research Question 3, "Does the implementation of the LLAMA intervention change teachers' ability to construct viable arguments and critique the arguments of others?" In the student achievement study design and this teacher argumentation study design, the treatment and control groups remain the same. The independent variable is the LLAMA intervention and the dependent variable is teacher argumentation and reasoning skills. The independent variable is the LLAMA intervention and the dependent variable is teacher argumentation and reasoning skills. The treatment and control teachers complete the Teacher Argument and Reasoning Assessment (TARA) as a pretest in Year 1 and a posttest in Year 2. For both the pretest and posttest teachers complete the posttest version of the Student Argument and Reasoning Assessment (i.e., the one with 9 items; herein referred to as the Teacher Argument and Reasoning Assessment [TARA]). The hypothesis is that teachers in the treatment group will improve significantly more in argumentation skills than teachers in the control group.

Study 4: LLAMA Learning Progression Study Design

There are 3 major components to the learning progression study design. In the first component, University of Idaho will gather classroom work or assessments from all treatment students. Treatment teachers will submit 13 pieces of student data from 13 time points from all students in Years 1, 2, and 3. The student work will address the 12 processes for students to master and 12 related conceptual pillars. In the second component, University of Idaho will draw a random sample of 10 treatment teachers to participate in an intensive case study. In Years 1, 2, and 3, these teachers will be observed and interviewed 3 times each year. Both research teams will complete a Classroom Argumentation Observation Protocol at each observation and videotape the observations. University of Idaho will interview the teachers using the Teacher Interview protocol and record the interviews. The recording and videotapes will allow for in-depth analysis. In the third component at the beginning of Years 1, 2,

and 3, University of Idaho will draw a random sample of 10 treatment students each year. The students will each complete 12 Cognitive Task-Based Interviews (Ginsburg, 1997), which represents one interview for each of the processes/conceptual pillars expressed in the learning progression. Each interview is conducted immediately after the students' teacher implements a lesson associated with the process/conceptual pillar. The interviews will be videotaped and transcribed.

Utilizing all student data collected during the 3 components University of Idaho will use a methodology similar to Lobato et al. (2012) to address Research Question 4, "To what extent does treatment student learning align with that hypothesized in the LLAMA learning progression?" to assess the degree to which students' learning aligns with that hypothesized in the learning progression. Lobato, Hohensee, Rhodelhamel, & Diamond (2012) assert that learners might have rudimentary ways of coming to know and reason that are important for their development that have been forgotten by experts. These 12 conceptions become pivotal intermediate conceptions when they can be leveraged toward more sophisticated ways of reasoning. The majority of studies highlight differences between novices' ways of reasoning and proving and that of experts. To address Research Question 5, "What pivotal intermediate conceptions are important for Grade 8 students in developing viable argumentation conceptions and practices?" University of Idaho will use retrospective analysis of all teacher and student data collected during the 3 components to develop models of student conceptions at various time points, based on the methods of Miles, Huberman, & Saldaña (2013). This analysis draws upon frameworks for student thinking developed from previous iterations of the intervention and will be used to develop learning trajectories (Ellis, Weber, & Lockwood, 2014) that describe plausible paths through which students acquire more sophisticated thinking. Research documenting barriers to teachers implementing the practice of teaching and learning with and through viable argumentation is limited, and perhaps absent from the literature. Therefore a grounded theory design (Strauss & Corbin, 1998) will be used to systematically generate a theory to identify factors influencing teachers' implementation of the learning trajectory to address Research Question 6, "What factors do teachers report as barriers to implementing the learning progression and the practice of teaching and learning through viable argumentation?"

Description of Students' Mathematics Learning Experiences

This chapter describes students' LLAMA mathematics learning experiences. The theory of action is if teachers incorporate the LLAMA intervention into their curriculum and assessments, then students will acquire the 12 conceptual pillars[1] and increase their argumentation skills and mathematics achievement. Consistent with the NSF-funded project CAREER: Proof in Secondary Classrooms: Decomposing a Central Mathematical Practice, LLAMA hypothesizes that teaching students to construct viable arguments (DRL1453493, National Science Foundation, n.d.) and critique the arguments of others can be accomplished by addressing subgoals for proving and viably arguing. The LLAMA learning progression is expressed as a sequence of conceptual pillars, processes that target these conceptual pillars, and assessable intermediate outcomes (AIOs), which are student behaviors comprising a coherent collection of argument practices and conceptions of viable argumentation.

> ***Conceptual Pillar 1:*** *Students conceive of viable argument as requiring explicitly stated features: a claim, a foundation, and a descriptive or explanatory link between the foundation and claim.* ***Process 1:*** *Introduce the LLAMA argument framework: claim, foundation, and narrative link as a reminder of the minimal features of a viable argumentation.* ***AIO:*** *Students use the LLAMA argument framework to construct and critique arguments.*

Generalizing activities are supported by cultures that encourage justifying (Ellis, 2011). Currently, middle grades students are not often asked to support the conjectures they generate (Bieda, 2010). Students' naïve conceptions of argumentation in nonmathematical contexts are often incommensurable with the concepts of proof and viable argument in mathematics, and students are unlikely to discover mathematics-specific argumentation and proving conventions on their own (Bieda, 2010; Fischbein, 1982; Lobato, Clarke, & Ellis, 2005). Existing research and teacher support materials for middle school curricula lack appropriate standards for proving at the middle grades level (Bieda, 2010; Stylianides, 2009). EngageNY materials (New York State Education Department, EngageNY, n.d.), for example, incorporate numerous proving opportunities and provide teachers with examples of worked proofs, but LAMP data suggests this support is insufficient for developing teachers' and students' conceptions of proof. As a starting point for viable argumentation, LLAMA uses an argument layout (modified from Toulmin, 1958, 2003) to give students and teachers a classroom standard for the minimally needed features for viable argumentation: an explicitly stated claim, a foundation that supports the claim, and a narrative link (warrant) that explains how the foundation is used to support the claim.

> ***Conceptual Pillar 2:*** *Students conceive of the mathematics register as communicating precise meanings. Students conceive of 2 types of claims in mathematics—generalizations and existence claims—and they are acutely aware of the domain of the claims they present.* ***Process 2:*** *Introduce the language of mathematics for making claims (e.g., for-all, or-any, if-then, and there exists).* ***AIO:*** *Students use the language of mathematics to state claims; distinguish between existence claims and generalizations; and identify domains of the claims.*

The mathematics register uses precise meanings of terms in ways that are different from their everyday uses (Schleppegrell, 2007). Many students do not give proper attention to words such as *every*

[1] Thirteen conceptual pillars were originally proposed. Several of the conceptual pillars were related. After careful review by research and PD team members, the conceptual pillars were reorganized resulting in 12 conceptual pillars without losing any information.

(Galbraith, 1981), yet such terms signify important mathematical meanings. The appropriate use of the mathematics register is important for learning (Schleppegrell, 2007) and is intertwined with the practice of mathematics itself (Ball & Bass, 2000). There are 2 types of claims in mathematics—for-all and there-exist—and based on these quantifiers, arguers choose a mode of argumentation (e.g., example, exhaustion, deduction). The argument mode must fit the claim type. However, students have difficulty identifying the claim type. For-all statements can sound like there-exist statements to a novice (Yopp, 2015). Students who fail to distinguish between the 2 types of claims may choose inappropriate modes of argument (Yopp, 2015).

> ***Conceptual Pillar 3:*** *Students conceive of viable arguments for existence claims as providing an example in the domain of the claim and demonstrating that the example has the desired properties.* ***Process 3:*** *Introduce providing an example in the domain of the claim and demonstrating that the example has the desired properties as a viable mode of argumentation for existence claims.* ***AIO:*** *Students construct and critique existence arguments using this mode of argumentation.*

Students can hold misconceptions about the role of existence arguments unless this mode of argumentation is addressed properly (see Yopp, 2013, 2014). The Common Core State Standards for Mathematics (CCSS-M) include numerous content targets for which existence arguments are appropriate. For example, 2 triangles are congruent if and only if there exists a sequence of rigid motions that map one triangle onto the other (CCSS-M Grade 8, G.2). A viable argument for this claim provides an example (the sequence of rigid motions) and demonstrates that the example has the desired properties (maps one triangle onto the other).

> ***Conceptual Pillar 4:*** *Students conceive of empirical arguments as insecure support for a generalization.* ***Process 4:*** *Introduce skepticism by creating cognitive disequilibrium when students generalize based on exploring a few cases and then discover a counterexample using activities similar to those in Stylianides and Stylianides (2009).* ***AIO:*** *Students express skepticism of empirical arguments and express an intellectual need for more secure modes of argumentation.*

The finding that students at all levels are convinced by empirical evidence is robust (Stylianides & Stylianides, 2009). Untrained students may produce a few examples when asked to prove a generalization (Balacheff, 1988; Healy & Hoyles, 2000; Bieda, Holden, & Knuth, 2006; Porteous, 1990). Students may believe that examples prove the claim. Skepticism arises when students acculturate to the practices of mathematicians (Brown, 2014) and when they overgeneralize and find a counterexample later (Brown, 2014; Stylianides & Stylianides, 2009).

> ***Conceptual Pillar 5:*** *Students conceive of exhaustion as eliminating the possibility of counterexamples for generalizations with finite domains.* ***Process 5:*** *Introduce exhausting all cases as a viable mode of arguing for generalizations with finite domains.* ***AIO:*** *Students construct and critique arguments using this mode of argumentation.*

Students with strong reasoning skills tend to build mental models for a claim and use the models to explore the claim (Johnson-Laird, 1983). When the domain of a claim is finite, students can eliminate alternative models (i.e., counterexamples) by checking all cases. Constructing models of all possible counterexamples improves adolescents' reasoning skills and supports adolescents in eliminating counterexamples to a claim (Johnson-Laird).

> ***Conceptual Pillar 6:*** *A general pillar encompassing several others. Students conceive of proof as eliminating the possibility of counterexamples.* ***Process 6:*** *A general process that lays groundwork for further processes. Introduce pragmatic (Cheng & Holyoak, 1985) and mental models (Johnson-Laird, 1983) reasoning strategies using Wason Selection Tasks (Wason, 1966). Give special attention to the mathematics words—and, or, if, none, some, all—to ease students' linguistic struggles. Encourage listing the premises and prior results to ease working memory burdens when reasoning. Encourage the construction of models of claims' conditions and negated conclusions to find or eliminate counterexamples.* ***AIO:*** *Students make valid logical inferences and express an intellectual need for arguments that involve valid logic.*

> ***Conceptual Pillar 7:*** *Students conceive of valid reasoning for generalizations with infinite or large finite domains as applying viable logical reasoning schemas that eliminate the possibility of counterexamples.* ***Process 7:*** *Leverage mental models reasoning strategies to eliminate the possibility of counterexamples to generalizations.* ***AIO:*** *Students construct tables of mathematical objects that meet the conditions of a claim and mathematical objects that do not meet the conclusion. Students use these constructions to find or eliminate the possibility of counterexamples.*

Johnson-Laird (1983) asserts that the goal of all logical reasoning is to eliminate the possibility of counterexamples to claims. A definition of proof as eliminating the possibility of counterexamples appears to be unique to the LLAMA intervention. This conception arose from LAMP data where students validated arguments as follows: "This argument is viable because they proved that there are no counterexamples by proving that this is true for all cases" (Grade 8 LAMP treatment student). This conception proved to be a pivotal intermediate conception (Lobato, Hohensee, Rhodelhamel, & Diamond, 2012) which leveraged students toward more advanced ways of thinking of proof. It is consistent with Weber's (2014) assertion that proof should be defined as a cluster concept that includes multiple definitions of proof for a variety of educational purposes.

Invoking pragmatic reasoning schemas (Cheng, Holyoak, Nisbett, & Oliver, 1986) and mental models reasoning schemas (Johnson-Laird, 1990) improves deductive reasoning (see Stylianides & Stylianides, 2008, for a discussion of these schemas and mathematics education). Pragmatic reasoning theory asserts abstracted, pragmatic rules such as permissive and obligation schemas are invoked when reasoning (Cheng & Holyoak, 1985). Modals such as can, may, and must evoke rules such as "if action A is to be taken, then precondition B must be satisfied"; "if precondition B is not satisfied, then A may not be taken"; and "if A occurs, B must also occur." Studies associated with pragmatics reasoning schemas theory have been associated primarily with the Reduced Array Selection Task (RAST) or Wason Selection Tasks (Wason, 1966). Subjects test a rule "p implies q" by checking a minimal number of cards among those showing p, q, not p, and not q. Subjects do poorly on these tasks but improve with training (Cheng et al., 1986; Evans, 1982). Activating pragmatic reasoning (Girotto, Light, & Colbourn, 1988) improves performance. Increasing comprehension of logical (e.g., and, or, if, none, some, all) terms also improves performance (Johnson-Laird, 1990). Mental models reasoning theory asserts that an arguer's ability to build models for claims and search for alternative models influences reasoning skill (Johnson-Laird & Byrne, 1991). Arguers construct mental models of the information presented in premises and then construct concise descriptions of the models.

These descriptions can be used to conclude something not stated in the premises. Arguers then search for alternative mental models (i.e., counterexamples) that refute these conclusions. If alternatives are ruled out, the conclusion is taken as true. Practice managing models and limiting the number of

premises improves reasoning (Anderson, Howe, & Tolmie, 1996; Case, 1984; Johnson-Laird, Oakhill, & Bull, 1986).

> ***Conceptual Pillar 8:*** *Students conceive of referents as representative of all possible examples in the domain of a claim.* ***Process 8:*** *Introduce approaches for constructing referents in the foundation of an argument as a means of expressing generality (e.g., generic examples, variable expressions and equations, diagrams, prior results, and definitions).* ***AIO:*** *Students construct and use referents to express generality in the foundations of their arguments and determine whether a referent is representative of all possible examples in the domain of a claim.*

Referents such as examples can be useful in developing mathematical intuition and proofs (Burton, 1999; Fischbein, 1982; Hanna & Jahnke, 1996; Küchemann & Hoyles, 2009; Pedemonte, 2008; Sandefur, Mason, Stylianides, & Watson, 2013; Yopp, 2011b). An example can be any instantiation of a claim's conditions and conclusions, like a number sentence, picture, or diagram (Yopp & Ely, 2015; Yopp, Ely, & Johnson-Leung, 2015). The key to using an example appropriately when crafting arguments is to seek and express conceptual insights (Sandefur et al., 2013), which are structural features linking the conditions of a claim to its conclusion (Yopp, 2014; Yopp, 2015). An example can be a referent in a viable argument for a generalization when the example expresses a conceptual insight. Examples become generic examples (Rowland, 2002) when they are used to represent all examples in the domain of a claim and when the arguer appeals to only features of the example shared by all possible examples in the domain of the claim (Yopp & Ely, 2015). Nongeneric example reasoning results in a nonviable argument and occurs when the arguer appeals to a feature that is special to the example. These distinctions are found even among Grade 5 students' work (Adams, Ely, & Yopp, in press). Replacing representative cases with a variable can help students use their empirical work to develop more general arguments (Stylianides, 2007).

> ***Conceptual Pillar 9:*** *Students conceive of a viable argument for a generalization as requiring a conceptual insight that applies to all possible examples in the domain of a claim.* ***Process 9:*** *Introduce practice of searching for conceptual insights that express links between conditions and conclusions.* ***AIO:*** *Students construct referents that express conceptual insights linking the conditions of a claim to its conclusion; students know that viable argument for generalizations require a conceptual insight that links the conditions of the claim to the conclusion.*

At this stage of the intervention, treatment students have learned to express conceptual insights in referents such as examples. The next stage is to leverage conceptual insights to develop a more viable conception of explaining why. To some, the power of proof in school mathematics lies in explaining why a claim is true (Hanna, 1990, 2000; Hersh, 1993; Schoenfeld, 1994). Generic examples and other referents can have this explanatory power (Balacheff, 1988; Lannin, 2005; Yopp, 2009, 2010). As students manipulate examples and other referents, they become aware that they are searching for what causes a statement to be true. These searches entail *abductive* reasoning (Ely et al., 2014; Pedemonte, 2008).

> ***Conceptual Pillar 10:*** *Students conceive of a viable argument for a generalization as appealing to and using prior results.* ***Process 10:*** *Introduce practice of recognizing established facts that an argument relies upon and putting new facts "on the list" to be able to use in future arguments.* ***AIO:*** *Students are able to recognize and identify pieces of prior knowledge that are used in an argument.*

Proofs use logical and prior results to demonstrate the truth of a claim. When the inferences are correct, an argument is called *valid*. To be *sound*, an argument must be valid and based on true assumptions. Mathematicians create sound arguments by noting the axioms, definitions, and theorems used in their arguments. Even without using terms like *axiom*, Stylianides (2007) notes how Ball develops Grade 3 students' awareness of a proof's reliance on prior knowledge when they appeal to truths that are "on the list". Krummheuer (1995) uses the idea of *prima facie*—facts taken as self-evident—to show how "axiomatic-type" thinking occurs in early grades. With LLAMA, students are encouraged to appeal to definitions, accepted truths, and previously established results throughout the intervention.

> ***Conceptual Pillar 11:*** *Students conceive of an indirect argument for a generalization as viable because it eliminates the possibility of counterexamples.* ***Process 11a:*** *Revisit Wason Selection Tasks (Wason, 1966) with an emphasis on indirect argumentation. Introduce the concept of eliminating counterexamples by demonstrating that mathematical objects satisfying "not the conclusion" cannot satisfy the conditions.* ***Process 11b.*** *Students compare and contrast the collection of counterexamples for a generalization and the collection of counterexamples for its contrapositive.* ***Process 11c.*** *Introduce contradiction as an argument that eliminates the possibility of counterexamples to generalization.* ***AIOa:*** *Students construct indirect arguments by building models for the properties of possible counterexamples and use these models to find a counterexample or to eliminate the possibility of counterexamples. Students assess indirect arguments (contrapositive and contradiction) by determining whether the arguments eliminate counterexamples.* ***AIOb:*** *Students validate the logical equivalence of a conditional claim and its contrapositive by affirming that eliminating the possibility counterexamples to a claim also eliminates the possibility of counterexamples to its contrapositive, and vice versa.* ***AIOc:*** *Students construct contradiction arguments by constructing the collection of all possible counterexamples (descripted by the mathematical properties) then demonstrating that supposing a counterexample exists leads to an absurd or impossible statement.*

Indirect reasoning arises spontaneously in mathematics courses for students at all ages (Antonini & Mariotti, 2008; Reid & Dobbin, 1998; Thompson, 1996). During the LAMP pilot study, researchers found that students conceived of indirect arguments differently than experts. Experts tend to validate indirect reasoning based on their knowledge of logical theory. LAMP students often affirmed indirect reasoning as viable by noting that the possibility of a counterexample had been eliminated. For example, a LAMP student argued for the claim by writing $(2k + 1)(2b + 1) = 2(2kb + b + k) + 1$ and writing "I've proved that there aren't any counterexamples, because any odd number that's multiplied by another odd number will have to be odd . . . so, it's impossible for that to be a counterexample for the original claim" (Grade 8 LAMP treatment student). The student's reasoning can be described by a combination of mental models and pragmatic reasoning schemas.

In general, students tend to do poorly on indirect reasoning tasks (Antonini, 2004; Antonini & Mariotti, 2008; Leron, 1985). LLAMA leverages pragmatic and mental models reasoning to address this problem. Students confirm rules in RAST tasks by eliminating all counterexamples (Wason, 1966). Students reason as follows: all possible counterexamples are of the form p and not q; if in all cases of not q we have not p, then counterexamples cannot exist. LAMP students also successfully constructed contradiction arguments (e.g., the square root of 15 is irrational) using this mode of reasoning (e.g., by arguing there cannot exist a quotient of 2 integers equal to this number).

> ***Conceptual Pillar 12:*** *Students conceive of viable argumentation activities as requiring a decision about what mode of argument to use.* ***Process 12:*** *Offer opportunities to practice the modes of argumentation described above and opportunities to choose among these modes of argumentation.* ***AIO:*** *Students make appropriate choices about modes of argumentation relative to the task.*

Stylianides and Stylianides (2008) assert that students require practice to become proficient at reasoning and argumentation. In LAMP, students needed to practice modes of argumentation in a variety of contexts to become proficient. The LLAMA lessons offer opportunities to consider multiple modes of argumentation in one lesson. For example, when solving a linear equation, if the student finds a solution, then 2 claims can be made: *there exists a solution* and *for all other real numbers, none are solutions*. A student can argue for the latter claim by noticing that the equation $3x + 2 = 3x + 4$ is equivalent to the statement $2 = 4$.

Enhanced Pedagogy

LLAMA asserts that making viable argumentation a daily feature of teaching and learning and a regular feature of assessment can increase student achievement. A similar hypothesis is expressed in the NSF-funded project Preparing Urban Middle Grades Mathematic Teachers to Teach Argumentation Throughout the School Year (DRL 1417895, NSF, n.d.). LLAMA asserts that this disciplinary practice builds solid mathematical practices within students. As students solve problems, they make explicit claims about their solutions and their solution approaches. By building the conceptual pillars, students increase their ability to construct viable arguments, critique the arguments of others, and deepen their understanding of mathematics, resulting in increases in their performance on state assessments such as SBAC. Teachers facilitate these practices and mindsets by encouraging students to articulate mathematical claims using the mathematics registry precisely (Ball & Bass, 2000; Yopp, 2014, 2015). Teachers encourage students to negotiate their claims, to develop shared generalizations (Ellis, 2011), to be explicit about their support for claims, and to communicate conceptual insights (Yopp, 2014, 2015). Teachers leverage students' pivotal intermediate conceptions (Lobato et. al., 2012) of viable argument toward more sophisticated arguments that align with the practices of mathematicians (Stylianides, 2007). Teachers facilitate a daily practice of making mathematical claims with the largest domains possible relative to the data and conceptual insights students articulate. Consistent emphasis on these practices during instruction and assessments creates a mindset that viable argumentation and proof are central to mathematics (Knuth, 2002; Wu, 1996) and an important tool for learning mathematics (Knuth, 2002; Yopp, 2011a).

Study Instruments

This section provides a list of the study instruments and describes which participants complete each instrument and when. Many of the instruments are used across research studies; Exhibit 2 shows the primary study in which the instrument is used. Details regarding instrument creation are included within the study design chapters. In Exhibit 2, TX refers to treatment teachers and CT refers to control teachers.

Exhibit 2: List of LLAMA Instrument and Participant Completing Instrument

Instrument	Participant Completing Instrument
Student Achievement Study (led by RMC Research)	
Smarter Balanced Assessment Consortium (SBAC)	RMC Research obtains SBAC data and student demographic data from the school districts for 5 school years: 2 years of baseline data (spring 2015; spring 2016) and data for Years 1–3 of the project (spring 2017; spring 2018; spring 2019). **Modification.** RMC gathered SBAC data from the students of active CT teachers in spring 2019 and planned to gather data in spring 2020; however, all state testing was canceled in spring 2020 due to COVID-19.
Mathematical Knowledge for Teaching (MKT) Assessment Middle School Patterns, Functions, and Algebra Content Knowledge 2007	TX and CT teachers complete this 1-hour assessment at the beginning of Year 1 (pre) and end of Year 2 (post). The MKT is a mediating variable for this study. **Modification.** The MKT was administered for a third time at the end of Year 3. The MKT was not administered in Year 4.
Implementation Measure	**Modification.** Originally this measure was a fidelity measure for the TX group only. The research team assigned a code to each teacher in Year 2, 3, and 4 specifying the extent to which the teacher utilizes argumentation in their classroom.
Student Argumentation Study (led by RMC Research)	
Student Argument and Reasoning Assessment (SARA, Pretest)	TX and CT students complete this assessment at the beginning of each school year in Years 1, 2, and 3. **Modification.** This assessment was also administered at the beginning of the school year in Year 4 by a subset of teachers.
Student Argument and Reasoning Assessment (SARA, Posttest)	TX and CT students complete this assessment at the end of each school year in Years 1, 2, and 3. **Modification.** This assessment was to be administered at the end of the school year in Year 4 by a subset of teachers; however, due to COVID-19 only one teacher returned post data.
Teacher Argumentation Study (led by RMC Research)	
Teacher Argument and Reasoning Assessment (TARA)	TX and CT teachers complete this assessment at the beginning of Year 1 (pre) and at the end of Year 2 (post). **Modification.** The TARA was administered for a third time at the end of Year 3 and for a fourth time at the end of Year 4 to the active CT teachers.
Argument and Reasoning Assessment Rubrics	UI uses the rubric to score all teacher and student reasoning assessments.
LLAMA Learning Progression Study (led by University of Idaho)	
Teacher Interview Protocol I	University of Idaho interviewed the 9 TX case study teachers 3 times each during Year 2.

Teacher Interview Protocol II	Because several of the case study teachers became inactive and some reported less use of teaching via viable argumentation, the research team selected 11 teachers to interview in Year 3 with this revised protocol. Teachers were selected with differing levels of implementation, based on self-report and coach rating and with different levels of mathematics knowledge, as measured by MKT.
Classroom Argumentation Observation Protocol	UI observes TX teachers twice a year in Years 1, 2, and 3 and 3 times each year for TX case study teachers. **Modification.** CT teachers and a subset of TX teachers were observed in Year 4.
Monthly LLAMA Survey	All TX and CT teachers complete the survey each month in Years 1, 2, and 3. **Modification.** The survey was changed from a weekly to a monthly administration to reduce the data collection burden on teachers. This survey was only administered to active CT teachers in Year 4.
Student Work Samples	In Years 1 and 2 TX and CT teachers uploaded student work samples via a tablet each month. TX and CT teachers uploaded 2 samples: 1 demonstrating rich student understanding of argumentation and another representing partial understanding. TX teachers uploaded a third sample representing 1 of the 12 conceptual pillars. In Year 3 TX and CT teachers uploaded 3 student work samples (limited understanding, moderate understanding, and strong understanding) via a tablet at 3 points during the school year (October, January, and May). Other data collection changes in Year 3 included asking TX and CT teachers to identify the argument type from the item they chose and to include the feedback teachers would have provided to students based on their work. Year 3 changes were made to decrease teacher burden in a way that still enabled the research team to gain a rich understanding of teachers' comprehension of the different argument types and how they interact with their students. Year 4 data collection was the same as Year 3 with one exception: teachers were not asked to submit samples in May because of COVID-19 school closures. **Modification.** The data collection plan specified in the proposal (13 pieces of data from every student corresponding to each of the original 13 pillars) was not feasible. Coaches may score the student work samples using a structured scoring form in Year 5.
Cognitive Task-Based Interviews	Coaches conduct interviews with 20 treatment students from the case study teachers' classes: 10 in Year 2 and 10 in Year 3. Each year, 12 interviews will be conducted with each student, 1 interview for each process conceptual pillar expressed in the learning progression. Each interview is conducted immediately after the student's teacher implements the signature lesson associated with the process/conceptual pillar. **Modification.** In Year 1, the project was funded too late to make this meaningful because the teachers may have covered too little. In Year 2, 10 students were selected from 1 case study teacher's class. The number of student interviews was reduced from 12 to 6 to reduce teacher burden. This change was made to construct a richer data set. The team selected a teacher who was known to implement LLAMA with fidelity and whose location allowed the students to be interviewed by all of the UI PIs. Choosing students from one teacher known to be implementing the program with fidelity allowed the team to focus on the learning of students who had all received the treatment. In Year 3, no students were interviewed; however, an in-depth case study including student interviews was conducted in Year 4.

Coaching Log	Coaches complete the log after each coaching session with a LLAMA teacher.
Other	
Professional Development Survey	Teachers taking LLAMA professional development during the school year complete this survey to provide formative data.
Attendance Data	University of Idaho and RMC Research track attendance electronically (forms not included in this report).
Summer Survey	Teachers participating in the Summer Institute complete this survey to provide formative data (forms not included in this report).
Participation database	RMC Research records all teacher and student information in an Access database.

Note. TX = treatment, CT = control.

Measuring the Implementation of the LLAMA Intervention

The University of Idaho and RMC Research Corporation proposed a late stage design and development study to the National Science Foundation (NSF) Discovery Research K–12 (DRK–12) program that addressed the learning strand by studying the Longitudinal Learning of Viable Argument in Mathematics for Adolescents (LLAMA) intervention, an effort to improve Grade 8 students' mathematics learning through the construction of viable arguments, a national standard of mathematical practice. The project seeks to demonstrate the effectiveness of the LLAMA intervention and contribute to the knowledge base of student mathematical learning.

The LLAMA design is based on a review of current research and builds upon a DRK–12 exploratory study, Learning Algebra and Methods for Proving (LAMP), which developed a well-defined theory, intervention, and collection of materials. The LAMP pilot study showed promising results with a small sample (i.e., less than 50 students per condition) in which the treatment students outscored control students on Smarter Balanced Assessment Consortium (SBAC) state tests. Treatment students also made significant pre-post gains on the LAMP-developed argumentation protocol and control group did not.

The **logic model** (see Exhibit 1) for this project shows that the designated resources (National Advisory Board, Leadership Team, and Research Team) will work to implement project activities (teacher recruitment, professional development, refining curriculum materials, conducting research) that will result in two major outcomes (enhanced math learning experiences for students and enhanced pedagogy) with the ultimate impacts of increased student argumentation skills, increased student math achievement, and increased use of argumentation in math classes across the nation.

The research team designed several studies to measure various aspects of the project (as described in other sections of the report). A central component of the research is focused on **one outcome**, **increased understanding of students' math learning experiences.** The **theory of action** for this **outcome** is:

> ***If students*** *experience the LLAMA intervention,*
>
> ***Then*** *students will acquire the 12 conceptual pillars/5 argument practices and increase their argumentation skills and math achievement.*

The theory of action within the proposal was "If teachers incorporate the LLAMA intervention into their curriculum and assessments, then students will acquire the 12 conceptual pillars and increase their argumentation skills and math achievement. Treatment students experience the LLAMA intervention and the practice of teaching and learning with and through viable argumentation as features of daily instruction and regular assessment." The theory of action was revised as the research team honed the definitions of the various aspects of this complex project.

What is the LLAMA Intervention?

This project is studying an **instructional intervention** and not teacher professional development. For this project the professional development is used to help teachers implement the instructional intervention. LLAMA is an instructional intervention that combines a learning progression and the practice of teaching and learning with and through viable argumentation to improve students' abilities to construct viable arguments and critique the arguments of others (National Governors Association

Center for Best Practices & Council of Chief State School Officers [NGACBP & CCSSO], 2010) as they learn mathematics content. The **LLAMA instructional intervention** includes 3 parts:

1. The teacher engages students in learning experiences targeting the learning of the 12 CPs as content and practices. This "engagement" can include activity-based instruction, direct instruction, etc. In other words, LLAMA does not prescribe any particular instructional format.
2. The teacher includes viable argumentation, as described in LLAMA, as a regular feature of instruction and a part of assessment throughout the school year. This does not require daily inclusion. However, a teacher should attend to viable argumentation at least weekly, barring a handful exceptions, such as preparation for skill-based assessment.
3. Teacher should include viable argumentation for generalizations frequently, meaning at least twice a month, and have students attend to whether or not counterexamples to generalizations exist and, when students believe a generalization is true, have students develop descriptions of counterexamples and argue that counterexamples are impossible.

Based on the literature (Munter et al., 2014) the **LLAMA instructional intervention** could be classified as an unprescribed intervention. An unprescribed intervention has "two characteristics: (a) the instructional sequence and pacing are not predetermined (e.g., no topical, weekly plans are provided for teachers to follow), and (b) the choice of tasks is not predetermined" (pg. 84). As Munter notes at the conclusion of the article Assessing Fidelity of Implementation of an Unprescribed, Diagnostic Mathematics Intervention,

> "Many potentially high-quality interventions are unprescribed, require considerable tailoring by implementers, and rely on teacher knowledge and professional development. The rigorous evaluation of such programs requires the development of reliable fidelity measures that are both feasible to use and true to program components. The use of such measures enables evaluators to link assessments of fidelity of implementation to outcomes in order to more accurately determine the relative strength of interventions (Cordray & Pion, 2006) and to provide feedback to developers that will help in improving programs' effectiveness (Dusenbury et al., 2003)." Pg. 110

How Do We Know If Students Experienced the LLAMA Intervention?

To test the theory of action, the research team needs to know if the students experienced the LLAMA Intervention.

Year 1

During Year 1 the research team created a fidelity measure but the research team was not satisfied that this document fully captured the needed information.

Year 2

At the conclusion of Year 2, the research team convened several meetings to determine how to measure the extent to which students experienced the LLAMA intervention. The research team reviewed all available data sources and convened several meetings. After reviewing the data sources, the LLAMA team determined that:

> *26% of the active treatment teachers and 18% of the original treatment sample implemented the LLAMA intervention at the category they had hoped. 100% of the active treatment teachers implemented some parts of LLAMA intervention.*

At the onset of the study, 34 teachers were in the treatment group. As of the end of Year 2, there were 25 teachers in the LLAMA treatment group, 6 of which were identified by coaches as high LLAMA implementers, (i.e., 24% of the active treatment teachers and 18% of the original treatment sample). To determine who qualified as a high LLAMA implementer, coaches reflected on all the data (e.g., coaching logs, coaching interactions, observations, student samples, teacher administered assessments, and interviews) and deemed a teacher a high implementer if the teacher implemented all twelve conceptual pillars and made argument, as the LLAMA team envisions it, a regular part of their instruction (i.e., an almost daily feature). The LLAMA team worked intensely with Cohort 2 teachers in Year 3 and plan to continue in Year 4 of the grant to increase the number of high LLAMA implementers.

Year 3

By the beginning of Year 3 the research team had made several efforts to formalize all aspects of this process. First, the research team created a clear formative tool, Cohort 2 Implementation Guideline document, to provide teachers and coaches with clear implementation guidelines. Second, the research team presented a summative measure to the National Advisory Board in summer 2018. The purpose of the tool was to determine which teachers to define as high LLAMA implementers and to serve as a replacement for the older fidelity measure.

Over the course of Year 3 the research team carefully reviewed the National Advisory Board feedback, reviewed articles on fidelity of implementation in educational contexts, created various versions of an implementation/fidelity measure; and continued to hone the theory of action and intervention definitions. The article Defining, Conceptualizing, and Measuring Fidelity of Implementation and Its Relationship to Outcomes in K-12 Curriculum Interventions (O'Donnell, 2008) provided some historical context to measuring fidelity in education environments:

> "Fidelity of implementation is a relatively new construct in K-12 curriculum intervention research, but its use in program evaluation dates back 30-35 years…Although seemingly well defined in the health literature (cf. Hansen, Graham, Wolkenstein, & Rohrbach, 1991; Kolbe & Iverson, 1981), fidelity of implementation is rarely reported in large-scale education studies that examine the effectiveness of K–12 core curriculum interventions, especially with regard to how fidelity enhances or constrains the effects of the intervention on outcomes (L. D. Dobson & Cook, 1980; NRC, 2004; U.S. Department of Education, 2006). Moreover, according to the NRC (2004), even less seldom is such a measure of fidelity to K–12 curriculum interventions used to adjust for or interpret outcome measures" (p. 34).

The research team used the guidelines as described in Defining, Conceptualizing, and Measuring Fidelity of Implementation and Its Relationship to Outcomes in K-12 Curriculum Interventions (O'Donnell, 2008) and Assessing Fidelity of Implementation of an Unprescribed, Diagnostic Mathematics Intervention (Munter et al., 2014) to provide a framework for the research team to conceptualize how to effectively assess the fidelity of the LLAMA intervention. These steps are outlined below along with a status update for each step.

- **Step 1:** Ensure that the fidelity of implementation criteria and instruments are based on the underlying theory of the program being evaluated.
- **Step 2:** Ensure the program constructs, variables, and implementation processes are operationally defined.
- **Step 3:** Develop instruments to document the implementation of core components and processes as defined in the previous step.
  STATUS: The LLAMA research team is in the process of addressing Steps 1-3.

- **Step 4:** Assess fidelity for all teachers. If this is not possible draw a random sample of teachers so the findings can be generalized.
  STATUS: The research team coded each teacher.
- **Step 5:** Test and report the reliability and validity of instruments and the fidelity of data collected.
  STATUS: This step will not be possible given the sample size of the project and the resources needed to create a reliable and valid fidelity instrument.
- **Step 6:** Indices should be combined where appropriate (Nelson et al., 2010, 2012) and each should be related to outcomes where possible. O'Donnell argued that too often researchers monitor the structural aspects of fidelity without assessing users' fidelity to program processes and, in so doing, fail to account for the variation in FOI that is most strongly related to outcomes (Mowbray et al., 2003).
  STATUS: This step will not be possible given the sample size of the project and the resources needed to create a reliable and valid fidelity instrument.

After working through the framework, the research team realized that developing a valid and reliable fidelity instrument that other researchers could use was beyond the scope of this project. At the end of Year 3 the research team used a process similar to that in Year 2 to code the teachers' implementation category. To determine the appropriate code, the research team reflected on all of the available data (e.g., coaching logs, research team interactions with teachers, observations, student samples, teacher administered assessments, teacher surveys, and interviews). The coding system is described below and based on the 3 components of the LLAMA intervention.

- **High Implementer:** A teacher was coded as a '4' if the data showed the teacher (a) engaged students in learning experiences targeting the learning of the 12 CPs, (b) included viable argumentation as a regular feature of instruction, and (c) included viable argumentation for generalizations frequently (i.e., at least twice a month).
- **Medium Implementer:** A teacher was coded as a '3' if the data showed the teacher (a) engaged students in learning experiences targeting the learning of the 12 CPs, (b) included viable argumentation sometimes in their instruction, and (c) included viable argumentation for generalizations sometimes.
- **Low Implementer:** A teacher was coded as a '2' if the data showed the teacher (a) engaged students in learning experiences targeting the learning of some of the 12 CPs, (b) included viable argumentation infrequently in their instruction, and (c) included viable argumentation for generalizations infrequently.
- **No Implementation:** A teacher was coded as a '1' if the data showed the teacher did not start the project or there was no evidence of the teacher implementing LLAMA in the classroom

At the end of Year 3, of the 34 treatment teachers 12% (n = 4) were coded high implementers, 32% (n = 11) were coded as medium implementers, 41% (n = 14) were coded as low implementers, and 15% (n = 5) were coded as no implementation. Of the 19 control teachers who completed through June of Year 3 of the project, 5% (n = 1) were coded as high implementers, 32% (n = 6) were coded as medium implementers, 53% (n = 10) were coded as low implementers, and 11% (n = 2) were coded as no implementation. This implementation variable will be used as appropriate throughout the analyses.

Year 4

In Year 4, there were 12 control teachers who completed two years of the project by spring 2020. UI will rated these teachers at the onset of Year 5. Of the 12 teachers, 17% were coded as high implementers

(*n* = 2), 25% were coded as medium implementers (*n* = 3), and 58% were coded as low implementers (*n* = 7). Exhibit 3 shows the implementation categories by cohort for project years 3 and 4.

Exhibit 3: Implementation Categories by Cohort

	n	High	Medium	Low	None
Year 3					
Cohort 1	34	12%	32%	41%	15%
Cohort 2	19	5%	32%	53%	11%
Year 4					
Cohort 2	12	17%	25%	58%	0%

Note. Implementation scores were given to Cohort 1 only after their second year of intervention (Project Year 3). Implementation scores were given to Cohort 2 at the end of their first and second year of intervention (i.e., Years 3 and 4 respectively).

LLAMA Professional Development

The LLAMA professional development (PD) provides teachers with the concepts and skills they need to proficiently engage students in the LLAMA intervention. Preliminary findings from LAMP suggest that teachers have difficulties with implementing lessons that involve viable argumentation and proof because they lack the necessary understandings of viable argumentation and proving and how these activities link to content learning (Yopp, Sutton, Espel, &Wang, 2015). To ensure fidelity of implementation, the project provides planning guides that link LLAMA's 12 conceptual pillars to Common Core content and Grade 8 lesson plans that develop particular LLAMA conceptual pillars with supporting Common Core content. The professional development and coaching also assist teachers to:

- Identify and create opportunities in the LLAMA materials and existing curriculum materials for students to engage in constructing viable arguments with learning Common Core mathematics content.
- Improve their knowledge of viable argumentation and proving in Grade 8 mathematics content.
- Use instructional practices that engage students in viable argument.
- Develop pacing calendars for implementing the LLAMA intervention while covering Common Core content.

The professional development increases teachers' ability to construct viable mathematical arguments and might increase teachers' mathematics content knowledge. Teachers are supported through coaching sessions, summer professional development, and academic year professional development. Curriculum materials were refined extensively in Years 1, 2, 3, and 4 by UI and participating teachers, and numerous new lesson plans were created (e.g., several lesson plans on mathematical modeling with and through viable argumentation). Curriculum materials will be finalized in Year 5.

> ***Target:*** *RMC Research will develop and maintain a* ***participation database*** *to support project management, assist in the computation of variables based on teachers' participation, and to provide information to the advisory board pertaining to teacher recruitment, retention, and professional development offering.* ***Status: Met***

RMC Research created a Microsoft Access database in Year 1 in collaboration with the University of Idaho. This database is used to track a myriad of research information, including professional development attendance.

Academic Year Professional Development

> ***Target:*** *All Treatment teachers will attend three 4-hour professional development sessions during the school year in Year 1 (12 hours total) and again in Year 2 (12 hours total). Control teachers will attend four 3-hour professional development sessions during the school year in Year 3 (12 hours total) and again in Year 4 (12 hours total). This professional development was adapted to be an online course.* ***Status: Nearly Met****. By the end of Year 1, there were a total of 28 active treatment teachers, all of which completed the Year 1 academic year professional development. By the end of Year 2, there were a total of 23 active treatment teachers and 92% attended all of the academic year professional development. By the end of Year 3, there were 19 control teachers and 53-100% completed some portion of the PD and 80% completed more than half. By the end of Year 4, there*

were a total of 12 active control teachers and 83% attended all of the academic year professional development.

***Target:** 30 treatment teachers will attend professional development (PD) in Years 1 and 2.* ***Status:** **Nearly Met**: The 30 treatment teachers did not complete the full 2-years of the PD because some teachers dropped out of the project. By the end of Year 1, there were a total of 28 active treatment teachers, all of which completed the Year 1 academic year professional development. By the end of Year 2, there were a total of 23 active treatment teachers and 92% attended all of the academic year professional development.*

***Modified Target:** All treatment teachers will complete the BbLearn courses in Year 1 and in Year 2. Control teachers will complete the courses in Year 3 and 4.*
***Status:** **Nearly Met**: Cohort 1 course completion ranged from 92% to 100% for active teachers. Cohort 2 course completion ranged from 53% to 100% for active teachers with 79% completing more than half of the course (i.e., through CP9).*

Cohort 1 (Treatment Teachers)

Rather than three 4-hour sessions in Year 1, this professional development was adapted to be an online course composed of lessons that corresponds to the 12 conceptual pillars and was offered by the University of Idaho through Blackboard Learn (BbLearn; https://bblearn.uidaho.edu). The videos supported teachers in becoming comfortable with using and identifying viable arguments through examples and guided exercises. After watching the video teachers engaged in online discourse focused on how the conceptual pillar can be implemented in their classroom with their LLAMA coach and other LLAMA teachers. Ongoing coaching sessions assisted teachers in implementing the LLAMA intervention and assessing student work. A **BbLearn Feedback Survey** was developed to gather formative data about the professional development. Teachers completed this course independently.

All 28 treatment group teachers who were active as of May 31, 2017 completed the Year 1 school year professional development, though only one had completed all 12 by the end of the 2016–2017 school year as planned. The other 27 teachers completed the sessions prior to attending the summer 2017 professional development. Because LLAMA was funded just as the 2016–2017 school year was beginning, the project was not able to begin implementing the professional development with the treatment teachers in September 2016 as intended.

Rather than three 4-hour sessions in Year 2, the LLAMA coaches led four 1-hour online guided discussions with teachers during the school year (October 2017, December 2017, February 2018, and April 2018). Each session focused on implementation strategies for different groups of conceptual pillars. The final session in April focused on Conceptual Pillar 12 and also gave teachers an opportunity to discuss integrating argumentation practice into their preparation for the SBAC. Teachers were asked to post responses to questions on a group discussion board about a month prior to the live discussion. Online discussions were facilitated by coaches and delivered via Zoom.

In Year 1 treatment teachers participated in the LLAMA professional development through BbLearn individually (i.e., at their own pace). Attendance for the Year 1 BbLearn PD was tracked by UI staff through the BbLearn platform. A spreadsheet record was sent to RMC Research on a weekly basis and was used to update the participant database with attendance. The project did not meet the target of 30 treatment teachers attending in Year 1; however, 24 of 25 (96%) treatment teachers active as of May

31, 2017 had completed the Year 1 professional development by the end of summer 2017. The remaining active treatment teacher completed the Year 1 professional development during Year 2.

The Year 2 professional development was delivered to teachers as online guided discussions, led 4 times. Twenty-two of the 25 active treatment teachers (88%) attended all 4 sessions or make-up sessions. Two of the remaining teachers attended 3 of 4 sessions; the last teacher completed alternate activities in lieu of the professional development, since they were pursuing a masters' degree concurrently. To make up the PD session, attendees viewed the session video and posted to the group discussion, rather than attending the live PD session. These teachers will be flagged in the analysis as receiving alternative PD. Exhibit 4 shows the professional development attendance for all intent-to-treat RCT teachers.

Exhibit 4: Cohort 1 Academic Year
Professional Development Attendance Completion Rates

Lesson	*n*	Intent-to-Treat Completion Rate[a]	Active[b] Cohort 1 Teachers Completion Rate	Case Study Teachers[c] Completion Rate
Year 1				
Conceptual Pillar 1	29	85%	100%	100%
Conceptual Pillar 2	29	85%	100%	100%
Conceptual Pillar 3	29	85%	100%	100%
Conceptual Pillar 4	29	85%	100%	100%
Conceptual Pillar 5	29	85%	100%	100%
Conceptual Pillar 6	28	82%	100%	100%
Conceptual Pillar 7	28	82%	100%	100%
Conceptual Pillar 8	28	82%	100%	100%
Conceptual Pillar 9	28	82%	100%	100%
Conceptual Pillar 10	28	82%	100%	100%
Conceptual Pillar 11	28	82%	100%	100%
Conceptual Pillar 12	28	82%	100%	100%
Year 2[d]				
October 2017 session	23	68%	92%	89%
December 2017 session	24	71%	96%	100%
February 2018 session	24	71%	96%	100%
April 2018 session	23	68%	92%	89%

[a]*n* = 34 Cohort 1 teachers. [b]Active as of May 31, 2018. *n* = 25. [c]*n* = 9 case study teachers.
[d]Teachers who made-up sessions, rather than attending the live session: **October 2017:** 6 of 23 (26%); **December 2017:** 5 of 24 (21%); **February 2018:** 4 of 24 (17%); **April 2018:** 6 of 23 (26%).

BbLearn Survey

In the Year 1 BbLearn PD, at the conclusion of groups of conceptual pillars, called sessions, teachers were asked to provide their feedback on the session they most recently completed via a Survey Monkey survey. There are 5 sessions: Session 1 (Conceptual Pillar 1), Session 2 (Conceptual Pillars 2–4), Session 3 (Conceptual Pillars 5–7), Session 4 (Conceptual Pillars 8–10), and Session 5 (Conceptual Pillars 11–12). Feedback on Session 5 was gathered from teachers informally at the summer professional development. Each survey collects formative data in terms of teachers' self-reported preparedness to access sample lessons, engage students in the practices described for the designated pillar, examine student work in terms of the designated pillar, and create new lesson plans which incorporate the designated pillar. The survey includes 3 open-ended items: (1) Is there any area in which you want more clarity or training? (2) What would have made this session more useful for you? and (3) Is there anything else you want us to know? The survey was developed collaboratively by RMC Research and University of Idaho. These survey data are used formatively to improve the LLAMA professional development. RMC Research prepared 3 briefs throughout Year 1 summarizing results from the BbLearn Survey (February 2017, March 2017, and June 2017) for the LLAMA leadership team to review during the monthly meeting. All briefs are available upon request. The data collection completion numbers as of June 26, 2017,[2] are shown in Exhibit 5. With the exception of the Session 4 Survey, survey completion was high.

Exhibit 5: BbLearn Survey Completion Rates

Cohort 1	CP 1 [a]	Session 1 Survey	Completion Rate	CPs 2–4[a]	Session 2 Survey	Completion Rate
Total Teachers	29	26	**90%**	28	23	**82%**
	CPs 5–7 [a]	**Session 3 Survey**	**Completion Rate**	**CPs 8–10 [a]**	**Session 4 Survey**	**Completion Rate**
Total Teachers	18	16	**89%**	14	8	**57%**

[a]The sample sizes for each pillar are based on the number of teachers that have completed that BbLearn session and not the total number of LLAMA teachers.

Respondents to the BbLearn Survey rated all sessions good to very good, and by Session 4 reported feeling moderately to extremely prepared to access sample lessons. Respondents reported feeling moderately prepared to engage students in the practices and to examine student work in relation to Sessions 1–3 but reported needing assistance in these areas in relation to Session 4 (half reported feeling a little prepared in both areas). In terms of creating new lesson plans which incorporate the sessions conceptual pillars, respondents indicated needing assistance for Sessions 1–4 (at least half reported feeling a little prepared).

Cohort 2 (Control Teachers)

The control teachers began the professional development in Year 3, beginning with a kick-off meeting in August 2018. Twenty of the 21 active control teachers (both RCT and non-RCT; 95%) attended either the live kick-off meeting or the make-up session.

The Year 3 professional development followed a similar format to the PD offered in Year 1: videos were hosted online, and teachers viewed course materials at their own pace. For the control teachers, course

[2]June 26 rather than May 31 was used as the survey completion date to align with the anticipated end date of the online course.

materials were hosted on a Google website (https://sites.google.com/view/llama-project/llama-pd) rather than BbLearn, to reduce the burden of accessing the UI platform, which required dual authentication to log in. In lieu of online discussion boards, 4 synchronous meetings were held in Year 3 to discuss course materials (October 2018, January 2019, March 2019, and May 2019). Online discussions were facilitated by coaches and delivered via Zoom.

The Year 4 professional development was delivered to active teachers (*n* = 12) as online guided discussions, led 3 times (October 2019, December 2019, and March 2020). A fourth session was planned, but due to the stresses of COVID-19 on teachers, LLAMA coaches decided to facilitate a discussion around what types of general support teachers needed as they switched to distance learning in lieu of an argumentation PD session.

Exhibit 6 shows the Year 3 and 4 professional development attendance for both intent-to-treat RCT control teachers as well as active non-RCT control teachers. Active control teachers' participation in the self-paced modules was high through Conceptual Pillar 7 (89% completed through this pillar), but only about half of the active control teachers viewed all 12 conceptual pillar videos in Year 3 (10 of 19; 53%). In Year 3 synchronous meetings had higher attendance earlier in the school year (84–89% of active teachers attended the first 2 sessions) than later in the school year (58–68% attended the last 2 sessions). Ten of the 12 active control teachers (83%) attended all 3 synchronous meetings in Year 4.

Exhibit 6: Cohort 2 Academic Year Professional Development Attendance Completion Rates

Lesson	*RCT Control Teachers* *n*	Intent-to-Treat Completion Rate[a]	*All Control Teachers*[b] *n*	Active[c] Cohort 2 Teachers Completion Rate
Year 3				
Conceptual Pillar 1	15	48%	21	100%
Conceptual Pillar 2	15	48%	21	100%
Conceptual Pillar 3	15	48%	20	100%
Conceptual Pillar 4	15	48%	20	100%
Conceptual Pillar 5	15	48%	20	100%
Conceptual Pillar 6	12	39%	17	89%
Conceptual Pillar 7	12	39%	17	89%
Conceptual Pillar 8	11	35%	15	79%
Conceptual Pillar 9	11	35%	15	79%
Conceptual Pillar 10	11	35%	14	74%
Conceptual Pillar 11	9	29%	12	63%
Conceptual Pillar 12	7	23%	10	53%
Synchronous Zoom Meetings				
October 2018 session	14	45%	17	89%
January 2019 session	13	42%	16	84%
March 2019 session	11	35%	13	68%

May 2019 session	9	29%	11	58%
Year 4				
October 2019 session	9	29%	11	92%
December 2019 session	9	29%	12	100%
March 2020 session	8	26%	10	83%

[a]*n* = 31 Cohort 2 teachers. [b]Includes both RCT and non-RCT control teachers. [c]Year 3: Active as of May 31, 2019 (*n* = 19, 14 RCT teachers; 5 non-RCT); Year 4: Active as of May 31, 2020. (*n* = 12 (9 RCT teachers; 3 non-RCT).

At the beginning of Year 3 there were 21 Cohort 2 teachers. By the start of Year 4 (fall 2019) the number of active Cohort 2 teachers was 12. These 12 Cohort 2 teachers were active throughout all years of the project including through Year 4. Many Cohort 2 teachers dropped out of the project during the summer between Year 3 and Year 4. The reasons for dropping out of the project include moving to a new position, loss and illness in family, and differences in pedagogical approaches to Grade 8 math.

In Year 4, in spring 2020, COVID-19 related school closure and transition to distance learning was a significant hurdle for teachers on psychological, pedagogical, and logistical levels. Many teachers faced uncertainty in their personal lives and were faced with teaching virtually with no time to prepare and for some with small children at home and no childcare. This situation shifted teachers focus from implementing argumentation to just being able to keep up with district demands and keeping students engaged in a virtual environment. As such, the last professional development session of the year was canceled and instead coaches brought teachers together virtually to discuss the stresses they were facing in terms of implementing distance learning and how teachers and coaches could support teachers through this difficult transition.

PD Feedback Surveys

A PD Feedback Survey was submitted by each teacher after viewing the course materials for each conceptual pillar to both serve as a record of participation and also to collect formative data about the professional development and the coaching. The PD Feedback Survey included 3 open-ended items: (1) What did you find most useful about the conceptual pillar video, (2) What would have made the conceptual pillar video more useful, and (3) What assistance would you like from your coach regarding this conceptual pillar?. Open-ended answers were provided to coaches to discuss so that they could determine what types of assistance to provide to specific teachers.

To assess the quality and usefulness of the synchronous Zoom meetings, 3 questions were added to the monthly survey in months where there was a meeting: 1 rating item asking the overall quality on a scale from 1 (*poor*) to 4 (*very good*), and 2 open-ended items (Is there any area in which you want more clarity or training?, and What would have made this session more useful to you?). Participants rated the synchronous meetings as *good* consistently over time (Year 3: January: *M* = 3.1, March: *M* = 3.0, May: *M* = 3.1; Year 4: October: *M* = 3.6, December: *M* = 3.5, March: *M* = 3.8). Open-ended answers from both surveys were provided to coaches so that they could determine what types of assistance to provide to specific teachers.

Coaching

> ***Target:*** *All teachers will be assigned a coach.* ***Status:*** *Met*
>
> ***Target:*** *Coaches will be trained by David Yopp.* ***Status:*** *Met*

The LLAMA project set the goal of delivering 10 coaching sessions (in person and online) to treatment teachers during each year of the project. To prepare for coaching, coaches individually read the text West, L., & Staub, F. C., 2003 Content-focused coaching: Transforming mathematics lessons. Portsmouth, NH: Heinemann. Coaches watched videos of LLAMA coaching sessions performed by Yopp and discussed coaching moves relative to those proposed by West and Staub. The coaching team developed a shared modeling for LLAMA teacher coaching. Coaching sessions include lesson planning and development, pacing calendars development, assistance adapting existing LLAMA lessons or crafting new lessons, and reflecting on student work. Each teacher is assigned a LLAMA coach, a University of Idaho Principal Investigator or Co-Principal Investigator, who will assist with implementation and use coaching practices akin to those described in West and Staub (2003). Every treatment and control teacher who was active in Year 1 was assigned a coach in Year 1 (n = 28 and n = 25, respectively).

Coaches received training from Yopp, a Principal Investigator on the NSF DRK-12 project Examining Mathematics Coaching. A coaching session must include 4 parts: plan, observe, debrief, and next steps. A session can happen in person, on the phone, or remotely, but it must include the 4 parts. Coaches complete a coaching log, either electronically or on paper, that tracks the date, duration, and method of delivery. The coaching logs are then entered into the participant database by UI and RMC project staff.

Prior to Year 3, the LLAMA team decided to assign a lead coach to each teacher and to utilize a more team-oriented coaching approach in Years 3 and 4 with coaches visiting a variety of teachers, and not just their assigned teacher. The project director assigned Cohort 2 teachers a lead coach to balance the number of teachers per coach and to ensure that coaches had teachers that were geographically in the same area to ease travel burden for coaches. Cohort 2 teachers were virtually introduced to their head coaches for Year 3 and informed that the LLAMA team would employ some team coaching during the 2018–2019 school year. This adapted approach has mitigated challenges with coverage and also allowed teachers to interact with coaches with different perspectives.

> ***Target:*** *Treatment teachers will receive* ***10 coaching sessions*** *(in-person and online) per year in Year 1 and Year 2. Control teachers will receive 10 coaching sessions (in-person and online) per year in Year 3 and Year 4.* ***Status:*** ***Partially Met:*** *While all active teachers received at least one coaching session per year, the average number of coaching sessions per teacher was lower than 10 for all but one year (Year 2 for Cohort 1).*

Cohort 1 (Treatment Teachers)

Due to the late funding date, none of the treatment teachers received 10 coaching sessions in Year 1. One teacher received 5 coaching sessions in Year 1; most teachers had 1 or 2 coaching sessions. The LLAMA coaches conducted a total of 59 coaching sessions with the treatment teachers during Year 1 (Exhibit 7). One session was conducted online; the rest were in person. All teachers active as of May 31, 2017 participated in at least 1 coaching session during the 2016–2017 school year.

Exhibit 7: Cohort 1 Year 1 Coaching Completion

Observation	*Intent to Treat*	*Active[a] Cohort 1 Teachers*	*Case Study Teachers*
0 sessions	5	0	0
1 session	11	10	4
2 sessions	11	11	4
3 sessions	3	3	0
4 sessions	3	3	0
5 sessions	1	1	1
6 sessions	0	0	0
7 sessions	0	0	0
8 sessions	0	0	0
9 sessions	0	0	0
10 sessions	0	0	0
Total teachers	**34**	**28**	**9**
Total number of Year 1 coaching sessions	**59**	**58**	**17**

Note. All Cohort 1 teachers: *n* = 34. Case Study teachers: *n* = 9.
[a]Active as of May 31, 2017.

During Year 2, the LLAMA coaches conducted a total of 295 coaching sessions with the 25 treatment teachers (Exhibit 8). Some sessions were as short as 5 minutes while others lasted several days. The vast majority of coaching sessions were in person (250) and a small number were conducted online (45). Of the teachers active as of May 31, 2018, more than half received at least 10 gold standard (3-part) coaching sessions (14 of 25; 56%); most received 9 or more coaching sessions (20 of 25; 76%); and all teachers received at least 4 gold standard coaching sessions during the 2017–2018 school year. A gold standard coaching visit has 3 required parts.

1. A pre-lesson conference during which a teacher communicates plan and intended outcomes to coach or asks for coach assistance.
2. Either an observation of class or student work and data from class.
3. A post lesson conference during which next steps are discussed/planned. Any format, e-mail, in-person, zoom, etc. is acceptable.

Exhibit 8: Cohort 1 Year 2 Coaching Completion

Observation	*Intent to Treat*	*Active[a] Cohort 1 Teachers*	*Case Study Teachers*
0 sessions	9	0	0
1 session	0	0	0
2 sessions	0	0	0
3 sessions	0	0	0
4 sessions	1	1	0
5 sessions	0	0	0
6 sessions	1	1	0
7 sessions	2	2	1
8 sessions[b]	2	2	1
9 sessions	5	5	2
10 or more sessions	14	14	5
Total teachers	**34**	**25**	**9**
Total number of Year 2 coaching sessions	**295**	**295**	**99**

Note. All Cohort 1 teachers: *n* = 34. Case Study teachers: *n* = 9.
[a]Active as of May 31, 2018.
[b]For 2 of a case study teacher's coaching dates, a coach taught in the case study teacher's coteacher's class (non LLAMA teacher that co-teaches the same students as the case study teacher); therefore, there are 10 days of coaching for the students, but only 8 coaching sessions for the case study teacher. This exhibit only captures gold standard sessions.

Cohort 2 (Control Teachers)

Cohort 2 received coaching beginning in Year 3. The LLAMA coaches conducted a total of 143 coaching sessions with the 19 active Cohort 2 teachers during Year 3 and a total of 74 coaching sessions with the 12 active Cohort 2 teachers during Year 4 (Exhibit 9). Some sessions were as short as 5 minutes while others lasted several days. The vast majority of coaching sessions were in person (188) and a small number were conducted online (29). Of the teachers active as of May 31, 2019, about a quarter received at least 10 gold standard (3-part) coaching sessions (5 of 19; 26%); more than half received 7 or more coaching sessions (12 of 19; 63%); and all teachers received at least 3 gold standard coaching sessions during the 2018–2019 school year. Of the teachers active as of May 31, 2020, 100% of in-person coaching sessions were gold standard, half received 6 or more coaching sessions, and all teachers received at least 3 gold standard coaching sessions during the 2019-2020 school year.

Exhibit 9: Cohort 2 Coaching Completion

	Year 3		Year 4	
Observation	**Intent to Treat[a]**	**Active[b] Cohort 2 Teachers**	**Intent to Treat[a]**	**Active[b] Cohort 2 Teachers**
0 sessions	16	0	0	0
1 session	0	0	0	0
2 sessions	0	0	0	0
3 sessions	3	3	0	2
4 sessions	0	0	0	1
5 sessions	1	4	3	3
6 sessions	1	0	3	3
7 sessions	2	2	0	0
8 sessions	3	4	1	1
9 sessions	1	1	0	0
10 or more sessions	4	5	2	2
Total teachers	**31**	**19**	**9**	**12**
Total number of coaching sessions	**116**	**143**	**61**	**71**

Note. All Cohort 2 teachers: *n* = 37. RCT control teachers: *n* = 31. Non-RCT control teachers: *n* = 6.
[a]Intent to Treat" only includes RCT control teachers.
[b]Year 3: Active as of May 31, 2019 Includes both RCT and non-RCT control teachers (14 RCT control teachers; 5 non-RCT control teachers)
[b]Year 4: Active as of May 31, 2020. Includes both RCT and non-RCT control teachers (9 RCT control teachers; 3 non-RCT control teachers).

Summer Professional Development

> ***Target:*** *Treatment teachers attended a 2-week summer professional development in 2017 and the control teachers attended a 2-week summer professional development in summer 2018.* ***Status:*** ***Met***
>
> ***Target:*** *Teachers come to the summer professional development with existing products to be refined during the professional development and with data, observations, and questions that support further learning and reflection.* ***Status:*** ***Met***

For both the summer 2017 (Cohort 1) and summer 2019 session (Cohort 2), professional development focused on increasing teachers' knowledge of and skill with the LLAMA intervention and developing a personalized plan of implementation for the upcoming school year. The summer professional development was positioned after the first implementation year so teachers would have experience with the intervention prior to the summer session. Teachers came to the summer professional development with existing products to be refined during the professional development and with data, observations, and questions that support further learning and reflection. The goal for the summer professional development was for teachers to deepen their understanding of the LLAMA intervention and have support from LLAMA coaches, as they made concrete plans for their implementation in the

upcoming school year.

LLAMA summer professional development consisted of multiple sessions. There was a session for each major content area outlined by Grade 8 CCSS-M (number systems, expressions and equations, functions, geometry, and statistics and probability). The purpose of these sessions is to promote teacher understanding of CCSS-M and how to use argumentation with each of the content areas of CCSS-M. Embedded in the sessions are strategies to support English language learners with LLAMA vocabulary and strategies for teachers to plan their implementation of LLAMA for the 2017–2018 (Cohort 1) and 2019-2020 (Cohort 2) academic year with the support of coaches and other teachers.

Cohort 1 (Treatment Teachers)

Cohort 1 teachers attended 2 weeks of LLAMA professional development in summer 2017. Twenty-five of the 34 intent-to-treat Cohort 1 teachers (74%) participated in the summer PD . The majority of the teachers who were active as of July 1, 2017 (18 of 27; 67%) attended summer professional development in Moscow from July 17 through July 28. To accommodate the diverse group of teachers, who have unique time constraints, the team provided additional sessions in Blackfoot, ID and Idaho Falls, ID for 7 of the 9 remaining teachers. Of the 2 remaining teachers one dropped; the other completed alternate professional development activities during the 2017–2018 school year to receive the same summer professional development content. All Cohort 1 teachers active at the time of the summer PD received either 2 weeks of summer PD or an equivalent alternate PD activity. The teachers who attended summer PD in Blackfoot or Idaho Falls and the teacher who made up the PD during the 2017–2018 school year will be flagged in the analysis as receiving alternative PD. The chair of the National Advisory Board (NAB) designed and administered a feedback survey to teachers who completed the summer PD. These results were shared with the project team.

Cohort 2 (Control Teachers)

Cohort 2 teachers attended 2 weeks of LLAMA professional development in summer 2019. Eleven of the 12 Cohort 2 teachers (92%) active as of summer 2019 participated in some or all of the PD sessions; active Cohort 2 teachers who missed some or all of the summer PD completed make-up work in early fall 2019. RMC Research administered a feedback survey to teachers who completed the summer PD. These results will be shared with the project team.

Formative Evaluation

The research team conducted a formative evaluation in Years 1- 4. The formative evaluation was included in the Year 4 report.

Study 1: Student Achievement Study—Original Study

Original design. RMC Research will conduct an experimental research study of the LLAMA intervention to address Research Question 1, "To what extent did students in the treatment group demonstrate greater improvement on state assessments than students in the control group?" The treatment group will consist of students whose teachers were randomly assigned to start participating in the LLAMA intervention in Year 1 and the control group consists of students whose teachers were randomly assigned to start participation in the LLAMA intervention in Year 3. In this design the independent variable is the LLAMA intervention and the dependent variable is state mathematics assessment scores (i.e., Smarter Balanced Assessment Consortium [SBAC] scores). The primary hypothesis is that students in the treatment group will improve significantly more in mathematics content learning measured by SBAC than students in the control group.

A hierarchical linear model (HLM) will be used as the primary analytic method. The study recognizes that that both mediating and moderating variables might have an impact on student achievement. Moderating variables are variables that exist at the time of the baseline and that may have an effect on outcomes (e.g., student gender, baseline achievement). Mediating variables are those that occur during the treatment time period and that may have an effect on the outcomes (e.g., number of coaching visits, hours of PD their teacher attended). At the time of the preproposal the team identified 3 hypotheses to examine the moderating effects in secondary analyses. The **first hypothesis** is that students in the treatment group will improve significantly more in mathematics content learning measured by SBAC than students in the control group. The **second hypothesis** is that treatment teachers will be most effective in their third year of project participation; therefore, participation year is included as a moderator of the effect of the intervention on student outcomes. The effect of LLAMA on student outcomes is expected to be strongest for students with a treatment teacher in Year 3, who will have had 2 prior years of practice implementing the intervention. The **third hypothesis** is that treatment teachers who implement the LLAMA intervention with high fidelity will have a greater impact on student achievement than teachers who implement the LLAMA intervention with lower fidelity. A fidelity measure (implementation measure) will be incorporated in the model as a moderating variable to assess the effect of the interaction between implementation fidelity and the intervention on student outcomes. To assess possible intervention mechanisms, the **fourth secondary analysis hypothesis** is that teacher content knowledge and practice mediate the relationship between the LLAMA intervention and outcomes. The Mathematical Knowledge for Teaching (MKT) assessment will be used to measure treatment and control teachers' baseline mathematical content knowledge for this moderating variable.

Major Modifications. The research team decided not to collect data in Year 3 from the treatment teachers and would not test the second hypothesis, "treatment teachers will be most effective in their third year of project participation." LLAMA implementation was lower than expected with the treatment teachers and the research team decided to focus all of the future efforts and resources on the Cohort 2 teachers. The original study was concluded at the end of Year 2 and every hypothesis was tested with the exception of Hypotheses 2.

SBAC Executive Summary

Research Question 1 is, "To what extent did students in the treatment group demonstrate greater improvement on state assessments than students in the control group?" The research team used an HLM model building approach to address the hypotheses. The main finding is that the LLAMA intervention does not have a significant effect on SBAC scores. The **first hypothesis** is that students in the treatment group will improve significantly more in mathematics content learning measured by SBAC than students in the control group. To test this hypothesis the research team used HLM Model 5 with covariates including student baseline scores, teacher implementation fidelity categories, teacher MKT scores and teacher TARA scores. **The hypothesis was not supported for Wave 1**.[3] In the 2016-2017 school year, there was no statistically significant program impact on student SBAC scores. **The hypothesis was partially supported for Wave 2.** The HLM results suggest there was a statistically significant program effect on only SBAC Claim 1 scores.

The **second hypothesis** was not tested because data were only collected for 2 years.

The **third hypothesis** of the LLAMA study is that treatment teachers who implement the LLAMA intervention with high fidelity will have a greater impact on student achievement than teachers who implement the LLAMA intervention with lower fidelity. To test this hypothesis the research team used an HLM model that included teacher implementation categories as a covariate to account for teacher differences in implementing the LLAMA intervention. **The hypothesis was not supported.** There is no statistically significant relationship between teacher LLAMA implementation categories and SBAC scores. Yet, for teachers in their second year of LLAMA there is some evidence that teachers coded as implementation Category 2 or 4 are having a non-significant positive effect on SBAC scores. However, the four teachers coded as implementation Category 3 are having a non-significant negative effect on student math achievement. The research team will need to conduct some exploratory analyses to further investigate this finding.

To assess possible intervention mechanisms, the **fourth secondary analysis hypothesis** is that teacher content knowledge and practice mediate the relationship between the LLAMA intervention and outcomes. To test this hypothesis the research team used HLM Model 4 which included MKT baseline scores as teacher math content knowledge measures and TARA baseline scores as covariates (because teacher practice data was not available) to examine the relationship between the LLAMA intervention and these two teacher outcomes. **This hypothesis was not supported.** There is no statistically significant relationship between teacher content knowledge and argumentative reasoning skills and student SBAC scores for Wave 1 or Wave 2.

Study Recruitment and Random Assignment

To prepare for teacher recruitment University of Idaho completed and submitted a human subjects protocol to the University of Idaho Office of Research Assurances Institutional Review Board (IRB). The IRB approved the protocol on July 1, 2016. University of Idaho reviewed the application each June. Recruitment for LLAMA began in September 2016 and concluded in November 2016. The University of Idaho research team was responsible for teacher recruitment. Eligible teachers taught Grade 8 mathematics and were age 22 or older. University of Idaho recruited teachers through a multi-pronged approach that included contacting organizations in which they had existing relationships (e.g., state agencies, school districts); and sending out letters and informational fliers to principals, district

[3]Wave 1: Grade 8 students had a LLAMA teacher during the 2016-2017 school year and baseline data was obtained from their Grade 7 year in 2015-2016. Wave 2: students that had a LLAMA teacher in 2017-2018 in Grade 8 and baseline data was collected in Grade 7 from 2016-2017.

superintendents, and teachers. If teachers were interested in participating, University of Idaho provided teachers with a memorandum of understanding that clearly explained the purpose of the project, what their involvement would entail from attending professional development to providing the researchers with data, the risks of participation, and the benefits of participation. A similar document was created for principals. Teachers who wanted to join the study were asked to complete an application and sign a consent form. Principals of these teachers also signed a consent form.

Nine teachers who completed an application were included in the random assignment but did not submit a consent form and were dropped from the project's active participants. Those teachers are included in RCT teacher counts and omitted from active teacher counts throughout this report.

All students in the participating school districts are included in the study, and SBAC data will be obtained from all students in the participating school districts. RMC Research realized from past studies that is it much easier to request data from all students and then cull the data set down to the students that are in the study rather than try to request a subset of data from the school district. Students were in the treatment group if they had a LLAMA teacher. Students were in the comparison group if their teacher was randomly assigned to Cohort 2.

Power Analysis

> ***Target:*** *Power analysis with Optimal Design Software (Spybrook et al., 2011) reveals 50 teachers are necessary in the study to achieve desired power of .80 for student achievement.* ***Status:*** *Met at the time of recruitment but not met at the time of analyses. The full data set included data from 33 teachers (Treatment n = 17, Control n = 16) but the analytic sample includes data from only 22 teachers (Treatment n=13, Control n=9). Therefore, the analyses are underpowered.*

Prior to recruitment, the research team conducted a power analysis to determine the sample size necessary to detect the impact of the intervention. The study team conducted a power analysis using Optimal Design software (Spybrook et al., 2011), made specifically for power analyses for hierarchical cluster randomized designs. Teachers, as clusters, were randomly assigned to each the treatment or control group. Sample and cluster size were chosen to achieve a high level of power, greater than .80. The study team chose conservative parameter estimates for the analyses to avoid overestimating power. The assumed minimum detectable effect for this study was 0.25 standard deviation. The intraclass coefficient was set as 0.25, and we chose 0.50 of posttest variance explained by pretest scores for this power analysis, assuming each teacher has 20 students. Power analysis revealed that 50 teachers are necessary to achieve desired power of .80 for student achievement. To account for possible attrition, this study oversampled with a target of 60 teachers from the 3 states. With 30 teachers and 20 students per each class, approximately 600 Grade 8 students will receive the LLAMA intervention in each year for a total of 2,400 students in Years 1–4, and with 30 control teachers and 20 students per each class, approximately 600 Grade 8 students will be in the control group each year.

> ***Target:*** *2400 participating treatment students (600 per year) and 2400 participating control students across 4 years (600 per year). Estimate arrived by approximating 60 teachers in the study and 20 students per teacher each year.* ***Status:*** *Met in the analytic sample for treatment teachers but not for control teachers. For control teachers the range of students by year spanned 178 to 345.*

Teacher Participant Demographics

There were 76 applicants in total: 65 applicants were accepted to participate prior to the random assignment (described later in this section); 5 applicants were deemed ineligible to participate, because they do not teach Grade 8 CCSS-M; and 5[4] applicants applied after the start of the project and were admitted to participate in the professional development and project activities.

Exhibit 10 shows the demographics of the 71 accepted applicants. There are about the same number of teachers from Idaho and Washington, and slightly less than half as many from Montana (44%, 42%, and 14%, respectively). The teachers all have Bachelor's degrees, are predominantly White (97%), and a majority are female (75%). Most teach in a middle or junior high school (88%), and many teach in a rural school (67%). Approximately three quarters (75%) have a background in mathematics (degree major or minor, endorsement, or certification in mathematics); more than half (54%) have a Master's degree.

Exhibit 10: All Recruited Teacher Participant Demographics

Item	All Teachers[a]	Item	All Teachers[a]
Total Recruited	71	**Ethnicity[b]**	
State		White	97%
Idaho	44%	Asian	3%
Montana	14%	American Indian	2%
Washington	42%	**Gender**	
School setting		Female	75%
Rural	67%	Male	25%
Suburban	20%	**Years of experience (*M*)**	
Urban	13%	Years teaching total	11.6
School type		Years teaching mathematics	10.3
K–8	3%	**Highest level mathematics courses completed**	
K–12	2%	100–199 (freshman)	10%
Jr/sr high	4%	200–299 (sophomore)	11%
Middle/junior high	88%	300–399 (junior)	19%
High school	2%	400–499 (senior)	26%
Alternative	2%	500+ (graduate)	34%
Education and credentials		**Course credits in mathematics (*M*)**	
Bachelor's	100%	Undergraduate credits	21
Master's	54%	Graduate credits	6
Doctorate	2%		
Degree major/minor, endorsement, or certification in mathematics	75%		

Note. All (including non-RCT teachers): n = 59–71.
[a]Including non-RCT. [b]May have listed more than 1.

[4]Six non-RCT teachers applied; however, 1 non-RCT teacher who applied never fully joined the project, so that teacher is not included in the counts throughout the report.

Random Assignment

Teachers were randomly assigned to either treatment or control groups in November 2016. The 65 eligible teachers who applied before November were assigned a random number from a random number generator (Rand in Excel).[5] Teachers were then ordered by the random number. The first 33 teachers were assigned to the treatment group; the second 32 were assigned to the control group. After the initial random assignment, 3 schools had both the treatment and control teachers. These 3 schools, and a group of teachers from the same school district who agreed to participate under the condition that they will be in the same group, were randomly reassigned as blocks to avoid contamination effect within school. The reassigned teacher list has 34 treatment teachers and 31 control teachers. RMC Research and University of Idaho held an informational webinar for all RCT teachers (https://drive.google.com/open?id=1bRIHk4HKGUHS68BwLjjM2YZ6m8RNa10_). The 65 teachers were from 54 schools. The treatment group (Cohort 1) began LLAMA professional development activities in Year 1 and the control group (Cohort 2) will delay participation in LLAMA professional development activities until Year 3.

> ***Target:*** *Recruit at least 60 Grade 8 teachers from rural, suburban, and urban schools in Idaho, Montana, and Washington.* ***Status:*** *Met*
>
> ***Target:*** *Randomly assign 30 treatment, 30 control.* ***Status:*** *Met*

Exhibit 11 shows the teacher demographics by study group.

> *Control teachers (Cohort 2) have significantly more graduate credits in mathematics than treatment teachers (p = .002); however, there were no other significant differences detected between the treatment and control groups.*

Exhibit 11: RCT Teacher Participant Demographics

Teachers	Cohort 1 Teachers (Treatment)	Cohort 2 Teachers (Control)	All RCT Teachers	Case Study Teachers[a]
Total Recruited	34	31	65	9
State				
Idaho	44%	42%	43%	33%
Montana	12%	16%	14%	33%
Washington	44%	42%	43%	33%
School Setting				
Rural	65%	69%	67%	67%
Suburban	16%	24%	20%	22%
Urban	19%	7%	13%	11%

[5]The 5 eligible teachers who applied after November 2016 were admitted to participate in the professional development and project activities, but will not be included in any RCT analyses. These teachers are referred to as **non-RCT teachers** throughout this report.

Teachers	Cohort 1 Teachers (Treatment)	Cohort 2 Teachers (Control)	All RCT Teachers	Case Study Teachers[a]
School Type				
Middle/junior high	91%	84%	88%	78%
Jr/Sr High	6%	3%	5%	11%
K–8	3%	3%	3%	11%
K–12	0%	3%	2%	0%
Alternative	0%	3%	2%	0%
High school	0%	3%	2%	0%
Experience (*M*)				
Years teaching	10.0	13.2	11.5	9.1
Years teaching math	8.9	12.3	10.5	8.9
Education and credentials				
Bachelor's	100%	100%	100%	100%
Master's	47%	63%	55%	56%
Doctorate	0%	3%	2%	0%
Other[b]	77%	84%	80%	89%
Course credits in mathematics (*M*)				
Undergraduate	23	20	21	22
Graduate	3	11	7	2
Highest level of mathematics course completed				
100–199[c]	12%	7%	9%	0%
200–299[d]	9%	13%	11%	13%
300–399[e]	24%	13%	19%	38%
400–499[f]	30%	19%	25%	25%
500+[g]	24%	48%	36%	25%
Ethnicity (may have listed more than 1)				
White	100%	96%	98%	100%
Asian	3%	0%	2%	0%
American Indian	0%	4%	2%	0%
Gender				
Female	73%	77%	75%	71%
Male	27%	23%	25%	29%

Note. Cohort 1: *n* = 29–34; Cohort 2: *n* = 26–31; All RCT teachers: *n* = 55–65; case study teachers (subset of Cohort 1): *n* = 7–9. Non-RCT teachers are not included in this table. Because of the skew of the distributions, Mann-Whitney U tests were used to assess significance between cohorts for the Years Teaching and Credits variables. Chi-squared tests were used to assess

significance for proportions. No significance tests were conducted for variables where $n < 5$.
[a]Subset of treatment group. [b]Degree major/minor, endorsement, or certification in mathematics.
[c]Freshman level. [d]Sophomore level. [e]Junior level. [f]Senior level. [g]Graduate level.

Data Collection

This study has 1 primary data source: SBAC scores (outcome measure). Other data sources that are included in some models as mediating or moderating variables include Student Demographics, Mathematical Knowledge for Teaching (MKT)[6] assessment, and Teacher Argumentative Reasoning Assessment (TARA). This section describes the data collection process for the SBAC data.

SBAC Data Collection

> ***Target:*** *RMC Research will obtain* ***SBAC data*** *for participating schools/districts.* ***Status:*** *Partially Met with 43% of the districts submitting data (23 of 53 districts) and 51% of the teachers (33 of 65 randomly assigned teachers).*

Barriers to data collection included rural school districts not having the SBAC data in a readily accessible format nor staff available to compile the data. School districts were also concerned regarding student confidentiality. For this study, student mathematics achievement is measured using the students' SBAC scores. The participating states (Idaho, Montana, and Washington) administer the SBAC computer-based summative test at the end of each school year. RMC Research in collaboration with University of Idaho developed a data request form in Year 1. This form specifies what should be included in each data file. The research team sent this form to school districts via email in spring 2017 and again in spring 2018. Districts provide the research team with 5 data files: the first 4 data files include data from spring 2015, spring 2016, spring 2017, and spring 2018. As of June 30, 2019, 23 districts submitted data (Exhibit 12). The research team gathered data from all students in the participating school districts and not just the students of RCT teachers. This approach will allow the research team to have flexibility in terms of design and analyses in follow-up analyses.

Exhibit 12: SBAC Completion RCT Districts

Status	Treatment	Control	Total
Number of Districts in LLAMA	24	29	53
Number with Data Submitted	11	12	23 (43%)
Attempting to Get Data	6	6	12 (23%)
Will Not Get Data	7	11	18 (34%)

> *As Exhibit 13 shows, SBAC Data has been provided for 100% of treatment teachers who are high implementers and 55% of treatment teachers who are medium implementers[7]. Of the randomly assigned teachers, there is SBAC data from 50% of the treatment teachers (17 of 34) and 52% of the control teachers (16 of 31).*

[6]Copyright © 2006 The Regents of the University of Michigan. For information, questions, or permission requests please contact Merrie Blunk, Learning Mathematics for Teaching, (734) 615–7632.
[7] Implementation categories are defined in the Measuring the Implementation of the LLAMA Intervention chapter.

Exhibit 13: SBAC Completion by Implementation Status

	Number of Teachers			SBAC Data Received		
Status	**Tx**	**Control**	**Total**	**Tx**	**Control**	**Total**
1-No LLAMA	5	37	42	20%	43%	40%
2-Low Implementer	14	0	14	43%		43%
3-Medium Implementer	11	0	11	**55%**		**55%**
4-High Implementer	4	0	4	**100%**		**100%**
Total	34	37	71	50%	43%	46%

Note. Six non-RCT control teachers are included in this table. All Cohort 2 (Control) teachers had no LLAMA implementation during the time period for the analytic file (2015-2016, 2016-2017, and 2017-2018 school years).

What Works Clearinghouse Guidelines

What Works Clearinghouse utilizes three steps for reviewing RCTs and QEDs that assign individual subjects to the intervention or comparison condition: [8]

- **Step 1:** Assess the study design,
- **Step 2:** Assess sample attrition, and
- **Step 3:** Assess equivalence of the intervention and comparison groups at baseline (prior to the intervention).

Step 1: Assess the study design

"To be eligible for the WWC's highest rating for group design studies, *Meets WWC Group Design Standards Without Reservations*, the study must be an RCT with low levels of sample attrition. A QED or high-attrition RCT is eligible for the rating *Meets WWC Group Design Standards With Reservations* if it satisfies the WWC's baseline equivalence requirement that the analytic intervention and comparison groups appear similar at baseline. A QED or high-attrition RCT that does not satisfy the baseline equivalence requirement receives the rating *Does Not Meet WWC Group Design Standards*."

This study is an RCT.

Step 2: Assess sample attrition

The What Works Clearinghouse (Institute of Education Sciences, 2008) definition of overall attrition is:

- **Overall Attrition:** "Attrition is defined as a failure to measure the outcome variable on all the participants initially assigned to the intervention and control groups. High overall attrition generally makes the results of a study suspect, although there may be rare exceptions."

[8] Page 5; https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_handbook_v4.pdf

Student outcome data was collected from 51% of the 65 randomly assigned teachers. Using both the conservative and liberal attrition standard, not collecting 45% of the data is still within an acceptable range

The What Works Clearinghouse (Institute of Education Sciences, 2008) definition of differential attrition is:

- **Differential Attrition:** "Differential attrition refers to the situation in which the percentage of the original study sample retained in the follow-up data collection is substantially different for the intervention and the control groups. Severe differential attrition makes the results of a study suspect, because it may compromise the comparability of the study groups."

Exhibit 14 shows that there are 65 RCT teachers included in the study. For each wave, the school district provided complete data sets for 13 treatment teachers and 9 control teachers (i.e., the data included both pre and post SBAC student data). The overall attrition rate for Wave 1 is 67.65% in the treatment group and 74.19% in the control group. For Wave 2, the attrition rate is 61.76% in the treatment group and 70.97% in the control group. The intervention group response rates are always higher than the control group response rate for both waves.

The level of overall nonresponse (greater than 30 percent) and the levels of differential nonresponse require the establishment of baseline equivalence of the analysis sample in order to warrant a rating of "meets evidence standards with reservations" (What Works Clearinghouse 2017).

Exhibit 14: SBAC Data Received by Teachers

	Number of Teachers			Response Rate		
	Tx	Control	Total	TX	Control	Total
RCT	34	31	65	-	-	-
Wave 1 Post*	11	8	19	32.35%	25.81%	29.23%
Wave 2 Post**	13	9	22	38.24%	29.03%	33.85%

*Students had a LLAMA teacher during the 2016-2017 school year. Their teacher had attended one year of LLAMA professional development.
**Students had a LLAMA teacher during the 2017-2018 school year. Their teacher had attended two years of LLAMA professional development.

Step 3: Assess equivalence of the intervention and comparison groups at baseline (prior to the intervention).

Baseline equivalence tests were conducted in this study for two study samples. The first sample is Wave 1 in which Grade 8 students had a LLAMA teacher during the 2016-2017 school year and baseline data was obtained from their Grade 7 year in 2015-2016. The second sample is Wave 2 in which the Grade 8

students had a LLAMA teacher during the 2017-2018 school year and baseline data was obtained from their Grade 7 year in 2016-2017.

In Wave 1 the treatment and comparison groups were equivalent at baseline.

Exhibit 15 shows the Wave 1 treatment and control group baseline scores in 2015-2016. For Wave 1 the control group on average scored higher than the treatment group in SBAC overall scores, Claim 1 and Claim 2 & 4 scores. HLM analysis was conducted to test the baseline equivalence between the two groups using a null model where students are nested within classroom. Exhibit 16 shows that there is no statistical difference between treatment and control groups in SBAC overall measure and sub-claim measures. Therefore, in Wave 1 the treatment and control groups were equivalent at baseline. To account for any baseline score differences, student SBAC scores/sub-claim scores were still included in the HLM analysis model as covariates at student level.

Exhibit 15: SBAC Baseline Mean Scores in Treatment and Control Groups in 2015-2016

	Treatment			Control		
SBAC	***N***	**Mean**	***SD***	***N***	**Mean**	***SD***
Overall	670	2540.95	101.79	473	2549.67	78.35
Claim 1	607	2543.05	108.16	391	2545.91	82.44
Claim 2/4	607	2529.80	123.37	391	2541.33	91.01
Claim 3	607	2534.03	120.08	391	2532.70	102.65

Exhibit 16. SBAC Baseline Equivalence Test between Treatment and Control Groups in 2015-2016

	SBAC overall		Claim 1		Claim 2 &4		Claim 3	
	Estimate	***SE***	**Estimate**	***SE***	**Estimate**	***SE***	**Estimate**	***SE***
Intercept	2550.98***	18.06	2545.03***	23.01	2541.12***	26.33	2535.59***	23.99
Intervention Effect	-21.25	23.56	-15.28	28.96	-22.73	33.15	-14.69	30.18

****p* < .001[9].

In Wave 2 the treatment and comparison groups were equivalent at baseline.

Exhibit 17 shows the Wave 2 treatment and control group baseline scores in 2016-2017. HLM analysis was conducted to test the baseline equivalence between the two groups using a null model where students are nested within classroom. Exhibit 18 shows that there is no statistical difference between treatment and control groups in SBAC overall measure and sub-claim measures in 2016-2017. For Wave

[9] *p*-value is an indicator that represents the likelihood that observed results occurred by chance. In education research, values of $p < .05$ (i.e., values indicating that observed results had a less than 5% chance of occurring by chance) are typically used to identify results that are statistically significant. Lower *p*-values indicate a smaller likelihood that observed results occurred by chance and are therefore associated with statistically significant findings.

2, the treatment and control group baseline scores were equivalent. To account for any baseline score differences, student SBAC scores/sub-claim scores were still included in the HLM analysis model as covariates at student level.

Exhibit 17: SBAC Baseline Mean Scores in Treatment and Control Groups in 2016-2017

	Treatment			Control		
SBAC tests	***N***	**Mean**	***SD***	***N***	**Mean**	***SD***
Overall	700	2554.67	135.79	426	2558.98	93.81
Claim 1	648	2559.84	108.72	341	2555.60	91.98
Claim 2 & 4	648	2545.04	125.23	341	2541.85	114.14
Claim 3	648	2552.63	119.17	341	2539.61	113.30

Exhibit 18. SBAC Baseline Equivalence Test between Treatment and Control Groups in 2016-2017

	SBAC overall		Claim 1		Claim 2 &4		Claim 3	
	Estimate	***SE***	**Estimate**	***SE***	**Estimate**	***SE***	**Estimate**	***SE***
Intercept	2551.95***	19.35	2542.04***	20.96	2528.55***	25.84	2520.43***	21.99
Intervention Effect	4.14	25.53	18.53	26.77	15.48	33.00	32.71	28.06

****p* < .001.

Analytic Sample

Exhibit 19 shows the composition of the analytic samples for this reporting period. To date there are two waves for the treatment group: one that received the treatment during the 2016-2017 school year (Wave 1) and another that received the treatment during the 2017-2018 school year (Wave 2). Students in Wave 1 had a teacher that participated in the LLAMA professional for one year, while students in Wave 2 had a teacher that participated in the LLAMA professional development for two years. As the pretest scores were used as student level covariates in the HLM analysis model, the final analytic sample included only students with both pretest and posttest SBAC scores. Exhibit 19 displays the analytic samples based on when the student had a LLAMA teacher. In the next report data will be available from students who had a LLAMA teacher in Grade 8 during the 2018-2019 school year (Wave 3). As noted previously, the full data set included data from 33 teachers (Treatment *n* = 17, Control *n* = 16) but the analytic sample only includes data from only 22 teachers (Treatment *n*=13, Control *n*=9); only these teachers had complete data from both the baseline and Grade 8 years.

Exhibit 19: HLM Analytic Sample by Wave

	Number of Students			
	Wave 1		Wave 2	
Study Group	Baseline (2015-2016)	Grade 8 (2016-2017)	Baseline (2016-2017)	Grade 8 (2017-2018)

Treatment	Total	678	678	710	710
	Overall	670	669	700	704
	Claim 1	607	606	648	704
	Claim 2 & 4[a]	607	606	648	704
	Claim 3	607	606	648	704
Control	Total	474	474	429	429
	Overall	473	468	426	420
	Claim 1	391	387	341	386
	Claim 2 & 4[a]	391	387	341	386
	Claim 3	391	387	341	386

Note. Student baseline data was collected from the year preceding the year the student had a LLAMA teacher. Wave 1 students had a LLAMA teacher in 2016-2017 in Grade 8 and baseline data was collected in Grade 7 from 2015-2016. Wave 2 students had a LLAMA teacher in 2017-2018 in Grade 8 and baseline data was collected in Grade 7 from 2016-2017.
[a]On the state test Claim 2 and 4 are reported together.

Findings

Two-level HLMs with students nested within teachers were used to estimate the impact of LLAMA on students' mathematics achievement. The primary student outcome in this study was student math achievement measured using the students' SBAC scores. The participating states administered the SBAC computer-based summative test at the end of each school year. Student level SBAC data were obtained. The SBAC data include the overall scale scores of the test and sub-scores for three Claims, including Claim 1 Concepts and Procedures, Claim 2 Problem Solving & Claim 4 Modeling and Data Analysis, and Claim 3 Communicating Reasoning. The data also include teacher identifiers, which the study used to nest students within classroom.

Multiple HLM models were tested sequentially to test different research hypotheses. Models 1-4 are preliminary models used to develop Model 5 which provides the main findings about program impact. The following 5 models were developed:

- **Model 1** was a baseline model with no covariate to identify if there is any baseline difference between the treatment and control groups (see baseline equivalence section);
- **Model 2** included the effect of the intervention and student baseline SBAC scores as covariate;
- **Model 3** added teacher implementation fidelity covariate to account for teacher differences in implementing the LLAMA intervention (the research team is still in the process of deciding the best model and analytic approach to address fidelity implementation. This analysis will be included in the next report);
- **Model 4** included student baseline measure and additional teacher covariates: teacher MKT baseline scores that accounted for teacher mathematics content knowledge and teacher TARA baseline scores that accounted for teacher argumentative reasoning skills.
- **Model 5** is the final analytic model that includes student baseline measure and three teacher level covariates: implementation fidelity levels, MKT baseline scores, and TARA baseline scores.

The final analytic model included the student's pretest score as a robust covariate at the student level (Level 1). At the class/teacher level (Level 2), the model will include LLAMA group assignment

(intervention = 1, control = 0), teacher baseline math content knowledge, measured by MKT assessments at the beginning of the study; and teacher baseline argumentative reasoning skills measured by the TARA assessments. RMC Research conducted a 2-level HLM to identify the mediating effect of levels of LLAMA implementation on student outcomes, controlling for participation in the LLAMA intervention and other covariates. All HLM analyses were conducted for Year 1 and Year 2 separately to examine if program effects on student outcomes vary by year.

This section shows the findings for two waves of student data. All hypotheses are addressed with the exception of hypothesis 2 (the treatment teachers will be most effective in their third year of project participation). Hypothesis 2 is not addressed in this report because 3 years of data were not collected (see data collection section).

Hypothesis 1: Students in the treatment group will improve significantly more in mathematics content learning measured by SBAC than students in the control group.

First, the research team ran descriptive statistics as a naïve presentation of the data (i.e., not correcting for baseline differences or controlling for other variables) to descriptively see the difference between the treatment and control group posttest scores. Exhibit 20 shows that in 2016-2017 school year (Wave 1), control students scored higher in the post SBAC overall assessment and all three sub-claim measures. In Wave 2, however, treatment students outperformed in the SBAC Claim 1 and Claim 3 measures (see Exhibit 21).

Exhibit 20: Wave 1 SBAC Post Test Comparison between Treatment and Control Groups in 2016-2017

	Treatment			Control		
SBAC tests	***N***	**Mean**	***SD***	***N***	**Mean**	***SD***
Overall	669	2550.91	112.40	468	2564.79	93.56
Claim 1	606	2546.53	116.97	387	2561.63	100.83
Claim 2 & 4	606	2545.62	131.46	387	2546.80	109.38
Claim 3	606	2544.62	143.49	387	2546.39	115.33

Exhibit 21: Wave 2 SBAC Post Test Comparison between Treatment and Control Groups in 2017-2018

SBAC tests	Treatment			Control		
	N	**Mean**	***SD***	***N***	**Mean**	***SD***
Overall	704	2574.62	115.34	420	2576.44	114.37
Claim 1	704	2582.88	121.95	386	2569.73	119.82
Claim 2 & 4	704	2552.12	134.64	386	2558.02	130.37
Claim 3	704	2566.81	139.63	386	2560.33	147.86

Next, HLM model 2 was developed to examine the impact of LLAMA intervention where students were nested in classrooms/teachers and student baseline SBAC scores were used as a covariate.

The results suggest, in the 2016-2017 school year, there was no statistically significant program impact on student SBAC scores (Exhibit 22). Similar findings were also found for Wave 2 (see Exhibit 23).

Exhibit 22. HLM Model 2 Results Examining the Impact of LLAMA on Wave 1 SBAC Scores in 2016-2017

	SBAC overall		Claim 1		Claim 2 &4		Claim 3	
	Estimate	***SE***	**Estimate**	***SE***	**Estimate**	***SE***	**Estimate**	***SE***
Intercept	170.77**	51.17	542.20***	67.09	702.54***	70.84	803.07***	80.58
Baseline SBAC score	0.94	0.02	0.79	0.03	0.73	0.03	0.69	0.03
Intervention Effect	-6.50	7.69	-13.84	11.73	1.85	12.92	-8.87	15.42

***p* < .01, ****p* < .001.

Exhibit 23. HLM Model 2 Results Examining the Impact of LLAMA on SBAC Scores in 2017-2018

	SBAC overall		Claim 1		Claim 2 &4		Claim 3	
	Estimate	***SE***	**Estimate**	***SE***	**Estimate**	***SE***	**Estimate**	***SE***
Intercept	1027.05***	51.22	391.97***	64.06	793.18***	67.41	971.25***	80.62
Baseline SBAC score	0.60***	0.02	0.85***	0.02	0.69***	0.03	0.62***	0.03
Intervention Effect	**14.15**	**17.42**	**14.70**	**15.80**	**4.15**	**19.49**	**14.41**	**26.05**

****p* < .001.

The final analytic HLM model (Model 5) with covariates including student baseline scores, teacher implementation fidelity categories, teacher MKT scores and teacher TARA scores, however, shows positive program impact on SBAC post-test Claim 1 scores in Wave 2 (see Exhibit 24), but for Wave 1, in the 2016-2017 school year, there was no statistically significant program impact on student SBAC scores (Exhibit 25). For Wave 2, HLM results suggest there was a statistically significant program effect on SBAC Claim 1 scores, controlling for student pretest and teacher level covariates including TARA and MKT baseline scores and implementation fidelity categories. Treatment students in Wave 2 significantly outperformed control students in Claim 1 post scores.

For Wave 1 there was no statistically significant program impact on student SBAC scores. For Wave 2, HLM results suggest there was a statistically significant program effect on SBAC Claim 1 scores, controlling for student pretest and teacher level covariates including TARA and MKT baseline scores and implementation fidelity categories.

Exhibit 24. HLM Analytic Model Results Examining the Impact of LLAMA on Wave 1 SBAC Scores in 2016-2017

	SBAC overall		Claim 1		Claim 2&4		Claim 3	
	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Intercept	151.23*	60.28	519.68***	88.29	631.87***	87.47	854.42***	95.85
Baseline SBAC score	0.95***	0.02	0.81***	0.03	0.75***	0.03	0.68***	0.03
Implementation effect								
Category 2	-0.83	18.86	24.13	38.07	-11.28	34.90	32.35	39.99
Category 3	-20.54	16.19	-12.22	33.57	-52.89	31.25	0.00	35.85
TARA	-0.28	1.78	-0.66	2.76	1.11	2.45	-0.35	2.80
MKT	-2.60	15.12	-15.14	23.78	3.64	20.98	-34.61	23.92
Intervention effect	**4.32**	**15.00**	**-18.00**	**33.83**	**38.24**	**31.28**	**-23.87**	**35.88**

*$p < .05$; ***$p < .001$.

Exhibit 25. HLM Analytic Model Results Examining the Impact of LLAMA on SBAC Scores in 2017-2018

	SBAC overall		Claim 1		Claim 2&4		Claim 3	
	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Intercept	1139.41***	69.95	340.22**	85.58	826.26***	88.37	1041.64***	117.45
Baseline SBAC score	0.57***	0.02	0.87***	0.03	0.69***	0.03	0.62***	0.03
Implementation effect								
Category 2	10.82	35.17	-33.38	33.79	-17.60	34.35	25.33	56.63
Category 3	-53.08	29.42	-59.35*	28.92	-82.94**	29.55	-58.62	48.36
TARA	-1.93	2.89	0.38	2.93	1.04	2.95	-3.93	4.93
MKT	-12.50	21.36	-6.78	17.66	-41.43*	18.11	-9.45	29.50
Intervention effect	**37.13**	**23.91**	**52.23***	**25.31**	**45.65**	**25.84**	**39.12**	**42.36**

* $p < .05$, ** $p < .01$, ***$p < .001$.

Hypothesis 3. The treatment teachers who implement the LLAMA intervention with high fidelity will have a greater impact on student achievement than teachers who implement the LLAMA intervention with lower fidelity.

> In the Year 3 report to NSF the research team included some preliminary analyses of Hypothesis 3. This is a more in-depth analyses of Hypotheses 3.

The **third hypothesis** of the LLAMA study is that treatment teachers who implement the LLAMA intervention with high fidelity will have a greater impact on student achievement than teachers who implement the LLAMA intervention with lower fidelity. To test this hypothesis the research team used an HLM model that included teacher implementation categories as a covariate to account for teacher differences in implementing the LLAMA intervention. **The hypothesis was not supported.** There is no

statistically significant relationship between teacher LLAMA implementation categories and student SBAC scores. Yet, for teachers in their second year of LLAMA[10] there is some evidence that teachers coded as implementation Category 2 or 4 are having a non-significant positive effect on SBAC scores. However, the four teachers coded as implementation Category 3 are having a non-significant negative effect on student math achievement. The research team will need to conduct some exploratory analyses to further investigate this finding.

Description of Implementation Categories

The research team gave each LLAMA teacher an implementation category code. Codes ranged from 1 to 4.

- **High Implementer:** A teacher was coded as a '4' if the data showed the teacher (a) engaged students in learning experiences targeting the learning of the 12 CPs, (b) included viable argumentation as a regular feature of instruction, and (c) included viable argumentation for generalizations frequently (i.e., at least twice a month).
- **Medium Implementer:** A teacher was coded as a '3' if the data showed the teacher (a) engaged students in learning experiences targeting the learning of the 12 CPs, (b) included viable argumentation sometimes in their instruction, and (c) included viable argumentation for generalizations sometimes.
- **Low Implementer:** A teacher was coded as a '2' if the data showed the teacher (a) engaged students in learning experiences targeting the learning of some of the 12 CPs, (b) included viable argumentation infrequently in their instruction, and (c) included viable argumentation for generalizations infrequently.
- **No Implementation:** A teacher was coded as a '1' if the data showed the teacher did not start the project or there was no evidence of the teacher implementing LLAMA in the classroom.

Analytic Sample

The analytic sample includes 21 teachers: 12 treatment[11] and 9 control for both waves. Exhibits 26 and 27 show the numbers of teachers and students included by implementation category.

Exhibit 26. Number of Teachers by Implementation Category and Study Group

	Wave 1		Wave 2	
Implementation Category	**Treatment**	**Control**	**Treatment**	**Control**
Category 1: None	1	9	1	9
Category 2: Low	4	0	3	0
Category 3: Medium	4	0	4	0
Category 4: High	3	0	4	0

[10] Wave 2 which is comprised of students that had a LLAMA teacher in 2017-2018 in Grade 8 and baseline data was collected in Grade 7 from 2016-2017.

[11] The treatment *n* is 12 rather than 13 because only 12 teachers had SBAC data from both waves. One teacher did not have SBAC data from 2016-2017.

Exhibit 27. Number of Students by Implementation Category and Study Group

	Wave 1		Wave 2	
Implementation Category	**Treatment**	**Control**	**Treatment**	**Control**
Category 1: None	88	474	80	429
Category 2: Low	281	0	229	0
Category 3: Medium	222	0	257	0
Category 4: High	87	0	144	0

Analytic Plan

To investigate the relationship between student SBAC achievements and teacher implementation of LLAMA intervention, especially for those high implementing teachers, this HLM model comprised two analyses. In the first analysis all 4 implementation categories are included and Category 1 is used as a reference group in the analysis. For the second analysis teacher implementation is recoded as a dichotomous variable where Category 1 and Category 2 indicate low implementation fidelity (0) and Category 3 and Category 4 refer to high implementation fidelity (1). See Exhibit 28 for the analytic sample for the second analysis. In both analyses, Implementation category was still included in the HLM model as a covariate, along with student pretest scores, to identify the extent to which SBAC overall scores and sub-claim scores for LLAMA participating students vary according to different categories of LLAMA implementation.

Exhibit 28. Number of Teachers in Each Group: Dichotomous Coding of Implementation Variable

Implementation Category	Wave 1		Wave 2	
	Treatment	**Control**	**Treatment**	**Control**
Low	5	9	4	9
High	**7**	0	8	0

Analyses 1 Results: Category 1 Used as the Reference Group

Exhibit 29 and Exhibit 30 present the implementation analyses results for Wave 1 and Wave 2 respectively for Analysis 1 where Implementation Category 1 was used as the reference group. The hypothesis was not supported. While implementation fidelity varied across participating teachers, there is no statistically significant relationship between teacher implementation categories and student SBA scores for Wave 1 or Wave 2. The overall model is not significant. Data tables with descriptive statistics are shown in Exhibits 31-34. For both waves, students taught by Category 3 teachers had the lowest baseline SBAC scores across all four measures although these differences were controlled for in the HLM analyses.

Exhibit 29. Analysis 1: LLAMA Implementation Categories HLM Results in 2016-2017 (Wave 1)

	SBA overall	**Claim 1**	**Claim 2&4**	**Claim 3**

	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Intercept	177.19**	51.56	551.85***	67.54	712.72***	71.21	816.89***	81.19
Baseline SBA score	0.94***	0.02	0.79***	0.02	0.72***	0.03	0.68***	0.03
Implementation effect								
Low	-2.30	18.58	-7.99	25.36	-4.19	26.82	-44.98	33.80
Medium	-14.46	18.58	-28.72	25.37	-27.30	26.84	-42.20	33.83
High	6.43	20.02	-2.25	30.57	17.58	33.16	-19.05	41.29
Intervention effect	-1.85	17.53	0.66	24.37	12.11	25.75	27.17	32.50

**$p < .01$, ***$p < .001$.

Exhibit 30. Analysis 1: LLAMA Implementation Categories HLM Results in 2017-2018 (Wave 2)

	SBA overall		Claim 1		Claim 2&4		Claim 3	
	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Intercept	1030.08***	51.23	393.96***	63.86	797.54***	67.21	976.33***	80.59
Baseline SBA score	0.60***	0.02	0.85***	0.02	0.69***	0.03	0.62***	0.03
Intervention Effect								
Low	3.45	43.98	-20.32	32.77	13.33	41.70	11.42	58.87
Medium	-28.27	42.57	-31.60	31.70	-24.66	40.34	-33.67	56.98
High	24.08	43.02	25.99	33.72	45.68	42.87	45.24	60.07
Intervention effect	15.44	40.19	26.02	30.42	-1.44	38.72	12.74	54.62

*$p < .05$, ***$p < .001$.

Exhibit 31. Descriptive Statistics:
Implementation Category by Treatment Group in 2016-2017 (Wave 1)

		SBA overall		Claim 1		Claim 2&4		Claim 3	
Implem.	**Group**	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***
None	Control	2564.79	93.56	2561.63	100.83	2546.80	109.38	2546.39	115.33
None	Treatment	2612.62	65.40	2610.61	67.38	2601.84	87.37	2619.91	99.32
Low	Treatment	2563.73	128.09	2558.88	133.00	2560.94	148.39	2550.45	161.42
Medium	Treatment	2508.25	88.57	2502.93	94.63	2498.80	106.52	2505.19	120.94
High	Treatment	2554.61	110.95	2564.04	100.73	2587.83	130.23	2560.75	129.35

Exhibit 32. Descriptive Statistics:
Implementation Category by Treatment Group in 2017-2018 (Wave 2)

Implem.	Group	SBA overall	Claim 1	Claim 2&4	Claim 3

		Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
None	Control	2576.44	114.37	2569.73	119.82	2558.02	130.37	2560.33	147.86
None	Treatment	2567.01	98.44	2579.91	107.87	2530.40	128.73	2557.40	113.02
Low	Treatment	2608.71	109.45	2608.81	113.77	2598.02	125.38	2612.06	134.41
Medium	Treatment	2521.61	103.90	2531.32	112.72	2495.19	122.64	2503.04	133.34
High	Treatment	2617.64	115.46	2633.83	124.15	2590.90	133.20	2611.99	127.48

Exhibit 33. Descriptive Statistics:
Baseline SBA Scores by Implementation Category in 2015-2016 (Wave 1)

		SBA overall		Claim 1		Claim 2&4		Claim 3	
Implem.	**Group**	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***
None	Control	2549.67	78.35	2545.91	82.44	2541.33	91.01	2532.70	102.65
None	Treatment	2600.83	57.52	2605.30	72.16	2595.65	67.59	2597.61	55.40
Low	Treatment	2552.80	113.56	2552.16	120.26	2537.09	138.32	2547.11	131.97
Medium	Treatment	2504.85	87.28	2505.78	91.72	2491.55	110.72	2491.20	111.39
High	Treatment	2533.70	96.03	2550.08	82.30	2553.46	83.71	2541.42	86.13

Exhibit 34. Descriptive Statistics:
Baseline SBA Scores by Implementation Category in 2016-2017 (Wave 2)

		SBA overall		Claim 1		Claim 2&4		Claim 3	
Implem.	**Group**	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***
None	Control	2558.98	93.81	2555.60	91.98	2541.85	114.14	2539.61	113.30
None	Treatment	2532.87	91.56	2535.59	94.35	2511.95	111.44	2534.86	113.20
Low	Treatment	2599.83	99.90	2600.63	104.61	2592.07	125.96	2595.02	115.25
Medium	Treatment	2509.12	174.60	2519.42	105.57	2501.34	110.02	2510.20	110.48
High	Treatment	2574.16	95.68	2588.82	93.27	2574.95	125.08	2577.51	115.13

Analyses 2 Results: Implementation as a Dichotomous Variable

Exhibit 35 and Exhibit 36 present the results from the implementation analyses with teacher Implementation category treated as a dichotomous variable. Again, there is no statistically significant relationship between teacher implementation and student SBA scores for Wave 1 or Wave 2.

Exhibit 35. Analysis 2: LLAMA Implementation Category HLM Results in 2016-2017 (Wave 1)

	SBA overall	Claim 1	Claim 2&4	Claim 3

	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Intercept	167.63**	51.81	527.95***	67.85	689.35***	71.85	805.54***	82.28
Baseline SBA	0.94***	0.02	0.79***	0.03	0.73***	0.03	0.69***	0.03
Implementation effect	**5.14**	**10.11**	**16.20**	**13.95**	**14.06**	**15.45**	**1.42**	**19.55**
Intervention effect	-3.65	9.61	-5.68	13.47	8.90	14.87	-8.14	18.86

**$p < .01$, ***$p < .001$.

Exhibit 36. Analysis 2: LLAMA Implementation Category HLM Results in 2017-2018 (Wave 2)

	SBA overall		Claim 1		Claim 2&4		Claim 3	
	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Intercept	1022.44***	56.32	397.95***	66.76	788.19***	71.42	962.78***	86.76
Baseline SBA score	0.60***	0.02	0.85***	0.02	0.69***	0.03	0.62***	0.03
Implementation effect	**5.90**	**24.58**	**-5.63**	**20.82**	**7.28**	**25.65**	**11.03**	**34.37**
Intervention effect	18.10	24.07	11.37	20.78	8.86	25.60	21.49	34.32

*$p < .05$, ***$p < .001$.

Hypothesis 4. Teacher content knowledge and practice mediate the relationship between the LLAMA intervention and outcomes.

Teacher practice data was not collected in the current study. Instead, teacher baseline TARA data was used to estimate teacher argumentative reasoning skills. HLM Model 4 was developed to include MKT baseline scores as teacher math content knowledge measures and TARA baseline scores as covariates to examine the relationship between the LLAMA intervention and these two teacher outcomes. Exhibit 37 and 38 present the Model 4 analysis results for Wave 1 and Wave 2, respectively.

> *There is no statistically significant relationship between teacher content knowledge and argumentative reasoning skills and student SBAC scores for Wave 1 or Wave 2*

Exhibit 37. HLM Model 4 Results Examining the Impact of LLAMA on SBAC Scores in 2016-2017

	SBAC overall		Claim 1		Claim 2&4		Claim 3	
	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Intercept	143.82*	59.62	499.72***	82.70	645.28***	86.81	821.17***	90.64
Baseline SBAC score	0.95***	0.02	0.82***	0.03	0.75***	0.03	0.69***	0.03

TARA	-0.51	1.75	-0.43	2.40	-0.37	2.59	0.35	2.44
MKT	-0.01	15.02	-12.50	22.92	-3.10	24.70	-29.94	22.80
Intervention effect	**-4.55**	**11.31**	**-16.93**	**16.67**	**5.17**	**17.95**	**-13.29**	**16.51**

**$p < .01$, ***$p < .001$.

Exhibit 38. HLM Model 4 Results Examining the Impact of LLAMA on SBAC Scores in 2017-2018

	SBAC overall		Claim 1		Claim 2&4		Claim 3	
	Estimate	***SE***	**Estimate**	***SE***	**Estimate**	***SE***	**Estimate**	***SE***
Intercept	1140.94***	71.88	376.63***	83.13	867.86***	91.65	1049.96***	112.45
Baseline SBAC score	0.58***	0.02	0.87***	0.03	0.69***	0.03	0.62***	0.03
TARA	-2.84	2.97	-2.63	2.57	-2.22	3.17	-5.31	4.27
MKT	-2.72	22.36	0.51	19.12	-30.71	23.61	-2.36	31.84
Intervention effect	**19.27**	**19.92**	**16.82**	**17.32**	**3.62**	**21.39**	**18.52**	**28.91**

***$p < .001$.

Study 1: Student Achievement Study—Cohort 2 SubStudy

The research team decided to conduct an additional SBAC study with the 12 highly engaged Cohort 2 teachers. The research team decided to conduct this additional study because the Cohort 2 teachers were highly engaged in the LLAMA professional development for the entire time period. There are two designs for this study.

SBAC Executive Summary

This section investigates the same hypotheses as the Cohort 1 Student Achievement Study but presents findings for Cohort 2 for two SubStudies. First, there is a within treatment teacher study (Substudy 1) that represents 9 of the 12 teachers and includes SBAC data from the 7th graders and 8th graders taught in the treatment year (Project Year 3; 2018-2019 school year) and the SBAC data from the 7th graders and 8th graders taught by the same teachers prior to LLAMA (Project Year 1; 2016-2017 school year and Project Year 2; 2017-2018 school year). Second, there is a quasi-experimental study (Substudy 2) that includes both treatment and comparison data. The treatment data is comprised of students in Grades 7-8 from 7 Cohort 2 teachers with treatment data in Project Year 3 (i.e., 2018-2019 School Year). The comparison data is comprised of students in the same grade levels who did not have a LLAMA teacher in Project Year 3 but were in the same district as the 7 treatment teachers. This design includes 7 teachers instead of 12 because only 7 teachers had comparison data.

For Substudy 2 the treatment and comparison groups did not have similar baseline scores in Project Year 2. To control for this difference, student scores from Project Year 2 (i.e., the school year prior to the Cohort 2 LLAMA intervention) were included in the HLM analysis model as covariates at the student level. Research Question 1 is, "To what extent did students in the treatment group demonstrate greater improvement on state assessments than students in the control group?" The research team used an HLM model building approach to address the hypotheses. The main finding is that the LLAMA intervention has a partial significant effect on SBAC scores. The **first hypothesis** is that students in the treatment group will improve significantly more in mathematics content learning measured by SBAC than students in the comparison group. **This hypothesis was partially supported by the results from SubStudy 2.** Controlling for significant baseline inequivalence, the HLM model shows a partial significant positive LLAMA effect on their mathematics achievement in SubStudy 2. The treatment students scored statistically higher in the SBAC Claim 3 measure. They did not show significant difference in overall measure and sub-claim 1 and sub-claim 2 & 4. **Hypothesis 1 was not supported for SubStudy 1.** There was a statistically negative program impact on student SBAC Claim 1 and Claim 2 & 4 scores for students that had a LLAMA teacher in Year 3. That said, this analysis compares different groups of students over time and cannot control for potential differences among the student groups.

The **second hypothesis** was not tested because data were only collected for 2 years.

The **third hypothesis** of the LLAMA study is that treatment teachers who implement the LLAMA intervention with high fidelity will have a greater impact on student achievement than teachers who implement the LLAMA intervention with lower fidelity. To test this hypothesis the research team used an HLM model that included teacher implementation categories coded as 1-4 as a covariate to account for teacher differences in implementing the LLAMA intervention. **The hypothesis was partially supported for SubStudy 1**. There is a significant relationship between teacher implementation scores of 3 and student SBAC overall scores and their scores in Claim 1 and Claim 2 & 4 for SubSutdy 1. **For SubStudy 2, hypothesis 3 was not supported.** LLAMA implementation category did not had a significant impact on student mathematics achievement measured by SBAC. Due to the small sample size within

each implementation category (i.e., some categories with only teacher), these results should be interpreted with caution.

To assess possible intervention mechanisms, the **fourth secondary analysis hypothesis** is that teacher content knowledge and practice mediate the relationship between the LLAMA intervention and outcomes. To test this hypothesis the research team used HLM Model 4 which included MKT baseline scores as teacher math content knowledge measures and TARA baseline scores as covariates to examine the relationship between the LLAMA intervention and these two teacher outcomes. SubStudy 1 results revealed that there is no statistically significant effect of teacher content knowledge and practice on student SBAC scores. **This hypothesis was not supported in SubStudy 1**. This hypothesis cannot be tested for Substudy 2 because MKT and TARA scores were not collected from the comparison teachers.

Study Recruitment

Study recruitment is described within the chapter **Study 1: Student Achievement Study**.

SBAC Data Collection

> ***Target:*** *RMC Research will obtain SBAC data for the participating schools/districts of the 12 active Cohort 2 teachers.* ***Status:*** *Partially Met with 75% (9 of 12) of districts providing data for Year 3. Due to COVID-19, state achievement testing was cancelled for the 2019-2020 school year, as such there was no SBAC data to collect for Year 4.*

Exhibit 39 shows the data collected for this study. The research team collected SBAC data in Years 1 through 3, but due to COVID-19 the SBAC was not administered in Year 4 during the 2019-2020 school year. By the end of Year 4, there were 12 active Cohort 2 teachers. For the within treatment study, 9 of the 12 active Cohort 2 teachers submitted data for Years 1-3 (75%). For the quasi-experimental study, 7 of the 12 active Cohort 2 teachers submitted data for their students and comparison data for students in their district without a LLAMA teacher for Years 1-3 (58%).

Exhibit 39: SBAC Completion Cohort 2 SubStudy

	SubStudy 1 Within Treatment Study		SubStudy 2 Quasi-Experimental Study	
Year	No. of Cohort 2 Teachers[a]	% of Complete Data Sets	No. of Cohort 2 Teachers[b]	% of Complete Data Sets
Year 1 (Spring 2017)	9	75%	7	58%
Year 2 (Spring 2018)	9	75%	7	58%
Year 3 (Spring 2019)	9	75%	7	58%
Year 4 (Spring 2020)[c]	N/A	N/A	N/A	N/A

[a]Reflects number of teachers with treatment data
[b]Reflects number of teachers with treatment and comparison data
[c]SBAC was not administered spring 2020

Analyses and Findings

Assess equivalence of the intervention and comparison groups at baseline (prior to the intervention)

SubStudy 1: Within Treatment Study

The 7th grade and 8th grade students taught by the same treatment teachers in 2016-2017 and 2017-2018 (business-as-usual years) were used as the comparison group for the 2018-2019 treatment group in the same grade levels. Baseline equivalence cannot be compared because data was only collected at one point in time from the students.

SubStudy 2: Quasi-Experimental Study

The 2017-2018 student SBAC data were used as baseline data in this study. Exhibit 40 shows treatment and comparison group baseline scores in spring 2018. T-test results revealed that, on average, the treatment students scored significantly higher in SBAC overall scores and sub-claim scores than the comparison students. HLM analysis was also conducted to test the baseline equivalence between the two groups using a null model where students are nested within teacher. Similar findings were found in the HLM analysis as displayed in Exhibit 41. There was a significant treatment effect on SBAC overall measure and sub-claim measures. To account for baseline inequivalence, , student scores from Project Year 2 (i.e., 2017-2018 school year prior to the Cohort 2 LLAMA intervention) were included in the HLM analysis model as covariates at the student level.

Exhibit 40: SBAC Baseline Mean Scores in Treatment and Comparison Groups in SubStudy 2

	Treatment			Comparison		
SBAC	***N***	**Mean**	***SD***	***N***	**Mean**	***SD***
Overall* **	404	2568.40	86.79	2472	2522.05	103.58
Claim 1* **	372	2565.04	87.43	2440	2524.87	110.56
Claim 2/4* **	372	2553.62	107.34	2440	2512.88	118.97
Claim 3* **	372	2553.19	108.71	2440	2502.23	124.99

****p* < .001.

Exhibit 41. SBAC Baseline Equivalence Test between Treatment and Comparison Groups

	SBAC overall		Claim 1		Claim 2 &4		Claim 3	
	Estimate	***SE***	**Estimate**	***SE***	**Estimate**	***SE***	**Estimate**	***SE***
Intercept	2522.04	2.04	2524.62	2.15	2502.01	2.45	2512.62	2.34
Intervention Effect	45.33***	5.40	34.95***	5.69	44.94***	6.49	35.47***	6.20

****p* < .001.

Analytic Sample

Exhibit 42 shows the composition of the analytic samples for the two SubStudies.

Exhibit 42: HLM Analytic Sample by SubStudy

		Number of Students			
		SubStudy 1 Within Treatment Study		SubStudy 2 Quasi-Experimental Study	
Study Group		Grades 7/8 (2017-2018)	Grades 7/8 (2018-2019)	Baseline (2017-2018)	Grades 7/8 (2018-2019)
Treatment	Total	-	726	410	410
	Overall	-	718	404	407
	Claim 1	-	680	372	374
	Claim 2 & 4[a]	-	680	372	374
	Claim 3	-	680	372	374
Comparison	Total	1418	-	2473	2473
	Overall	1395	-	2472	2470
	Claim 1	1207	-	2440	2438
	Claim 2 & 4[a]	1207	-	2440	2438
	Claim 3	1207	-	2440	2438

Note. [a]On the state test Claim 2 and 4 are reported together.

Findings

Two-level HLMs with students nested within teachers were used to estimate the impact of LLAMA on students' mathematics achievement, measured using the students' SBAC overall scores and sub-claim scores. Multiple HLM models were conducted sequentially to test different research hypotheses. Models 1-4 are preliminary models used to develop Model 5 which provides the main findings about program impact. The following 5 models were developed. For SubStudy 1: Within Treatment Study, the analysis models did not include student baseline model as students taught by the same teachers in Year 1 and Year 2 data were used to form the comparison group. For SubStudy 2: Quasi Experimental, teacher mathematics content knowledge and teacher TARA baseline scores were not available for the comparison teachers. Therefore, only Models 1-4 were included in SubStudy 2.

- **Model 1** was a baseline model with no covariate to identify if there is any baseline difference between the treatment and comparison groups (see baseline equivalence section);
- **Model 2** included the effect of the intervention and student baseline SBAC scores as covariate;
- **Model 3** added teacher implementation fidelity covariate to account for teacher differences in implementing the LLAMA intervention;
- **Model 4** included student baseline measure and additional teacher covariates: teacher MKT baseline scores that accounted for teacher mathematics content knowledge and teacher TARA baseline scores that accounted for teacher argumentative reasoning skills.

- **Model 5** is the final analytic model that includes student baseline measure and three teacher level covariates: implementation fidelity categories, MKT baseline scores, and TARA baseline scores.

Hypothesis 1: Students in the treatment group will improve significantly more in mathematics content learning measured by SBAC than students in the control group.

Naïve descriptive statistics, as displayed in Exhibit 43, reveal the difference between the treatment and comparison group posttest SBAC scores for SubStudy 1: Within Treatment. The comparison group scored slightly higher than the treatment group across all SBAC sub-claim measures. The HLM null model analysis results showed no intervention effect on SBAC scores between the two groups (Exhibit 44). For SubStudy 2: Quasi-Experimental, however, the treatment group scored significantly higher in the overall SBAC measure and three sub-claims in the t-test analysis (Exhibit 45) and the null HLM model (Exhibit 46).

Exhibit 43: Substudy 1 SBAC Comparison between Treatment and Comparison Groups

	Treatment			Comparison		
SBAC tests	***N***	**Mean**	***SD***	***N***	**Mean**	***SD***
Overall	718	2581.41	104.42	1395	2580.09	106.10
Claim 1	680	2577.61	110.91	1207	2581.86	111.17
Claim 2 & 4	680	2569.60	130.79	1207	2570.76	131.21
Claim 3	680	2569.63	124.26	1207	2571.38	125.12

Exhibit 44: Substudy 2 SBAC Post Test Comparison between Treatment and Control Groups in Year 3

	Treatment			Comparison		
SBAC tests	***N***	**Mean**	***SD***	***N***	**Mean**	***SD***
Overall***	407	2583.81	96.34	2470	2537.57	109.55
Claim 1***	374	2574.38	101.37	2438	2539.29	116.32
Claim 2 & 4***	374	2574.13	121.03	2438	2519.94	129.59
Claim 3***	374	2568.10	117.48	2438	2524.49	128.55

***$p < .001$.

Exhibit 45. HLM Null Model Results Examining the Impact of LLAMA on Substudy 1 SBAC Scores

	SBAC overall		Claim 1		Claim 2 &4		Claim 3	
	Estimate	***SE***	**Estimate**	***SE***	**Estimate**	***SE***	**Estimate**	***SE***
Intercept	2586.26***	12.97	2581.49***	14.17	2568.22***	14.01	2568.81***	16.71
Intervention Effect	0.30	4.65	-0.48	5.16	1.58	5.86	3.0514	6.10

****p* < .001.

Exhibit 46. HLM Null Model Results Examining the Impact of LLAMA on Substudy 2 SBAC Scores

	SBAC overall		Claim 1		Claim 2 &4		Claim 3	
	Estimate	***SE***	**Estimate**	***SE***	**Estimate**	***SE***	**Estimate**	***SE***
Intercept	2537.57	2.17	2539.29	2.32	2524.49	2.58	2519.94	2.60
Intervention Effect	46.24***	5.77	35.09***	6.36	43.61***	7.06	54.19***	7.14

****p* < .001.

For the quasi-experimental study, HLM model 2 was performed to examine the impact of LLAMA intervention with student baseline SBAC scores as a covariate. *The results from SubStudy 2: Quasi-Experimental Study partially support hypothesis 1. C*ontrolling for significant baseline inequivalence, the HLM model shows a partial significant positive LLAMA effect on their mathematics achievement. The treatment students scored statistically higher in the SBAC Claim 3 measure. They did not show significant difference in overall measure and sub-claim 1 and sub-claim 2 & 4 (Exhibit 47).

Exhibit 47. HLM Model 2 Results Examining the Impact of LLAMA on Substudy 2 SBAC Scores in Year 3

	SBAC overall		Claim 1		Claim 2 & 4		Claim 3	
	Estimate	***SE***	**Estimate**	***SE***	**Estimate**	***SE***	**Estimate**	***SE***
Intercept	299.57	39.01	440.64	31.50	822.19	36.81	795.74	40.48
Baseline SBAC score	0.89***	0.01	0.83***	0.01	0.68***	0.01	0.69***	0.02
Intervention Effect	5.72	3.21	2.26	3.99	9.39	5.37	26.68***	5.60

****p* < .001.

Next, the final analytic HLM model (Model 5) with teacher level covariates were conducted for SubStudy 1 to further examine the LLAMA impact. **The results from SubStudy 1 did not support hypothesis 1.** Exhibit 48 shows that, controlling for LLAMA implementation fidelity category, teacher mathematics content knowledge measured by MKT, and their argumentative skills measured by the TARA, the 2018-2019 treatment students significantly underperformed their 2016-2017 and 2017-2018 comparison group in SBAC Claim 1 and Claim 2 & 4. This model was not run for Substudy 2 because the MKT and TARA were not collected from the comparison teachers.

Exhibit 48. HLM Analytic Model Results Examining the Impact of LLAMA on Substudy 1 SBAC Scores

	SBAC overall		Claim 1		Claim 2&4		Claim 3	
	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Intercept	2506.44	56.82	2518.54	64.73	2495.92	61.10	2492.05	75.98
Implementation Fidelity								
Category 3	21.83*	9.72	34.31**	10.89	51.52***	12.34	18.64	12.90
Category 4	-1.70	17.82	3.54	19.10	22.09	21.65	-15.35	22.62
TARA	5.79	3.92	5.11	4.57	4.81	4.29	5.97	5.36
MKT	14.34	24.32	20.50	26.99	31.29	25.47	21.15	31.68
Intervention effect	-10.59	7.27	-18.61*	8.28	-27.05**	9.38	-5.07	9.80

$^{**}p < .01$; $^{***}p < .001$.

Hypothesis 3. The treatment teachers who implement the LLAMA intervention with high fidelity will have a greater impact on student achievement than teachers who implement the LLAMA intervention with lower fidelity.

The research team hypothesized that treatment teachers who implement the LLAMA intervention with high fidelity will have a greater impact on student achievement than teachers who implement the LLAMA intervention with lower fidelity. To test this hypothesis the research team used an HLM model that included teacher implementation scores as a covariate to account for teacher differences in implementing the LLAMA intervention.

Analytic Sample

The analytic sample includes 9 treatment teachers for the SubStudy 1 and 7 treatment teachers for the SubStudy 2 (Exhibit 39). All treatment teachers have implemented the LLAMA intervention in their classroom with a fidelity score of 2, 3 or 4. The research team gave each LLAMA teacher an implementation category code in Year 5. Codes ranged from 1 to 4.

- **High Implementer:** A teacher was coded as a '**4'** if the data showed the teacher (a) engaged students in learning experiences targeting the learning of the 12 CPs, (b) included viable argumentation as a regular feature of instruction, and (c) included viable argumentation for generalizations frequently (i.e., at least twice a month).
- **Medium Implementer:** A teacher was coded as a '**3'** if the data showed the teacher (a) engaged students in learning experiences targeting the learning of the 12 CPs, (b) included viable argumentation sometimes in their instruction, and (c) included viable argumentation for generalizations sometimes.
- **Low Implementer:** A teacher was coded as a '**2'** if the data showed the teacher (a) engaged students in learning experiences targeting the learning of some of the 12 CPs, (b) included viable argumentation infrequently in their instruction, and (c) included viable argumentation for generalizations infrequently. **No Implementation:** A teacher was coded as a '**1'** if the data showed the teacher did not start the project or there was no evidence of the teacher implementing LLAMA in the classroom.

Analytic Plan

Descriptive statistics demonstrate that SubStudy 1 students taught by teachers with LLAMA implementation scores of 4 scored highest in SBAC overall scores and all sub-claims while students taught by teachers with implementation scores of 2 scored the lowest (Exhibit 49). For SubStudy 2, however, student scores increased by their teachers' implementation scores in SBAC Claim 1 and Claim 2 & 4 (Exhibit 50). HLM Model 4 was then conducted to investigate the relationship between student SBAC achievements and teacher LLAMA implementation for SubStudy 1 and SubStudy 2 (Exhibits 51 and 52). **The hypothesis was partically supported for SubStudy 1**. There is a significant relationship between teacher implementation scores of 3 and student SBAC overall scores and their scores in Claim 1 and Claim 2 & 4 for SubSutdy 1. **For SubStudy 2, hypothesis 3 was not supported.** LLAMA implementation category does not had a significant impact on student mathematics achievement measured by SBAC.

Exhibit 49. Descriptive Statistics:
Implementation Fidelity by Treatment Group in Substudy 1

Impl Category	Group	SBA overall		Claim 1		Claim 2&4		Claim 3	
		Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
1	Comparison	2580.09	106.10	2581.86	111.17	2571.38	125.12	2570.76	131.21
1	Treatment	-	-	-	-	-	-	-	-
2	Treatment	2568.04	94.54	2568.56	100.78	2551.37	123.83	2564.95	116.54
3	Treatment	2589.89	110.62	2583.19	118.70	2582.35	125.06	2574.51	140.06
4	Treatment	2558.52	57.12	2572.11	56.57	2536.11	87.59	2535.56	94.92

Exhibit 50. Descriptive Statistics:
Implementation Fidelity by Treatment Group in Substudy 2

Impl Score	Group	SBA overall		Claim 1		Claim 2&4		Claim 3	
		Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
1	Comparison	2537.57	109.55	2539.29	116.32	2519.94	129.59	2524.49	128.55
1	Treatment	-	-	-	-	-	-	-	-
2	Treatment	2560.31	79.11	2558.32	89.64	2530.33	117.93	2569.80	88.02
3	Treatment	2593.84	103.08	2580.41	108.67	2585.22	116.80	2579.02	134.37
4	Treatment	2564.68	53.91	2577.72	54.06	2546.36	82.21	2542.40	85.59

Exhibit 51. Model 4: LLAMA Implementation Fidelity HLM Results in Substudy 1

	SBA overall		Claim 1		Claim 2&4		Claim 3	
	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Intercept	2586.11	13.32	2579.84	15.34	2568.82	14.78	2569.85	17.16
Implementation Fidelity								
3	25.88**	9.70	28.63**	10.57	46.14***	11.96	13.15	12.51
4	18.83	34.48	36.32	36.84	16.86	39.52	-5.92	42.82
Intervention effect	-16.38*	7.79	-18.21*	8.27	-26.28**	9.36	-4.71	9.79

***$p < .001$.

Exhibit 52. HLM Model 4 Results Examining the Impact of LLAMA on Substudy 2 SBAC Scores

	SBAC overall		Claim 1		Claim 2&4		Claim 3	
	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Intercept	297.70	27.62	439.18	31.52	825.08	36.85	794.70	40.54
Baseline SBAC score	0.89***	0.01	0.83***	0.01	0.68***	0.01	0.69***	0.02
Implementation Fidelity								
Category 3	2.45	6.92	2.99	8.46	30.23**	11.37	-10.29	11.90
Category 4	25.82	13.27	33.08*	15.94	41.21	21.42	-12.22	22.41
Intervention effect	2.41	6.06	-1.97	7.31	-13.52	9.82	34.36***	10.27

*$p < .05$; **$p < .01$; ***$p < .001$.

Hypotheses 4. Teacher content knowledge and practice mediate the relationship between the LLAMA intervention and outcomes.

Teacher content knowledge and practice was estimated using their 2018 baseline MKT scores and TARA scores as covariates in HLM Model 5 for SubStudy 1. Exhibit 53 indicates that there is no statistically significant effect of teacher content knowledge and practice on student SBAC scores. **This hypothesis was not supported in SubStudy 1.** This hypothesis cannot be tested for Substudy 2 because MKT and TARA scores were not collected from the comparison teachers.

Exhibit 53. HLM Model 5 Results Examining the Impact of LLAMA on SBAC Scores in Substudy 1

	SBAC overall		Claim 1		Claim 2&4		Claim 3	
	Estimate	***SE***	**Estimate**	***SE***	**Estimate**	***SE***	**Estimate**	***SE***
Intercept	2505.70	55.41	2518.61	62.71	2500.98	61.25	2486.92	74.55
TARA	5.46	3.84	4.72	4.44	4.37	4.32	5.62	5.28
MKT	14.13	23.61	18.02	25.99	24.51	25.32	22.95	30.90
Intervention effect	0.44	4.65	-0.33	5.17	1.92	5.87	3.33	6.11

***$p < .001$.

Study 2: Student Argumentation Study-Original Study

RMC Research conducted an experimental research study using a pre-post design and post-only design to address Research Question 2, "Does the implementation of the LLAMA intervention change the treatment students' ability to construct viable arguments and critique the arguments of others?" The treatment group consists of students whose teachers were randomly assigned to start participating in the LLAMA intervention in Year 1 and the control group consists of students whose teachers were randomly assigned to start participation in the LLAMA intervention in Year 3. The independent variable is the LLAMA intervention and the dependent variable is student argumentation and reasoning skills. In the pre-post design, treatment and control students in Years 1, 2, and 3 completed the Student Argument and Reasoning Assessment at the beginning (pre) and end (post) of each school year. The pretest has 5 items: 4 that measure the ability to construct viable arguments, and 1 that assesses the ability to critique others' arguments. These items address mathematical content at the Grade 7 level to ensure the Grade 8 students have the mathematical knowledge necessary to adequately complete the assessment as a pretest at the beginning of their Grade 8 year (i.e., this approach ensures the assessment is measuring argumentation skills and not mathematical content knowledge). The posttest includes the same 5 items as the pretest and 4 additional items that address mathematical content that is taught to Grade 8 students during the school year. In the **pre-post design, the hypothesis** is that students in the treatment group will improve significantly more in argumentation skills than students in the control group (using the 5 items that are on both the pre and post). In the **post-only design, the hypothesis** is that students in the treatment group will score significantly higher on the posttest than students in the control group for the 4 items that are only included on the posttest.

Major Modifications. The data collection for the original study occurred as planned; however, based on the estimated number of scorers (4), targeted timeline to finish Year 1 SARAs (September 2018), and the number of assessments to score (approximately 3,000), RMC Research estimated each scorer would have to score 250 assessments a month to complete all Year 1 SARAs. Due to time and resource restraints, the LLAMA team decided to score a sample of the SARAs rather than all of the SARAs collected. The sample consisted of a subset of Year 2 SARAs which focused on 6 of the 34 Cohort 1 teachers (treatment) and 6 matched Cohort 2 teachers (comparison). Details regarding the sampling are included within this chapter.

SARA Executive Summary

Research Question 2, "Does the implementation of the LLAMA intervention change the treatment students' ability to construct viable arguments and critique the arguments of others?" **The first hypothesis is** that students in the treatment group will improve significantly more in argumentation skills than students in the control group (using the 5 items that are on both the pre and post). **The second hypothesis** is that students in the treatment group will score significantly higher on the posttest than students in the control group for the 4 items that are only included on the posttest. To test both hypotheses, the research team adhered to WWC guidelines as closely as possible in order to address potential issues related to attrition and baseline equivalence.

The first hypothesis was supported. To test the hypothesis and account for any group differences in the SARA pretest, a more nuanced analysis of MANOVA was used to estimate the treatment effect on student pre-post growth scores for Problems 1-5 (i.e, problems that are on both the pre and post assessment, items 6-9 are only on the post). Results show there was a statistically significant difference between the treatment group and control group on the growth scores of combined dependent variables of five SARA growth items, F (5, 317) =.868, p =.000.

Next, Multivariate analysis of covariance (MANCOVA) analysis was conducted to further account for teacher variance of LLAMA implementation fidelity. LLAMA treatment status was used as the independent variable. Student growth score was calculated as the score difference between the pre and post SARA assessments. Teacher implementation category (from 1-4 as described in previous chapters) was used as a covariate. Significant differences were observed between the treatment and control groups, F (5,316) = 5.809, p = .000. When controlling for teacher implementation fidelity categories, participation in LLAMA program was still positively and significantly associated with student argumentative skills.

The second hypothesis was supported. To test the second hypothesis and account for any group differences in the SARA pretest, a more nuanced analysis of MANOVA was used to estimate the treatment effect on items 6-9. The hypothesis was supported. The multivariate analysis of variance (MANOVA) was conducted to compare post test scores for Problems 6-9. The dependent scores are the four problem scores in Form B from the posttest. The independent variable is the study treatment status: treatment vs. control. The results of the MANOVA analysis show there were statistically significant differences between the treatment group and control group for Problems 6-9, F (4, 318) = 3.963, p = .004.

SARA Methods

Study Recruitment and Random Assignment

Study recruitment and random assignment is described within the chapter Study 1: Student Achievement Study.

What Works Clearinghouse Guidelines

What Works Clearinghouse utilizes three steps for reviewing RCTs and QEDs that assign individual subjects to the intervention or comparison condition: [12]

- **Step 1:** Assess the study design,
- **Step 2:** Assess sample attrition, and
- **Step 3:** Assess equivalence of the intervention and comparison groups at baseline (prior to the intervention).

Step 1: Assess the study design

"To be eligible for the WWC's highest rating for group design studies, *Meets WWC Group Design Standards Without Reservations*, the study must be an RCT with low levels of sample attrition. A QED or high-attrition RCT is eligible for the rating *Meets WWC Group Design Standards With Reservations* if it satisfies the WWC's baseline equivalence requirement that the analytic intervention and comparison

[12] Page 5; https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_handbook_v4.pdf

groups appear similar at baseline. A QED or high-attrition RCT that does not satisfy the baseline equivalence requirement receives the rating *Does Not Meet WWC Group Design Standards*."

This study is an RCT.

***Step 2:** Assess sample attrition*

The What Works Clearinghouse (Institute of Education Sciences, 2008) definition of differential attrition is:

- **Differential Attrition:** "Differential attrition refers to the situation in which the percentage of the original study sample retained in the follow-up data collection is substantially different for the intervention and the control groups. Severe differential attrition makes the results of a study suspect, because it may compromise the comparability of the study groups."

Differential attrition is 22% (74%-52%) for the intent to treat study groups (i.e., all teachers/students initially recruited for the study). For the intent to treat sample, 74% of the treatment group submitted pre/post data and 52% of the control group. The differential attrition for active teachers is smaller (i.e., those who have not dropped out of the project). For the active teachers in Year 1 the differential attrition is 17%: 89% of the treatment group submitted pre/post assessments and 72% of the control group. For Year 2, the differential attrition was 0%: 100% of both the treatment and control group submitted pre/post assessments.

The What Works Clearinghouse (Institute of Education Sciences, 2008) definition of overall attrition is:

- **Overall Attrition:** "Attrition is defined as a failure to measure the outcome variable on all the participants initially assigned to the intervention and control groups. High overall attrition generally makes the results of a study suspect, although there may be rare exceptions."

For the intent to treat study groups (i.e., all teachers/students initially recruited for the study), overall attrition was low for the treatment group with only 26% not submitting pre/post data but overall attrition was higher for the control group with 48% not submitting pre/post data.

Differential attrition cannot exceed 11% and this study is at 22%. While the differential attrition is not in an acceptable range, the overall attrition is within an acceptable range. In order to Meet WWC Group Design Standards With Reservations this study will need to show that the sample members who remain in the intervention and comparison groups in the analysis were similar on important characteristics at baseline.

***Step 3:** Assess equivalence of the intervention and comparison groups at baseline (prior to the intervention).*

SARA pretest results were compared using the independent samples *t* test to examine the baseline equivalence between treatment students taught by the LLAMA participating teachers and business-as-usual control students. Differences of SARA scores between all treatment and control students are presented in Exhibit 54.

Overall, treatment students outperformed the control students in Problem 1 and Problem 3 and scored lower than control students in Problems 2, 4, and 5. The group differences were statistically significant for Problem 2 and 4. Therefore, baseline equivalence was established for three out of five SARA items. In the MANCOVA analysis, student growth scores (score differences between the pre and post SARA assessments) were used to account for any group performance differences at the outset of the study.

Treatment and control students' baseline scores were not equivalent at baseline. The analysis will account for any group performance differences at the outset of the study.

Exhibit 54. Student Pretest SARA Scores by Treatment Group

	Treatment (*N* = 200)		Control (*N* = 123)		
SARA	Mean	*SD*	Mean	*SD*	Mean Difference
Problem 1	0.61	0.91	0.45	0.66	0.16
Problem 2	0.61	0.67	0.85	0.77	-0.24**
Problem 3	1.40	1.27	0.93	1.18	0.47**
Problem 4	0.46	0.92	0.56	1.05	-0.11
Problem 5	0.35	0.76	0.46	0.79	-0.11
Problem 6	-	-	-	-	-
Problem 7	-	-	-	-	-
Problem 8	-	-	-	-	-
Problem 9	-	-	-	-	-

Note. Statistically significant based on independent samples *t* test results;
*$p < .05$, **$p < .01$. Pretest data for Problems 6-9 were not available for analysis.

Instrument Development & Interrater Reliability

Target: *Use* ***Student Argument and Reasoning Assessment****, Version 1 (pretest) and Version 2 (posttest).* ***Status:*** *Met*

Target: *The research team will develop an* ***Argument and Reasoning Assessment Rubric****.*
Status*: Met*

The research team developed the Student Argument and Reasoning Assessment (SARA) to measure students' abilities to construct viable arguments and critique others' arguments. The SARA was originally developed and validated in the LAMP pilot study (NSF Award Number: 1317034). Items were developed by reviewing prior research on proof/proving (e.g. Healy & Hoyles, 2000; Knuth, 2002b), state assessments, and feedback from the external advisory board. The pretest assessment has 5 items: 4 items measure the ability to construct viable arguments, and 1 item assesses the ability to critique others' arguments. Specifically Item 1 was designed to elicit a direct argument. Item 2 was designed to elicit an indirect argument or a direct argument. Item 3 was designed to elicit a counterexample

argument, and Item 4 was designed to elicit an exhaustive argument. Item 5 was designed to assess students' ability to see the generalization in a specific example and recognize that the structure in the example applied to all cases. These items address mathematical content at the Grade 7 level to ensure the Grade 8 students have the mathematical knowledge necessary to adequately complete the assessment as a pretest at the beginning of their Grade 8 year (i.e., this ensures the assessment is measuring argumentation skills and not mathematical content knowledge). The posttest assessment includes the same 5 items as the pretest and 4 additional items that address mathematical content that is taught to Grade 8 students during the school year—at the onset of the school year the students would not have the content knowledge to respond to these items on a pretest.

Exhibit 55 shows the two types of ratings each SARA received during scoring. Total SARA scores range from 0-15; scores per item ranged from 0-3. For the interrater reliability training only the second rating, viable argumentation, was utilized.

Exhibit 55: SARA Ratings and Rating Scales

Rating Type	Rating Scale
Read Correctly: measures students' understanding of mathematical objects/definitions and of the format/structure/instructions of the task	0: No evidence of understanding 1: Some understanding 2: Demonstrates understanding
Viable Argumentation: measures students' demonstration of a viable argument	0: No elements of a viable argument 1: Limited elements of a viable argument 2: Elements of a viable argument 3: Viable argument

For the 5 items included on both the pretest and posttest version, LAMP established content validity through an expert panel and assessed interrater reliability using single rater Intraclass Correlation Coefficients (ICC) in SPSS, which suggested that raters moderately agreed upon results (Item 1 ICC = 0.47; Item 2 ICC = 0.48; Item 3 ICC = 0.92; Item 4 ICC = 0.45; Item 5 ICC = 0.56). During the LAMP project, the team refined the SARAs, developed a scoring rubric, and scored the SARAs. Developing the scoring rubric required a significant amount of time, which left little time for formal interrater reliability training. Additionally, the scorers had considerably diverse backgrounds, which resulted in enough of a gap in perspective that the ratings differed substantially. The research team believed that the interrater reliability was lower than expected, primarily due to a lack of time dedicated to training raters, rather than issues with the SARAs or rubric.

At the onset of the LLAMA project the team had 3 major goals pertaining to the SARAs:

1. Refine and revise the LAMP Scoring Rubric for the LLAMA project.
2. Ensure high interrater reliability among coders.
3. Score the LLAMA Student Argument and Reasoning Assessment (SARAs).

RMC Research worked with University of Idaho to establish and begin implementing a multi-stage plan to address these three SARA goals. What occurs during each stage is described in detail in the Longitudinal Learning of Viable Argument in Mathematics for Adolescents Remote Interrater Reliability

Training Manual [13]. Exhibit 56 shows the stages, which SARA was used in each stage, who comprised the scoring team, which item was scored, and the intraclass correlation for each item that was scored. As shown in Exhibit 56, the ICCs were low during the LAMP scoring but greatly improved over time.

To date, the project has met all three goals for this study. The team met Goal 1 by refining and revising the LAMP Scoring Rubric for the LLAMA project. The team has met Goal 2 by attaining high interrater reliability among coders (i.e., at least .70 ICC or higher). This is a substantial improvement from the LAMP scoring. The team met Goal 3 by scoring all the SARAs for this study.

Exhibit 56. Inter-Rater Reliability Estimates for Argument and Reasoning Student SARAs

Stage	Assess.	Scoring Team	Item Number on SARA								
			1	2	3	4	5	6	7	8	9
LAMP ICC	LAMP	LAMP	.47	.48	.92	.45	.56	NS	NS	NS	NS
Stage 1 ICC (July 2017)	LAMP	LLAMA	.91	.86	.84	.87	.74	.89	.93	.83	.80
Stage 3 ICC (January 2018)	LAMP	NAB	.84	.67	.90	.91	.74	.89	.91	.85	.85
Stage 2 Round 1 ICC (October 2017)	LLAMA	LLAMA	.57	.61	.95	.62	.47	NV	NV	NV	NV
Stage 2 Round 2 ICC (June 2018)	LLAMA	LLAMA	.90	.73	.92	.86	.84	.99	NV	NV	NV
Stage 2 Round 3 ICC (August 2018)	LLAMA	LLAMA	.74	.77	.92	.88	.86	.93	NV	NV	NV
Stage 2 Round 4 ICC (February 2019)	LLAMA	LLAMA	.82	NS	NS	NS	NS	NS	NS	NS	NS
Stage 2 Round 5 ICC (March 2019)	LLAMA	LLAMA	.88	NS	NS	NS	NS	NS	NS	NS	NS
Stage 2 Round 6 ICC (June 2019)	LLAMA	LLAMA	NS	.80	NS	NS	NS	NS	NS	NS	NS

Note. NS= Not Scored because calibration round focused on a specific item. NV= ICC calculated but not enough variance so the ICC is invalid. Stage 1 and Stage 3: Items 1–5: (n = 27), Items 6–9: (n = 14). Stage 2 Round 1: Items 1–5: (n = 60), Items 6–9: (n = 30). Stage 2 Round 2: Items 1–9: (n = 48). Stage 2 Round 3 n = 50. Stage 2 Round 4 n = 50. Stage 2 Round 5 n = 50. Stage 2 Round 6 n = 50. LAMP, Stage 1, and Stage 3 ICCs are based on scoring of the same LAMP SARAs. Stage 2 Rounds 1-3 are based on scoring of a sample of LLAMA Year 1 SARAs. Stage 2 Rounds 4-6 are based on scoring of a sample of LLAMA Year 2 SARAs. Items 6-9 (Version B LAMP SARAs) were not scored during the first LAMP scoring.

[13] Qureshi, C., Wang, X., Lewis, C., Yopp, D., & Hiebert Larson, J. (2019). Longitudinal Learning of Viable Argument in Mathematics for Adolescents Remote Interrater Reliability Training Manual. RMC Research Corporation and University of Idaho.

Data Collection

> ***Target:*** *Treatment and control students in Years 1, 2, and 3 will complete the* ***Student Argument and Reasoning Assessment Version 1*** *at the beginning (pre) and end (post) of each school year; they will complete* ***Version 2*** *at the end of each school year.* ***Status:*** *Met*

RMC Research prepared the materials for teachers and mailed them a pre administration packet in December 2016 for Year 1 and August 2017 for Year 2 and August 2018 in Year 3. The packet included a copy of the parent consent form, Student Assent Information Sheets, a Research Study Assent Form, copies of pre SARAs, and instructions for administering the SARA. Teachers read the Student Assent Information Sheet to the students, had the students sign the Research Study Assent Form, and had students complete a paper version of the SARA. Teachers mailed the completed Research Study Assent Forms and SARAs to RMC Research. This process was repeated in spring 2017, spring 2018, and spring 2019 for the post administration, except that assent forms were only administered to new students at the post administration.

Data Collection Completion

As shown in Exhibits 57–58, participation in this data collection activity was high for the active treatment and control teachers in Years 1-3 (i.e., 72%-100%), but was lower (48%-74%) using the intent-to-treat sample that includes all teachers recruited for the study.

Exhibit 57: Intent to Treat RCT Teachers Submitting Data

	Treatment		Control	
Time Period	**Teachers**	**Completion**	**Teachers**	**Completion**
Total recruited	**34**		**31**	
Submitted Year 1 pre	28	82%	24	77%
Submitted Year 1 post	26	76%	18	58%
Submitted Year 2 pre	25	74%	16	52%
Submitted Year 2 post	25	74%	16	52%
Submitted Year 3 pre	19	56%	15	48%
Submitted Year 3 post	19	56%	15	48%

Note. Although Year 3 data were collected as planned, Year 3 data were not included in this study.

Exhibit 58: Active RCT Teachers Submitting Data

	Treatment		Control	
Time Period	**Teachers**	**Completion**	**Teachers**	**Completion[a]**
Year 1				
Total active May 31, 2017	**28**		**25**	
Submitted pre	27	96%	25	100%
Submitted post	25	89%[a]	18	72%
Year 2				
Total active May 31, 2018	**25**		**16**	
Submitted pre	25	100%	16	100%
Submitted post	25	100%	16	100%
Year 3				
Total active May 31, 2019	**22**		**15**	
Submitted pre	19	86%	15	100%
Submitted post	19	86%	15	100%

Note. Although Year 3 data were collected as planned, Year 3 data were not included in this study.
[a]Five additional non-RCT teachers also submitted Year 2, six non-RCT teachers submitted Year 3 pre and post assessments, and 3 non-RCT teachers submitted Year 4 pre assessments; those are not included in this percentage.

Student Participants and Consent Information for RCT Classes

Exhibit 59 shows the total number of students participating in this study (student assented; parents did not withdraw consent) is 1,721 for Year 1: 1,032 in the treatment group and 689 in the control group. The total number of participating students for Year 2 was 1,521: 997 in the treatment group and 524 in the control group (Exhibit 60). Using a two-sample test for equality of proportions with continuity correction, there were no significant differences between cohorts for parent refusals for either Year 1 or Year 2, nor were there significant differences between cohorts for student refusals in Year 2; however, significantly more Cohort 1 students than Cohort 2 students refused in Year 1 (14% and 11%, respectively, $p = 0.016$). The parent and student consent process are described in the Student Achievement Study chapter. Exhibit 61 shows student participation for Year 3, the analyses on the consent process was not conducted for Year 3 since Year 3 data were not included in this study.

Exhibit 59: Student Participants and Consent Information for Year 1 RCT Classes that Submitted Data

Study Condition	Students[a]	Active Student Participants[b]	No. Parents Refused	% Students w/ Parents Refusal	Students Refused[a,c]	
Cohort 1 (treatment)	1,229	1,032	29	2%	177	14%
Cohort 2 (control)	782	689	13	2%	83	11%
Total	**2,011**	**1,721**	**42**	**2%**	**260**	**13%**

Note. One non-RCT teacher also administered student assessments to 30 students; they are not included in this table or the analysis.
[a]Unduplicated count. Two students (who both withdrew assent) were in 2 Cohort 1 teachers' classes.
[b]Active student participants include all students who assented and whose parents did not withdraw consent.
[c]Three hundred sixty three students (Cohort 1: 213; Cohort 2: 150) had teachers who did not send these students' assent forms to the research team; the research team assumes that the teachers followed the proper assent procedures but forgot to mail the assent forms to the research team. These students are counted as giving assent.

Exhibit 60: Student Participants and Consent Information for Year 2 RCT Classes that Submitted Data

Study Condition	Students[a]	Active Student Participants[b]	No. Parents Refused	% Students w/ Parents Refusal	Students Refused[a,c]	
Cohort 1 (treatment)	1,282	997	35	3%	264	21%
Cohort 2 (control)	679	524	20	3%	143	21%
Total	**1,961**	**1,521**	**55**	**3%**	**407**	**21%**

Note. Five non-RCT teachers also administered student assessments to 125 students; they are not included in this table or the analysis.
[a]Unduplicated count. Two students (who both withdrew assent) were in 2 Cohort 1 teachers' classes.
[b]Active student participants include all students who assented and whose parents did not withdraw consent.
[c]One hundred fifty-eight students (Cohort 1: 93 Cohort 2: 65) had teachers who did not send these students' assent forms to the research team. These students are counted as withdrawing assent.

Exhibit 61: Student Participants and Consent Information for Year 3 RCT Classes that Submitted Data

Study Condition	Students[a]	Active Student Participants[b]	No. Parents Refused	% Students w/ Parents Refusal	Students Refused[a]	
Cohort 1 (treatment)	1,193	877	23	2%	296	25%
Cohort 2 (control)	868	700	18	2%	157	18%
Total	**2,061**	**1,577**	**41**	**2%**	**453**	**22%**

Note. Six non-RCT teachers also administered student assessments to 329 students; they are not included in this table or the analysis.
[a]Unduplicated count.
[b]Active student participants include all students who assented and whose parents did not withdraw consent.

Assessment Completion for RCT Intent to Treat

Largely due to 10 RCT teachers dropping from the project prior to pretest assessment administration, the SARA completion rates for an intent-to-treat model are low (Exhibits 62-64): only 61% of all possible students completed a pretest and 51% completed a posttest in Year 1; only 45% of all possible students completed a pretest and 39% completed a posttest in Year 2; and only 37% of all possible students completed a pretest and 33% completed a posttest in Year 3.

Exhibit 62: Student Argument and Reasoning Assessment Completion: Year 1 RCT Intent-to-Treat Completion Rates

Study Condition	Students[a]	Pretests	Completion Rate	Posttests	Completion Rate	Matching[b]	Completion Rate
Cohort 1 (treatment)	1,461	903	62%	782	54%	629	52%
Cohort 2 (control)	1,053	635	60%	488	46%	423	40%
Total	**2,514**	**1,538**	**61%**	**1,270**	**51%**	**1,052**	**42%**

Note. One non-RCT teacher also administered student assessments to 30 students; they are not included in this table or the analysis. These student counts include nonconsenting students. The combined average class size of Year 1 LLAMA classes, used for mean imputation to arrive at the completion rate, was ~39 (38.7) students.
[a]Unduplicated count. Two students (who both withdrew assent) were in 2 Cohort 1 teachers' classes. Mean imputation was used to estimate class size for teachers who did not submit rosters or class counts (Cohort 1: 6 teachers; Cohort 2: 7 teachers; mean classroom of 38.7 students).
[b]Matching refers to students who completed both a pre- and a post-assessment.

Exhibit 63: Student Argument and Reasoning Assessment Completion: Year 2 RCT Intent-to-Treat Completion Rates

Study Condition	Students[a]	Pretests	Completion Rate	Posttests	Completion Rate	Matching[b]	Completion Rate
Cohort 1 (treatment)	1,704	920	54%	772	45%	753	43%
Cohort 2 (control)	1,383	462	33%	426	31%	269	19%
Total	**3,087**	**1,382**	**45%**	**1,198**	**39%**	**1,022**	**33%**

Note. Five non-RCT teachers also administered student assessments to 125 students; they are not included in this table or the analysis. These student counts include nonconsenting students. The combined average class size of Year 2 LLAMA classes, used for mean imputation to arrive at the completion rate, was ~47 (46.9) students.
[a]Unduplicated count. Mean imputation was used to estimate class size for teachers who did not submit rosters or class counts (Cohort 1: 9 teachers; Cohort 2: 15 teachers; mean classroom of 46.9 students).
[b]Matching refers to students who completed both a pre- and a post-assessment.

Exhibit 64: Student Argument and Reasoning Assessment Completion: Year 3 RCT Intent-to-Treat Completion Rates

Study Condition	Students[a]	Pretests	Completion Rate	Posttests	Completion Rate	Matching[b]	Completion Rate
Cohort 1 (treatment)	2,102	825	39%	739	35%	699	33%
Cohort 2 (control)	1,838	647	35%	542	29%	523	28%
Total	**3,940**	**1,472**	**37%**	**1,281**	**33%**	**1,222**	**31%**

Note. Six non-RCT teachers also administered student assessments to 329 students; they are not included in this table or the analysis. These student counts include nonconsenting students. The combined average class size of Year 3 LLAMA classes, used for mean imputation to arrive at the completion rate, was ~61 (60.6) students.
[a]Unduplicated count. Mean imputation was used to estimate class size for teachers who did not submit rosters or class counts (Cohort 1: 15 teachers; Cohort 2: 16 teachers; mean classroom of 60.6 students).
[b]Matching refers to students who completed both a pre- and a post-assessment.

Assessment Completion for Students of Active Teachers

Exhibit 65 shows the student SARA completion rates for the LLAMA teachers who were active as of May 31, 2017 in Year 1; Exhibit 66 shows the student SARA completion rates for the LLAMA teachers who were active as of May 31, 2018; and Exhibit 67 shows the student SARA completion rates for the LLAMA teachers who were active as of May 31, 2019. The completion rates are much higher for assenting students of active teachers: 78% of active students completed a pretest and 64% completed a posttest in Year 1; 92% of active students completed a pretest and 79% completed a posttest in Year 2; and 93% of active students completed a pretest and 81% completed a posttest in Year 3.

Exhibit 65: Student Argument and Reasoning Assessment Completion: Year 1 RCT Active Student Participant Completion Rates

Study Condition	Active Student Participants[a]	Pretests	Completion Rate	Posttests	Completion Rate	Matching[b]	Completion Rate
Cohort 1 (treatment)	1,201	885	74%	760	63%	629	52%
Cohort 2 (control)	749	635	85%	488	65%	453	60%
Total	**1,950**	**1,520**	**78%**	**1,248**	**64%**	**1,082**	**55%**

Note. One non-RCT teachers also administered student assessments to 30 students; they are not included in this table or the analysis. One Cohort 1 teacher who dropped from the project submitted pre and post assessments from 26 students; those assessments are not included in this table but will be included in the analysis.
[a]Active student participants include all students who assented and whose parents did not withdraw consent.
[b]Matching refers to students who completed both a pre- and a post-assessment.

Exhibit 66: Student Argument and Reasoning Assessment Completion: Year 2 RCT Active Student Participant Completion Rates

Study Condition	Active Student Participants[a]	Pretests	Completion Rate	Posttests	Completion Rate	Matching[b]	Completion Rate
Cohort 1 (treatment)	984	920	93%	772	78%	753	77%
Cohort 2 (control)	524	462	88%	426	81%	394	75%
Total	**1,508**	**1,382**	**92%**	**1,198**	**79%**	**1,147**	**76%**

Note. Five non-RCT teachers also administered student assessments to 125 students; they are not included in this table or the analysis.
[a]Active student participants include all students who assented and whose parents did not withdraw consent.
[b]Matching refers to students who completed both a pre- and a post-assessment.

Exhibit 67: Student Argument and Reasoning Assessment Completion: Year 3 RCT Active Student Participant Completion Rates

Study Condition	Active Student Participants[a]	Pretests	Completion Rate	Posttests	Completion Rate	Matching[b]	Completion Rate
Cohort 1 (treatment)	877	825	94%	739	84%	699	80%
Cohort 2 (control)	700	647	92%	542	77%	523	75%
Total	**1,577**	**1,472**	**93%**	**1,281**	**81%**	**1,222**	**77%**

Note. Six non-RCT teachers also administered student assessments to 329 students; they are not included in this table or the analysis.
[a]Active student participants include all students who assented and whose parents did not withdraw consent.
[b]Matching refers to students who completed both a pre- and a post-assessment.

Data Collection Decision Rules

In several instances either the teacher or the student deviated from the instructions. Exhibit 68 shows how each of these cases were handled in terms of counting or excluding students from the completion rates and analytic sample.

Exhibit 68: Student Argument and Reasoning Assessment Data Collection Decision Rules

Data Collection Issue	How This Case was Handled
The research team received a parent consent form for a student who was not on the teachers' rosters.	Because the parent consent form did not note the students' teachers, the research team assumed that the student was not in a LLAMA teacher's classroom and excluded that student from the count of students.
On the class roster, a teacher marked that a parent had withdrawn consent for their student. The research team did not receive a parent consent form.	The research team counted the student as having a parent who refused and will remove the assessment from the analytic sample.
A teacher sent an assessment, but no assent form, for one or more students in the class.	After making every effort with the help of the teacher to recover assent forms for these students, the research

	team will remove from the analytic sample the assessments from students who are missing assent forms.
One teacher sent a complete class list with assent noted for each student.	The research team will verify that each student who sent an assessment is included on the class list and will assume that proper assent procedure was followed, excepting the teacher mailing the hard copies. The students noted on the class list as granting assent will be included in the analytic sample.
One teacher sent an incomplete class list that listed names only of students who withdrew assent for one or more research activities.	The research team verified that the teacher administered the assent forms but did not mail them. The students in this class who are not noted as withdrawing assent will be assumed to have granted assent and their assessments will be included in the analytic sample.
Two students are in more than 1 teacher's class.	Those students are only counted once in the completion table and will be randomly assigned one of the 2 teachers in the analytic sample.
One student had 2 assent forms: 1 with nothing checked, and another with everything checked.	The research team counted the student as withdrawing assent for all items and will remove the assessment from the analytic sample.
The student completed 2 assent forms: 1 with the pretest and 1 with the posttest. The pre- grants assent and the post- withdraws consent.	The research team counted the student as withdrawing assent for both the pre- and posttests and will remove the assessments from the analytic sample.
The teacher noted on the assessment that the student did not complete the assessment or missed the second day of testing.	The research team counted the assessment as complete for calculating the completion rates but will remove the assessment from the analytic sample.
The teacher wrote the students' names on the assent forms for which no boxes were checked.	Because the students with boxes checked on the assent form wrote their own names, the research team assumed that the teacher instructed the students to only complete the assent form if they withdraw assent. These students (for whom the teacher wrote their names) are counted as giving assent. Their assessments will be included in the analytics sample.
For 1 student the name on the assent form was crossed out and the teacher wrote the name. The boxes were scribbled over.	The research team counted the student as withdrawing assent for all items and will remove the assessment from the analytic sample.
One teacher removed student names and coded the assessments for the pretests. Students wrote their names on the posttests, and the teacher also wrote a code on the posttests. No assent forms were sent. For 2 students whose parents refused, the codes are known, but do not match from pre to post.	The research team assumed that the teacher followed the proper assent procedures but forgot to mail the assent forms to the research team. The students are counted as giving assent, unless we received a parent refusal. All assessments for this teacher will be removed from the analytic sample for pre-post paired analysis.
One student was marked absent for the pretest, and the teacher did not send an assent form with the posttest.	The research team assumes this student did not grant assent and will remove the posttest from the analytic sample.
The teacher neither sent assent forms nor a class list.	The research team cannot be sure that the teacher upheld the assent process. These classes' assessments will be removed from the analytic sample.

Sampling

> *Based on the estimated number of scorers (4), targeted timeline to finish Year 1 SARAs (September 2018), and the number of assessments to score (approximately 3,000), RMC Research estimated each scorer would have to score 250 assessments a month to complete all Year 1 SARAs. Due to time and resource restraints, the LLAMA team decided to score a subset of Year 2 SARAs as part of a substudy which focused on 6 of the 34 Cohort 1 teachers (treatment) teachers and 6 matched Cohort 2 teachers (comparison).*

The 6 Cohort 1 treatment teachers were chosen based on their high level of LLAMA implementation and the increased likelihood of detecting a shift in their students' assessment scores from pre to post. As shown in Exhibit 69, the 6 high implementers resembled the overall treatment group with a few exceptions.

- A larger percentage of high implementers were from smaller schools with 200 students or less (33% vs. 12%).
- There were no high implementers from urban school settings compared to 21% of the overall treatment group.
- Though 44% of the treatment group were from Washington, only 17% of the Washington participants were high implementers.

Exhibit 69: Treatment Group Compared to High Implementers

	Treatment (*n* = 34)		Treatment High Implements (*n* = 6)	
Variable	***n***	**%**	***n***	**%**
State				
WA	15	44%	1	17%
ID	15	44%	4	67%
MT	4	12%	1	17%
School Setting				
Rural	22	65%	5	83%
Urban	7	21%	0	0%
Suburban	5	15%	1	17%
Grade Span				
5-8	4	12%	0	0%
6-8	17	50%	3	50%
7-8	4	12%	1	17%
7-9	3	9%	0	0%
7-12	1	3%	0	0%
6-12	2	6%	1	17%
PK-8	2	6%	1	17%

	Treatment (*n* = 34)		Treatment High Implements (*n* = 6)	
Variable	***n***	**%**	***n***	**%**
KG-8	1	3%	0	0%
Title I				
Yes	25	74%	5	83%
No	9	26%	1	17%
Student Enrolled				
0-200	4	12%	2	33%
201-400	9	26%	2	33%
401-600	8	24%	1	17%
601-800	7	21%	1	17%
801-1,016	6	18%	0	0%
Race				
0%-20% White	1	3%	0	0%
21%-40% White	6	18%	1	17%
41%-60% White	6	18%	1	17%
61%-80% White	8	24%	2	33%
81%-100% White	13	38%	2	33%
0%-20% Hispanic	20	59%	3	50%
21%-40% Hispanic	5	15%	1	17%
41%-60% Hispanic	6	18%	2	33%
61%-80% Hispanic	2	6%	0	0%
81%-100% Hispanic	1	3%	0	0%

To identify an appropriate match for each Cohort 1 teacher, a sampling frame was prepared consisting of all comparison teachers (*n* = 36). First, the research team excluded from the sampling frame comparison teachers who were non RCT (*n* = 5), did not teach Grade 8 during Year 1 or Year 2 (*n* = 2), or did not submit both pre and post SARAs in Year 1 (*n* = 21). A total of 22 unique teachers were excluded from the sampling frame, leaving 15 comparison teachers that could be matched with the treatment teachers. From this list of 15 teachers, each treatment teacher was matched with a comparison teacher based on the following school level variables. All data to create these variables was obtained from the Institute of Education Science's National Center for Education Statistics website. To ensure that each school level variable was weighted equally, each variable was constructed on a scale of 0 to 1..

- **State (Washington, Idaho, and Montana)—**Coded as 0 exact match or 1 no match.
- **School setting (rural, suburban, urban)—**Coded as Rural = 0, Suburban = .5, and Urban = 1.
- **Grade span (Grades 5-8, Grades 6-8, Grades 7-8, Grades 6-12, Grades KG-8, Grades PK-8)—**Coded as 0 exact match or 1 no match.
- **Title I—**Coded as 0 exact match or 1 no match.

- **Enrollment (Total number of students enrolled)—**Total number of students enrolled in school.
- **Hispanic**—% Hispanic students.
- **White—**% White students.

Using an exact matching approach,[14] a proximity score was then calculated for each comparison teacher using the variables noted above in the following formula.

Proximity score = (State) + Absolute Value(School Setting TX-School Setting CT) + (Grade Span)+(Title 1) + Absolute Value((Enrollment TX – Enrollment CT)/Enrollment TX) + Absolute Value(Hispanic TX – Hispanic CT) + Absolute Value(White TX – White CT).

The lower the value, the closer the match between the treatment and comparison teacher. Using this approach, a comparison teacher was selected for each treatment teacher. If a comparison teacher was matched with more than one treatment teacher, the treatment teachers' proximity scores for the next matches were compared and whomever' s next match had the larger proximity score kept the comparison teacher and the other treatment teacher was paired with their next match. Exhibit 70 shows that the treatment and comparison group school characteristics were similar.

Exhibit 70: High Implementers & Comparison

	High Implementers (*n* = 6)		Comparison (*n* = 6)	
Variable	***n***	***%***	***n***	***%***
State				
WA	1	17%	2	33%
ID	4	67%	3	50%
MT	1	17%	1	17%
School Setting				
Rural	5	83%	5	83%
Urban	0	0%	0	13%
Suburban	1	17%	1	17%
Grade Span				
5-8	0	0%	0	0%
6-8	3	50%	3	50%
7-8	1	17%	1	17%
7-9	0	0%	0	0%
7-12	0	0%	0	0%
6-12	1	17%	1	17%
PK-8	1	17%	1	17%
KG-8	0	0%	0	0%

[14] College board report

	High Implementers (*n* = 6)		Comparison (*n* = 6)	
Variable	***n***	***%***	***n***	***%***
Title I				
Yes	5	83%	4	67%
No	1	17%	2	33%
Student Enrolled				
0-200	2	33%	2	33%
201-400	2	33%	2	33%
401-600	1	17%	0	0%
601-800	1	17%	1	17%
801-1,016	0	0%	1	17%
Race				
0%-20% White	0	0%	0	0%
21%-40% White	1	17%	0	0%
41%-60% White	1	17%	1	17%
61%-80% White	2	33%	1	17%
81%-100% White	2	33%	4	67%
0%-20% Hispanic	3	50%	4	67%
21%-40% Hispanic	1	17%	2	33%
41%-60% Hispanic	2	33%	0	0%
61%-80% Hispanic	0	0%	0	0%
81%-100% Hispanic	0	0%	0	0%

Scoring for Substudy 1

The SARA assessments were blindly scored by University of Idaho in Year 3. As of June 30, 2019, all Year 2 matching pre/post SARAs of Substudy 1 teachers (*n* = 646) have been scored.

Findings

This chapter includes the results from the SARA substudy based on SARA responses collected from 200 students taught by 6 treatment teachers and 123 control students taught by 6 control teachers in study Year 2.

Pre-Post Comparisons: T Tests

The first hypothesis is that students in the treatment group will improve significantly more in argumentation skills than students in the control group. First, the research team conducted a naïve analysis (i.e., not correcting for baseline differences or controlling for other variables) using paired samples *t* test as a repeated measures test[15] to analyze SARA score changes over time. Exhibit 71

[15] The repeated measures analysis included only students with multiple years data. Therefore, the sample sizes in the repeated measures analysis is different from those in the group comparison analysis.

presents the overall SARA score changes in the control group. The results indicate that control students' SARA scores increased for all problems, yet only three problems, Problems 2, 3, and 5, demonstrated statistically significant pre-post improvement. As compared to the control students, the treatment group's SARA scores significantly increased for all five problems with a bigger magnitude of improvement (see Exhibit 72 and 73).

Exhibit 71. Students of Control Teachers' SARA Performance Over Time

	Pretest (*N* = 123)		Posttest (*N* = 123)		Pre-Post
SARA	Mean	SD	Mean	SD	Mean Difference
Problem 1	0.45	0.66	0.56	0.81	0.11
Problem 2	0.85	0.77	1.15	0.99	0.3***
Problem 3	0.93	1.18	1.28	1.26	0.35**
Problem 4	0.56	1.05	0.72	1.08	0.16
Problem 5	0.46	0.79	0.73	0.91	0.27**

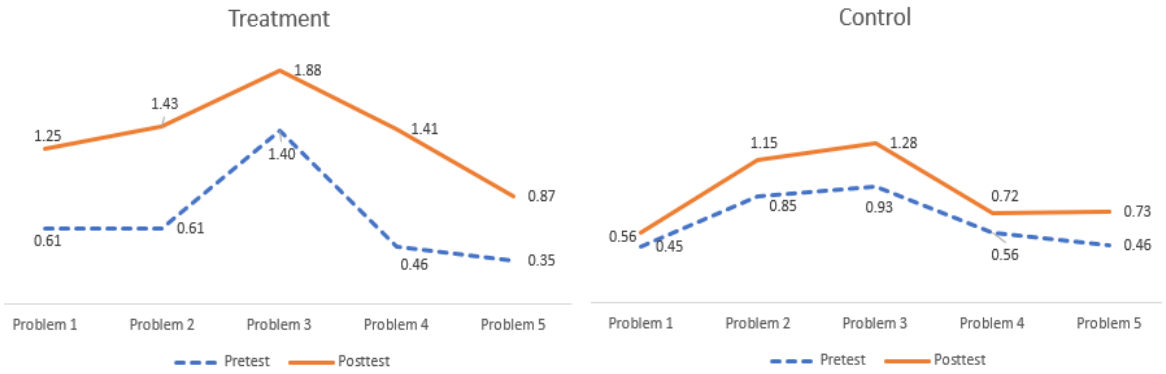
Note. Statistically significant based on paired samples *t* test results; **$p < .01$, ***$p < .001$.

Exhibit 72. Treatment Group SARA Performance Over Time

	Pretest (*N* = 25)		Posttest (*N* = 25)		Pre-Post
SARA	Mean	SD	Mean	SD	Mean Difference
Problem 1	0.61	0.91	1.25	1.23	0.64***
Problem 2	0.61	0.67	1.43	1.11	0.82***
Problem 3	1.40	1.27	1.88	1.24	0.48***
Problem 4	0.46	0.92	1.41	1.30	0.95***
Problem 5	0.35	0.76	0.87	1.06	0.52***

Note. Statistically significant based on paired samples *t* test results; ***$p < .001$.

Exhibit 73. SARA Score Changes in Study Groups



MANCOVA

To test the hypothesis and account for any group differences in the SARA pretest, a more nuanced analysis of MANOVA was used to estimate the treatment effect on student pre-post growth scores for Problems 1-5 (i.e, problems that are on both the pre and post assessment, items 6-9 are only on the post). The hypothesis was supported. Results show there was a statistically significant difference between the treatment group and control group on the growth scores of combined dependent variables of five SARA growth items, F (5, 317) =.868, p =.000.

Implementation Fidelity

Next, Multivariate analysis of covariance (MANCOVA) analysis was conducted to further account for teacher variance of LLAMA implementation fidelity. LLAMA treatment status was used as the independent variable. Student growth score was calculated as the score difference between the pre and post SARA assessments. Teacher implementation category (from 1-4 as described in previous chapters) was used as a covariate. Significant differences were observed between the treatment and control groups, F (5,316) = 5.809, p = .000. When controlling for teacher implementation fidelity categories, participation in LLAMA program was still positively and significantly associated with student argumentative skills.

Comparison Post Only Items (Items 6-9)

The second hypothesis is that that students in the treatment group will score significantly higher on items 6-9 than students in the control group. Items 6-9 were only included on the post assessments. First, the research team conducted a naïve analysis (i.e., not correcting for baseline differences or controlling for other variables of student post SARA by conducting independent samples t test). Exhibit 74 shows that treatment students outperformed control students for all four post only items. The treatment students scored significantly higher on each item.

Exhibit 74. Student Posttest SARA Scores by Treatment Group

SARA	Treatment (*N* = 200) Mean	*SD*	Control (*N* = 123) Mean	*SD*	Mean Difference
Problem 6	0.64	1.19	0.32	0.86	0.32**
Problem 7	0.11	0.48	0.01	0.09	0.1**
Problem 8	0.37	0.79	0.15	0.36	0.22**
Problem 9	0.39	0.84	0.20	0.52	0.19*

Note. Statistically significant based on independent samples *t* test results; *$p < .05$, **$p < .01$, ***$p < .001$.

MANOVA

To test the second hypothesis and account for any group differences in the SARA pretest, a more nuanced analysis of MANOVA was used to estimate the treatment effect on items 6-9. The hypothesis was supported. The multivariate analysis of variance (MANOVA) was conducted to compare post test scores for Problems 6-9. The dependent scores are the four problem scores in Form B from the posttest. The independent variable is the study treatment status: treatment vs. control. The results of the MANOVA analysis show there were statistically significant differences between the treatment group and control group for Problems 6-9, F (4, 318) = 3.963, p = .004.

Item Level Analyses

Because the MANOVA results suggest significant differences between the posttest SARA performance in the treatment and control groups, pairwise comparison of the individual item scores were conducted using the analysis of variance (ANOVA) tests. A Bonferroni correction was applied to reduce the Type I error (i.e., "false positives"). Results of the pairwise analyses are shown below in Exhibit 75. Results suggest that students of treatment teachers outperformed students of control teachers on all items 6-9. At this item level MANOVA analyses, this second hypothesis is supported.

Exhibit 75. Pairwise Comparisons for Treatment and Control Posttest SARA Scores: Form B

Posttest Problem	Difference	Lower Bound[a]	Upper Bound[a]	*p*	Sig.
Problem 6	0.318	0.076	0.560	0.010	*
Problem 7	0.102	0.016	0.188	0.020	*
Problem 8	0.224	0.075	0.373	0.003	**
Problem 9	0.195	0.029	0.360	0.021	*

Note. [a]Adjusted by Bonferroni correction.
* $p < 0.05$, ** $p < 0.01$.

Study 2: Student Argumentation Study—Substudy 1 of Active Cohort 2 Teachers

The original SARA study produced promising findings, so the LLAMA team decided to conduct additional studies to examine student argumentation in greater depth. In Year 3, during the summer 2019, the LLAMA team decided to conduct Substudy 1 with the 12 Cohort 2 teachers that were active through the end of Year 4. Substudy 1 provides 2 years of data prior to LLAMA implementation (Years 1 and 2) and one year of data after LLAMA implementation (Year 3) from 12 highly engaged teachers.

Study Recruitment

Study recruitment is described within the chapter **Study 2: Student Argumentation Study**.

Instrument Development and Interrater Reliability

Study recruitment is described within the chapter **Study 2: Student Argumentation Study**.

Data Collection

Exhibit 76 shows the number of pre and post sets of matching SARAS per teacher per year for this study.

Exhibit 76: Number of Matching Sets of Pre and Post SARAs Per Teacher for Substudy 1 of Active Cohort 2 Teachers

Teacher	Year 1	Year 2	Year 3	Total
Teacher 1	21	30	23	74
Teacher 2	14	8	13	35
Teacher 3	40	35	27	102
Teacher 4	42	25	40	107
Teacher 5	3	12	8	23
Teacher 6	37	31	73	141
Teacher 7	41	45	34	120
Teacher 8	22	26	33	81
Teacher 9	0	9	34	44
Teacher 10	0	19	64	85
Teacher 11	28	41	52	121
Teacher 12	26	44	44	114
Total	274	325	445	1,047

Sampling

Based on the number of SARAs scored to date, the estimated number of scorers (4), targeted timeline to finish scoring, and the number of assessments to score, the research team developed a sampling plan for this study that would result in a feasible number of SARAs to score.

The research team decided to score all matching pre/post SARAs of active Cohort 2 teachers for Year 3 and a random sample of half of matching pre/post SARAs from Years 1 and 2. This would ensure the research team has some SARAs from each year but the total pool of data prior to LLAMA beginning for the Cohort 2 teachers/students data will be approximately the same size as the Year 3 data. Exhibit 77 shows the number of pre and post sets of matching SARAS per teacher per year sampled for this study.

Exhibit 77: Number of Matching sets of Pre and Post SARAs Per Teacher for Substudy 1 of Active Cohort 2 Teachers

Teacher	Year 1	Year 2	Year 3	Total
Teacher 1	11	15	23	**49**
Teacher 2	7	4	13	**24**
Teacher 3	19	17	27	**63**
Teacher 4	19	12	40	**71**
Teacher 5	0	18	8	**26**
Teacher 6	17	15	73	**105**
Teacher 7	18	22	34	**74**
Teacher 8	9	13	33	**55**
Teacher 9	0	5	34	**39**
Teacher 10	0	9	64	**73**
Teacher 11	14	19	52	**85**
Teacher 12	21	44	44	109
Total	**135**	**193**	**445**	**773**

Scoring

The LLAMA team completed the scoring of these SARAs by September 1, 2020.

Findings

This chapter includes the results from SARA responses collected from 445 students taught by 12 treatment teachers in Study Year 3 and 328 students taught by the same group of teachers as a control group during Year 1 and Year 2. Detailed data tables are in Appendix A.

Pre-Post Distributions of Individual Item Scores

Ideally, a multivariate test such as MANOVA would be used to assess overall pre-post differences between the treatment and the control groups. Although MANOVA is generally robust to violations in the normality assumption with large sample sizes, students' scores for most individual items are strongly skewed to the right (i.e., 50–80% of the students scored a "0", see Exhibit 78) on both the pre-test and the post-test—resulting in fewer than 20 students scoring a "2" or a "3"—which, in addition to a challenge of unequal variance among groups, precludes using MANOVA in this analysis.

Exhibit 78. Pretest and Posttest Item Score Distributions by Group

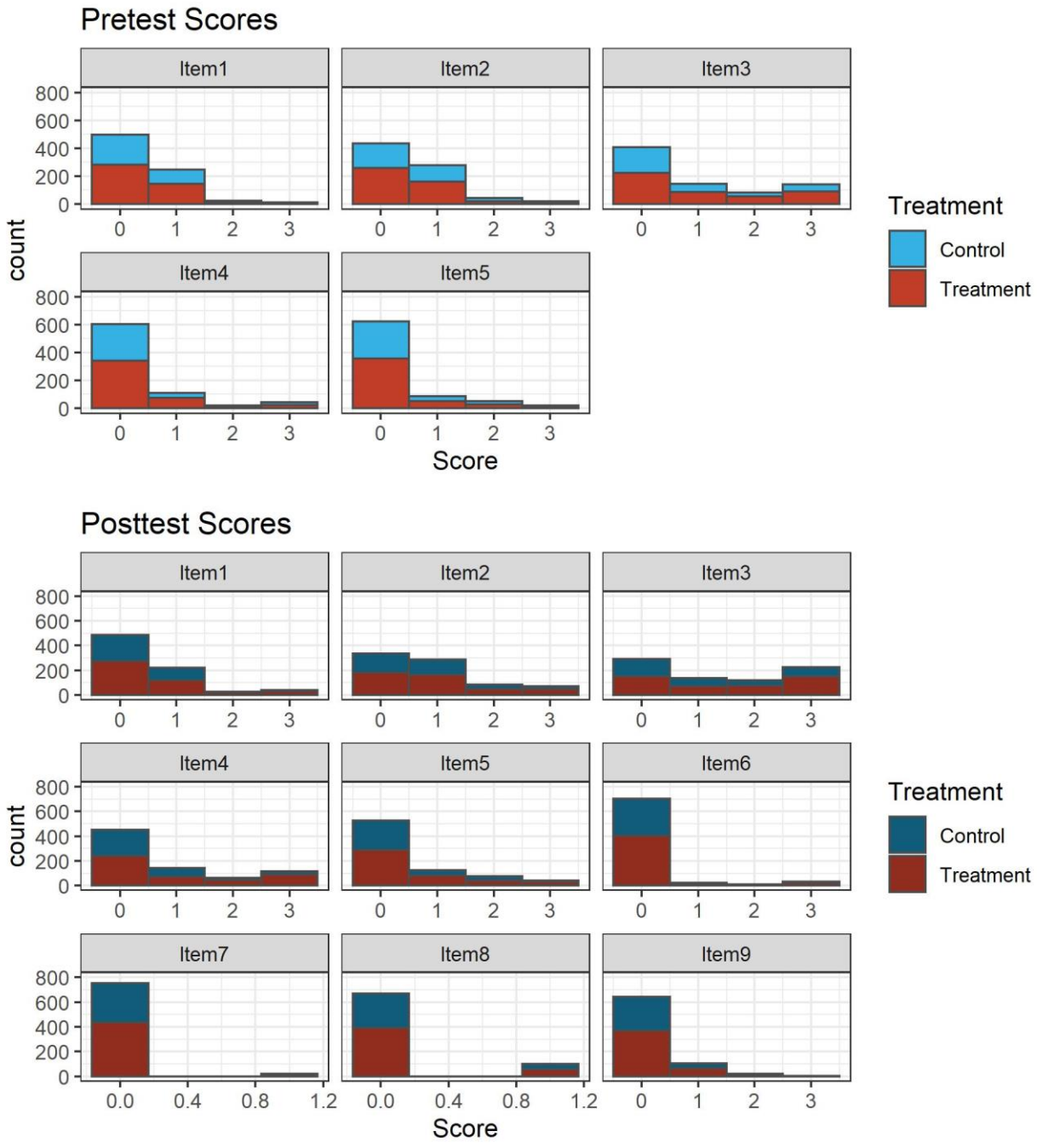
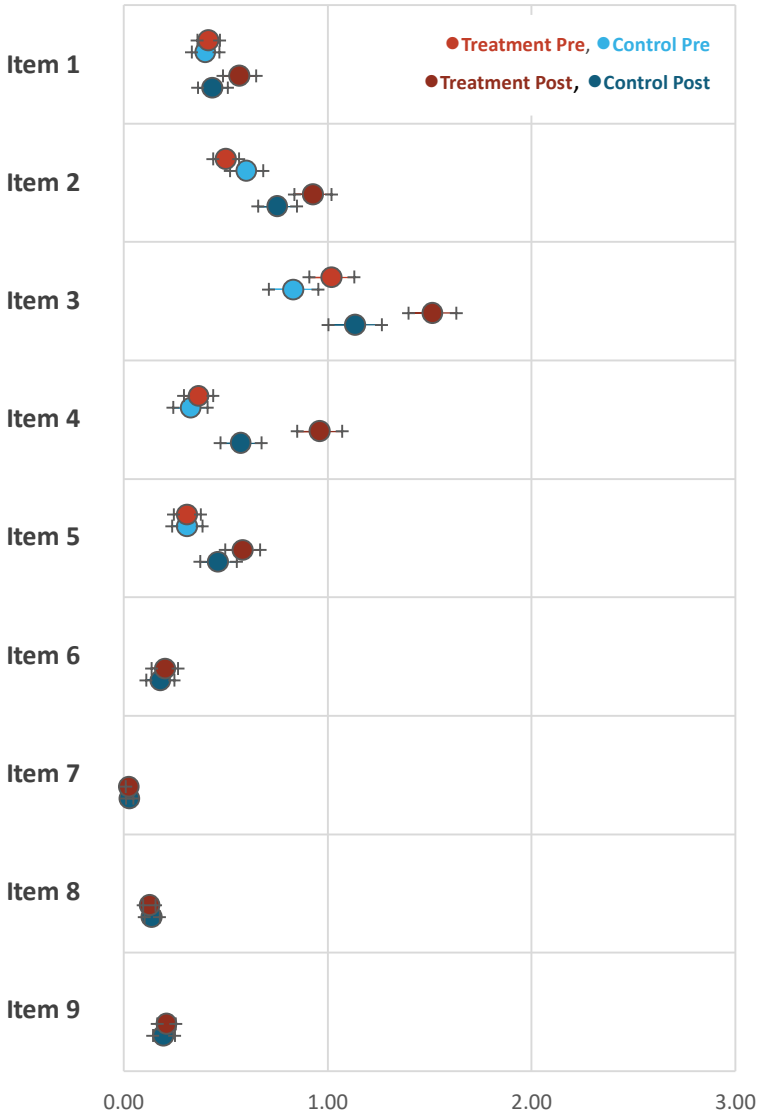


Exhibit 79 plots the means and confidence intervals for each item on the pretest and posttest for each group (treatment and control). Descriptively, Items 2, 3, and 4 show the most variation in scores, both between the groups as well as between the pre- and post-scores.

Exhibit 79. Pretest and Posttest Item Means by Group



Internal Consistency of Pretest Items

Using Cronbach's alpha[16] to test the internal consistency of pretest items provides weak evidence for combining individual items into a single scale score: the maximum value of alpha was 0.50 when Item 1 was excluded from the scale (Items 1–5 on the pretest only, using data from both the treatment and control students), suggesting that an analysis of a calculated "overall score" for each student may also not be a valid approach in analyzing student assessments.

[16]Cronbach's alpha is a measure of internal consistency. Values range from 0 to 1 with higher values indicating stronger internal consistency. An alpha value greater than 0.80 is the recommended threshold to analyze the items as a scale score.

Pre-Post Comparisons: Items 1–5

Given the limitations listed above, the research team conducted an analysis of each item separately. The function **misty::multilevel.icc** in R was used to determine whether or not there was a strong enough intraclass correlation coefficient (ICC) to merit including teacher as a random variable in the model. The ICC was extremely weak within groups, indicating that a random effect may not be necessary for a linear model. Therefore, to test the hypothesis that students in the treatment group improved significantly more in argumentation skills than students in the control group, a simple linear regression was applied for items common to both the pre- and posttest (Items 1–5) comparing posttest scores (the dependent variable) using treatment group as an independent variable and pretest score as a covariate. Differences were considered statistically significant when treatment group was a significant predictor ($p < 0.05$).

As shown below in Exhibit 80 and consistent with the plotted means and confidence intervals in Exhibit 79, students from the treatment group performed significantly better than students in the control group for Items 1–4, with the strongest evidence of a between-group difference for Item 3 and Item 4.

Exhibit 80. Pretest and Posttest Item Score Comparisons by Group

	Control Group				Treatment Group				
	Pretest		Posttest		Pretest		Posttest		
Item	*M*	*sd*	*M*	*sd*	*M*	*sd*	*M*	*sd*	*p*
Item 1	0.40	0.622	0.44	0.683	0.42	0.604	0.57	0.866	0.022*
Item 2	0.60	0.752	0.75	0.868	0.50	0.680	0.93	0.981	0.001**
Item 3	0.83	1.125	1.13	1.200	1.02	1.190	1.51	1.266	<0.001***
Item 4	0.33	0.778	0.57	0.932	0.37	0.767	0.96	1.200	<0.001***
Item 5	0.31	0.700	0.46	0.827	0.31	0.706	0.58	0.925	0.052

Note. Control Group: $n = 328$. Treatment Group: $n = 445$. *$p < 0.05$. **$p < 0.01$. ***$p < 0.001$.

Considerations for Potential Next Steps

Student scores are coded as ordinal categorical data, analyzed as continuous variables in this report. While treating evenly-spaced categorical data as continuous is a widely accepted approach, there are alternative methods to answer different types of questions that may be well suited for these data, given the large percentage of "0" scores.

- Because the data are strongly right-skewed, it may be useful to conduct a non-parametric test, such as a Chi-squared test of independence to test distribution differences among groups or a logistic regression to examine the likelihood that a student might score a "0" on the posttest.
- This report restricted the analysis to the overall scale scores. Further analysis may incorporate the "Reading" variable, either as a stratification variable or as a control variable.

The teachers' category of implementation was not ready at the time of this analyses but was ready by the time RMC research was prepared to send the report chapter to UI (see Exhibit 81). The team will need to decide if it is appropriate to include implementation category as a variable in follow-up

analyses. Since this study only includes data through Year 3 it would be the most appropriate to include the teachers' implementation category for Year 3 and not Year 4.

Exhibit 81: Category of LLAMA Implementation

	Year 3		Year 4	
Teacher	**No of Teachers**	**%**	**No of Teachers**	**%**
High Implementer	2	17%	2	17%
Medium Implementer	5	42%	3	25%
Low Implementer	5	42%	7	58%
No Implementation	0	0%	0	0%
Total	12	100%	12	100%

As the research team embarked on the SARA analyses this year, we realized a challenge with the SARA data is that the distributions are varying widely from year to year and from teacher to teacher--which makes it challenging to apply the same methods for each analysis. The statistical assumptions for different tests hold sometimes, but it depends on who is included in the analysis. An interesting observation is that for the original SARA study, the treatment group had posttest means that were higher than 1 for all but one items. In contrast, the treatment group for this study had posttest means that were all below one, with the exception of one item.

The LLAMA team should consider why the Cohort 1 students in the original study had higher posttest scores on average than the Cohort 2 students in this study.

Comparison Post Only Items (Items 6-9)

The second hypothesis is that students in the treatment group will score significantly higher than students in the control group on Items 6–9. (Items 6–9 were only included on the post assessments.) To test this hypothesis, the research team conducted an independent samples *t*-test comparing post-scores by treatment group (treatment and control). Differences in scores were very small between groups and non-significant ($p < 0.05$).

Exhibit 82. Posttest Item Score Comparisons by Group

	Control Group		Treatment Group		
Item	***M***	***sd***	***M***	***sd***	***p***
Item 6	0.18	0.636	0.20	0.697	0.611
Item 7	0.03	0.164	0.02	0.155	0.809
Item 8	0.14	0.345	0.13	0.332	0.632
Item 9	0.20	0.494	0.21	0.492	0.712

Note. Control Group: $n = 327$. Treatment Group: $n = 445$.
$^{*}p < 0.05$. $^{**}p < 0.01$. $^{***}p < 0.001$.

Study 2: Student Argumentation Study—Substudy 2 of Case Study Teachers

As noted earlier in the report, the original SARA study produced promising findings, so the LLAMA team decided to conduct additional studies to examine student argumentation in greater depth. In Year 3, during the summer 2019, the LLAMA team decided to conduct an additional study with the case study teachers. The case study teachers are teachers that had high LLAMA implementation in their classes. Substudy 2 includes matching sets of pre and post SARA data from these 3 high implementing case study teachers across Years 1, 2, and 3. This study does not include a comparison group. In terms of category of LLAMA implementation, all of the teachers received a rating of a 4-high implementer. A teacher was coded as a '4' if the data showed the teacher (a) engaged students in learning experiences targeting the learning of the 12 CPs, (b) included viable argumentation as a regular feature of instruction, and (c) included viable argumentation for generalizations frequently (i.e., at least twice a month).

Study Recruitment

Study recruitment is described within the chapter **Study 2: Student Argumentation Study**.

Instrument Development and Interrater Reliability

Study recruitment is described within the chapter **Study 2: Student Argumentation Study**.

Data Collection

This study includes 247 matching sets of SARA data from the selected Cohort 1 and Cohort 2 teachers. There are two years of data from the Cohort 1 teachers and one year of data from the Cohort 2 teacher. There are 23 matching SARAs from one Cohort 1 teacher and 180 from the other Cohort 1 teacher. There are 44 matching SARAS from the Cohort 2 teacher in Year 3.

Scoring

The LLAMA team completed the scoring of these SARAs by September 1, 2020.

Findings

This chapter includes the results from pre and post sets of SARA responses collected from 247 students taught by three teachers across three study years. As Exhibit 83 shows, SARA responses were collected from 78 students taught by all three teachers in their first year of LLAMA implementation, from 77 students taught by two teachers in their second year of implementation, and from 92 students taught by two teachers in their third year of implementation.

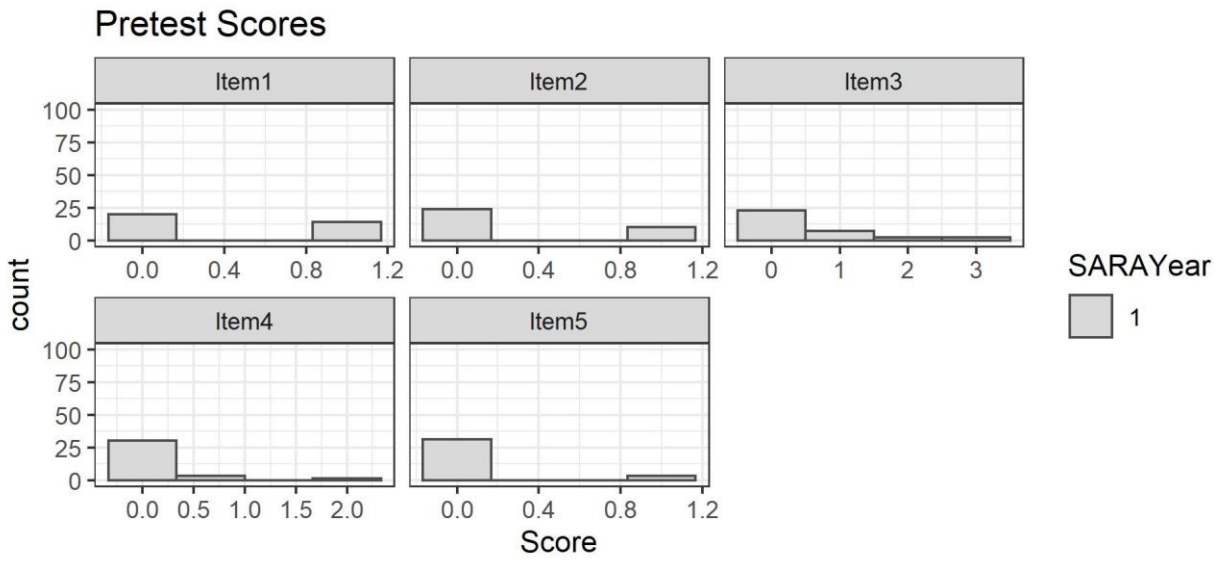
Exhibit 83: Number of Matching Sets of Pre and Post SARAs Per Case Study Teacher by Number of Years in LLAMA

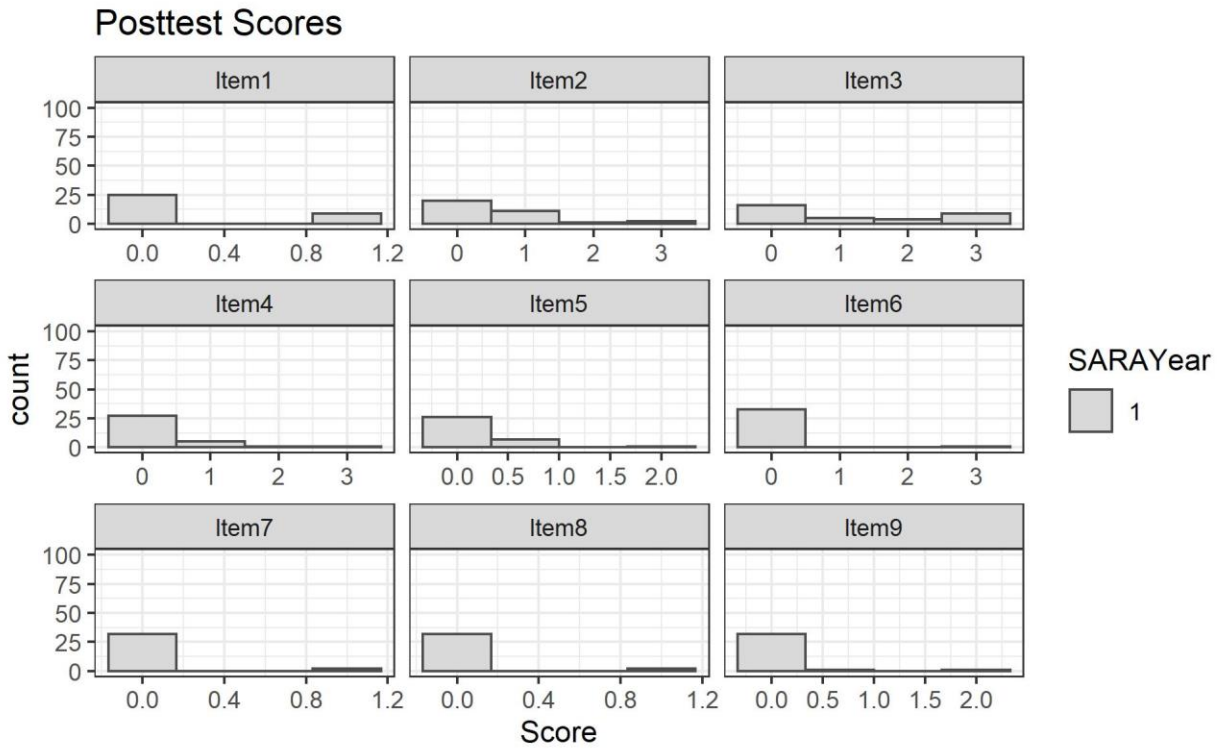
Implementation Year	Number of Matching SARA Sets
Year 1	78
Year 2	77
Year 3	92
Total	247

Pre-Post Distributions of Individual Item Scores

The research team assessed the pre-post distribution of individual items scores and analyses of data from Year 1 may be problematic due to the lack in variance for most items pre and post. Note that at least 1 student scored a 2 or 3 at pre in Year 1 on Items 3 and 4. The team decided it would be appropriate to run a paired samples *t*-test.

Exhibit 84. Pretest and Posttest Item Score Distributions, Year 1

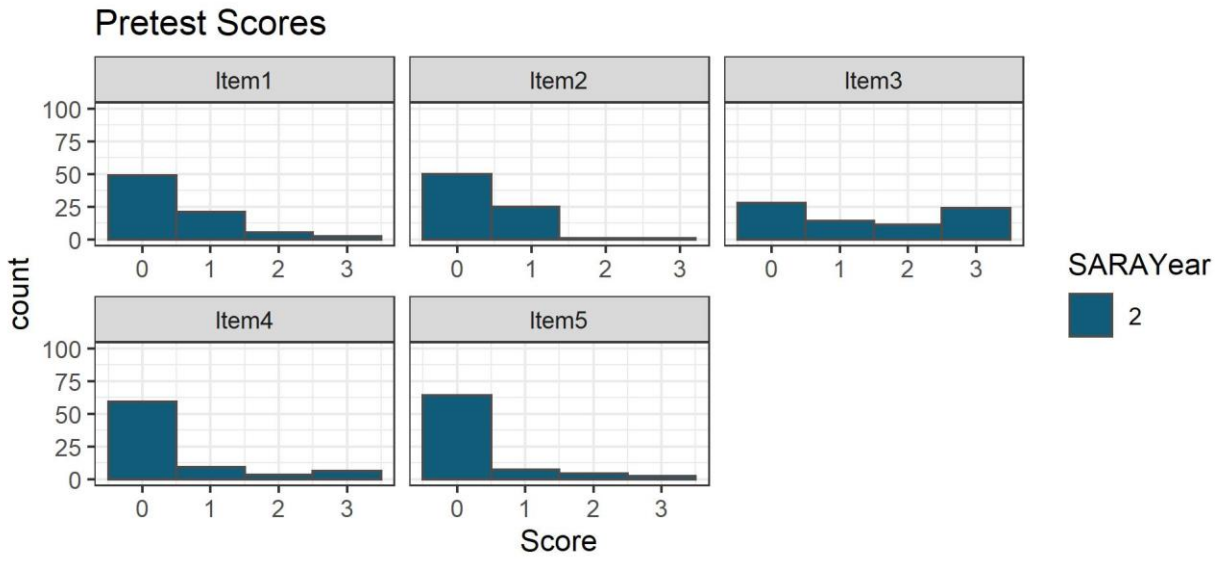


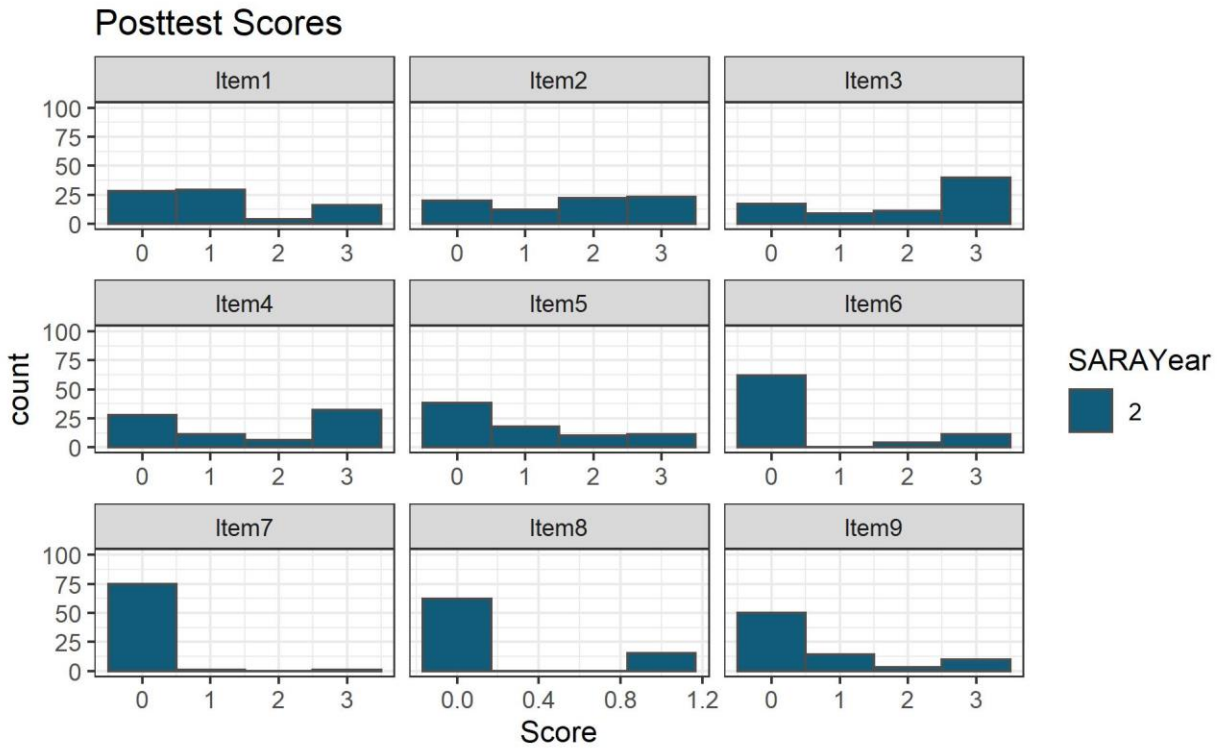


Year 2

Variance is much better in Years 2 and 3, but the sample size is too small to run a MANOVA. The research team decided it would be appropriate to run a paired samples *t*-test.

Exhibit 85. Pretest and Posttest Item Score Distributions, Year 2

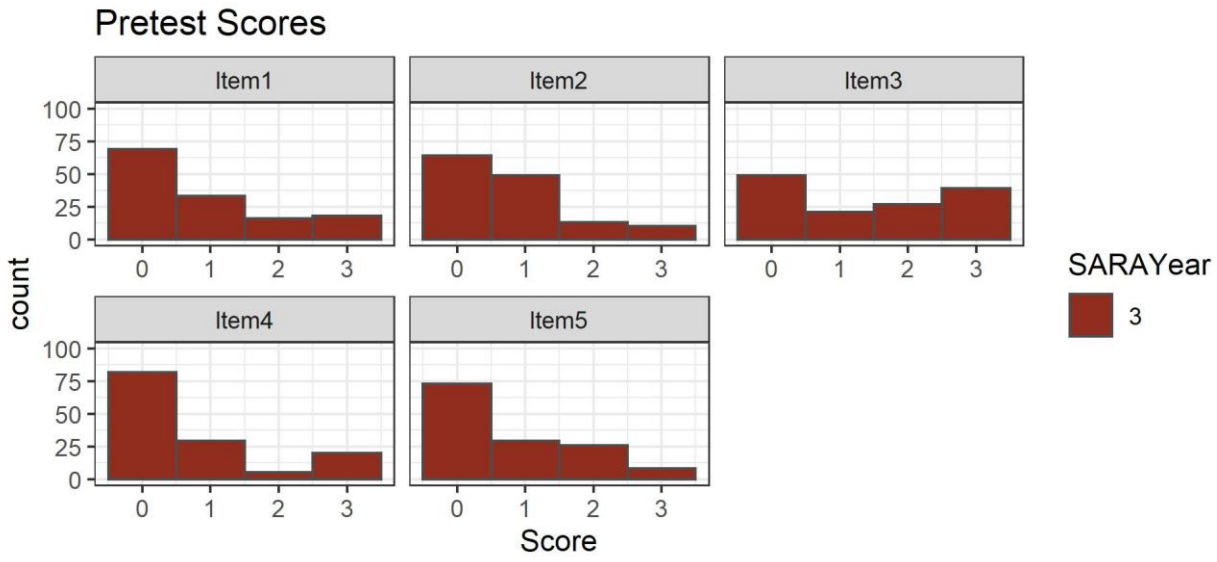




Year 3

Variance is much better in Years 2 and 3, but the sample size is too small to run a MANOVA. The research team decided it would be appropriate to run a paired samples *t*-test.

Exhibit 86. Pretest and Posttest Item Score Distributions, Year 3



Pre-Post Comparisons: T Tests

Paired samples *t*-test was conducted as a repeated measures test to analyze SARA score changes between the pre and post tests. Exhibits 87-90 present the overall SARA score changes by LLAMA implementation year. The results indicate that in the first year of LLAMA implementation, SARA scores significantly improved in the post test for Problem 2 and 3, whereas in the second and third year of implementation post test scores were significantly higher in all problems. The results may suggest that student growth in argumentative reasoning improves by teacher's experience implementing LLAMA. Exhibit 88 indicates that the largest growth magnitude was observed in the second year of LLAMA implementation. SARA pretest scores were slightly higher in the third year of teachers implementing LLAMA than the first two years, so the growth magnitude was not as large as that in the second implementation year.

Exhibit 87. First Year Implementation SARA Performance Over Time

	Pretest (*N* = 78)		Posttest (*N* = 78)		Pre-Post
SARA	Mean	SD	Mean	SD	Mean Difference
Problem 1	0.54	0.60	0.54	0.88	0.00
Problem 2	0.47	0.66	1.10	1.05	0.63***
Problem 3	0.92	1.20	1.33	1.27	0.41*
Problem 4	0.40	0.73	0.49	0.85	0.09
Problem 5	0.42	0.80	0.60	0.92	0.18

Note. Statistically significant based on paired samples *t* test results; *$p < .05$, ***$p < .001$.

Exhibit 88. Second Year Implementation SARA Performance Over Time

	Pretest (*N* = 77)		Posttest (*N* = 77)		Pre-Post
SARA	Mean	SD	Mean	SD	Mean Difference
Problem 1	0.48	0.72	1.10	1.12	0.62***
Problem 2	0.39	0.57	1.62	1.17	1.24***
Problem 3	1.40	1.27	1.96	1.24	0.56**
Problem 4	0.43	0.89	1.55	1.35	1.12***
Problem 5	0.27	0.68	0.92	1.10	0.65***

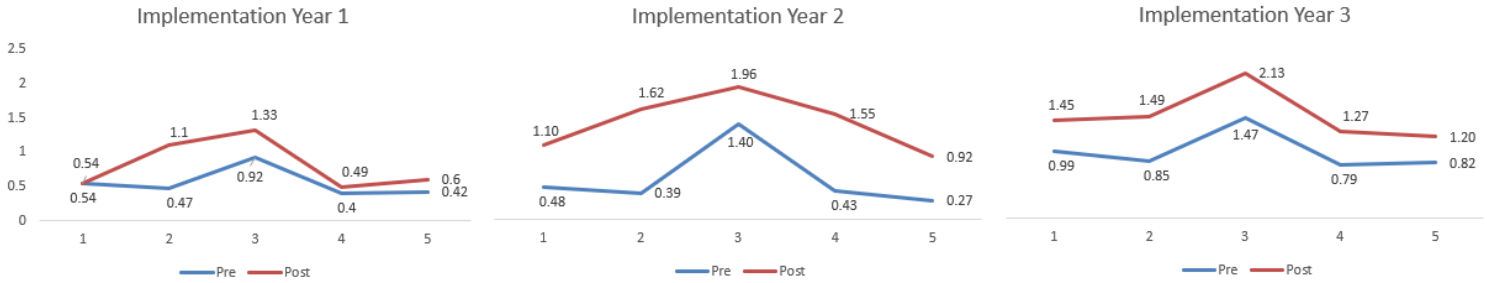
Note. Statistically significant based on paired samples *t* test results; **$p < .01$, ***$p < .001$.

Exhibit 89. Third Year Implementation SARA Performance Over Time

	Pretest (*N* = 92)		Posttest (*N* = 92)		Pre-Post
SARA	Mean	SD	Mean	SD	Mean Difference
Problem 1	0.99	1.21	1.45	1.35	0.46**
Problem 2	0.85	0.96	1.49	1.12	0.64***
Problem 3	1.47	1.20	2.13	1.11	0.66***
Problem 4	0.79	1.17	1.27	1.34	0.48**
Problem 5	0.82	0.96	1.20	1.16	0.38**

Note. Statistically significant based on paired samples *t* test results; **$p < .01$, ***$p < .001$.

Exhibit 90. SARA Score Changes by Implementation Year



Comparison Post Only Items (Items 6-9)

Items 6-9 were only included on the post assessments. The research team conducted a naïve analysis (i.e., not correcting for baseline differences or controlling for other variables of student post SARA) of ANOVA. Exhibit 91 shows that students demonstrated significant post test score differences across years for Problem 6 and Problem 9. Post hoc multiple comparison results indicated that, for Problem 6, student post test scores significantly increased between the implementation Year 2 and Year 3 and between implementation Year 1 and Year 3. For Problem 9, significant score improvement was observed only between implementation Year 1 and Year 2. The results are mixed. The results suggest that student growth for Problem 6 improves as teachers' experience implementing LLAMA but the same pattern does not hold for Problems 7, 8 and 9. The LLAMA team should consider what is different about Problem 6 and why more teachers' experience implementing LLAMA had a greater impact on Problems 1-5 compared to problems 6-9.

Exhibit 91. Student Posttest SARA Scores by Teacher Implementation Year

	Year 1 (*N* = 78)		Year 2 (*N* = 77)		Year 3 (*N* = 92)	
SARA	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***
Problem 6***	0.26	0.81	0.53	1.11	1.23	1.45
Problem 7	0.05	0.22	0.05	0.36	0.04	0.21
Problem 8	0.10	0.31	0.19	0.40	0.11	0.48
Problem 9***	0.18	0.45	0.65	1.05	0.08	0.31

Note. Statistically significant based on ANOVA test results; ***$p < .001$.

Study 2: Student Argumentation Study—Substudy 3 Year 4 Case Study

As noted earlier in the report, the original SARA study produced promising findings, so the LLAMA team decided to conduct additional studies to examine student argumentation in greater depth. The LLAMA team decided to collect pre and post data in Year 4 from the Cohort 2 teachers (*n* = 12). The LLAMA team hypothesized that the students would show significant gains pre to post, and more so than in prior years because the Cohort 2 teachers would have been taught argumentation in their classes for two years. Due to COVID-19 post data were only collected from 13 students in one Cohort 2 teacher's class; therefore, this became a descriptive case study of one teacher's data.

Study Recruitment

Study recruitment is described within the chapter **Study 2: Student Argumentation Study**.

Instrument Development and Interrater Reliability

Study recruitment is described within the chapter **Study 2: Student Argumentation Study**.

Data Collection

Exhibit 92 shows the number of pre and post matching SARAS per teacher per year for this study.

> *The research team planned to include data from the pre and post assessments administered during Year 4 (2019-2020 school year). The pre assessments were administered as planned; however, due to COVID -19 only one teacher was able to collect post assessments.*

Exhibit 92: Number of SARAs for Substudy 3

Teacher	Pre	Post	Matching Pre and Post
Teacher 1	63	0	0%
Teacher 2	24	0	0%
Teacher 3	22	0	0%
Teacher 4	47	0	0%
Teacher 5	51	0	0%
Teacher6*	0	0	0%
Teacher 7	15	0	0%
Teacher 8	81	0	0%
Teacher 9	50	0	0%
Teacher 10	39	13	33%
Teacher 11	36	0	0%

Teacher	Pre	Post	Matching Pre and Post
Teacher 12	30	0	0%
Total	458	13	3%

*This teacher submitted 44 assessments, however did not send student assent forms we the assessments are not counted and will not be included in the analysis.

Sampling

There is no sampling for this study. The LLAMA team planned to score all of the data collected (i.e., a 13 matching pre-post sets of SARAs).

Scoring

The LLAMA team scored the data in alternating pairs. Interrater reliability was assessed descriptively based on the exact rater agreement and adjacent rater agreement on each assessment item. Single rater intra-class correlation coefficient (ICC) was not estimated due to the small sample size and lack of variance in student scores. As Exhibit 93 shows, while adjacent agreement was high, exact agreement was only high for Items 2 and 6 (i.e., 70% or higher). Based on these findings, RMC Research decided to use the mean score of each pair in the analyses.

Exhibit 93: Exact and Adjacent Agreement by Problem

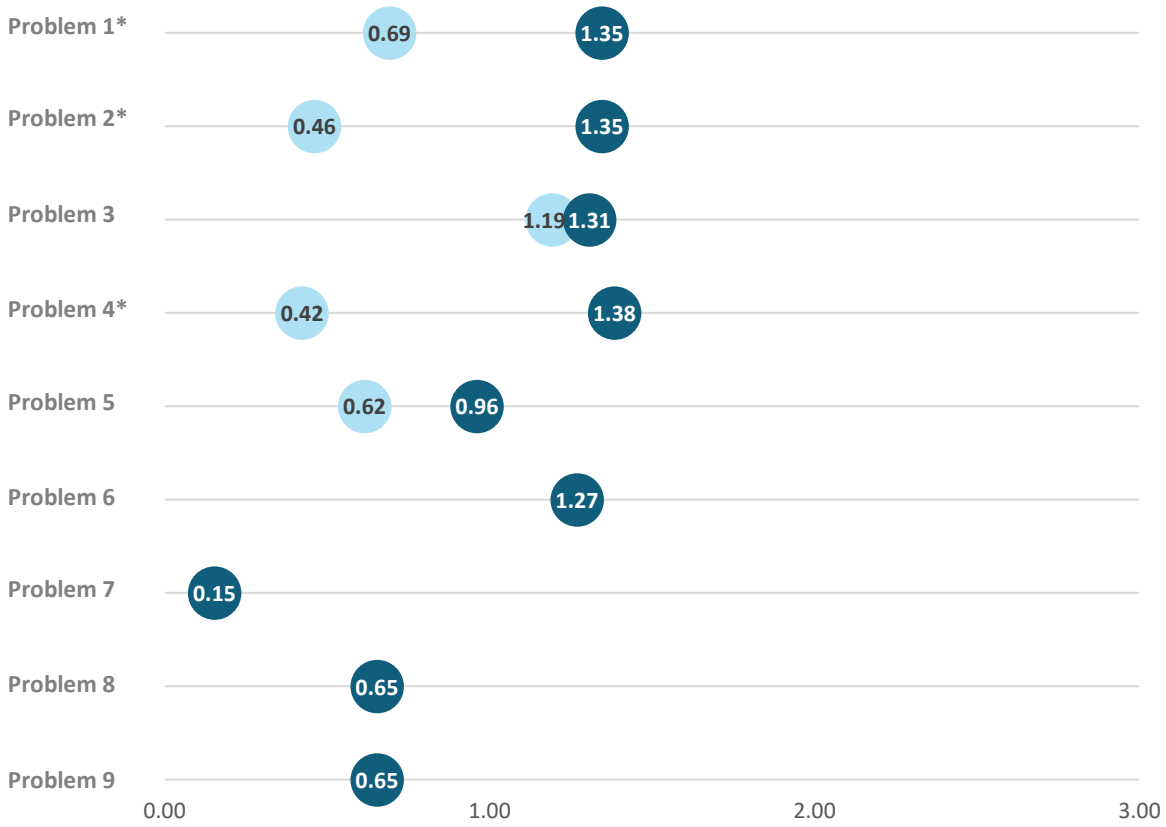
	Exact	Adj	Pair
Problem 1	65%	100%	A
Problem 2	81%	92%	B
Problem 3	62%	96%	C
Problem 4	54%	96%	B
Problem 5	42%	92%	C
Problem 6	92%	92%	A
Problem 7	69%	100%	A
Problem 8	62%	100%	B
Problem 9	54%	92%	C

Findings

RMC Research employed paired *t*-tests using the mean scores by item to assess change over time. Findings should be interpreted with caution due to the small sample size.

> Students showed an increase in scores for all problems between pre and post, and a significant increase for Problems 1, 2, and 4. Problems 6-9 were only on the post-SARA.

Exhibit 94: Pre and Post Mean Scores by Item



Note. Pre Post *n* = 13. Rating Scale Scores: 0 = No elements of a viable argument. 1 = Limited elements of a viable argument. 2 = Elements of a viable argument. 3 = Viable argument.
Problems 6-9 were only on the post-SARA.
*Significant differences at $p < .05$. Differences assessed using paired *t*-tests. Results should be interpreted with caution due to the small sample size.

Study 3: Teacher Argumentation Study

RMC Research conducted an experimental research study to address Research Question 3, "Does the implementation of the LLAMA intervention change teachers' ability to construct viable arguments and critique the arguments of others?" In this design, teachers were randomly assigned to the treatment and control group. The independent variable is the LLAMA intervention and the dependent variable is teacher argumentation and reasoning skills. The treatment teachers received the LLAMA intervention in Year 1 and Year 2, whereas the control group did not. The treatment and control teachers completed the Teacher Argument and Reasoning Assessment (TARA) as a pretest in Year 1 (prior to the start of the intervention) and a posttest at the end of Year 2. For both the pretest and posttest, teachers completed the posttest version of the Student Argument and Reasoning Assessment (i.e., the one with 9 items; herein referred to as the Teacher Argument and Reasoning Assessment [TARA]). **The hypothesis is** that teachers in the treatment group will improve significantly more in argumentation skills than teachers in the control group. This hypothesis was supported both with the original and revised rubric. The results of the Teacher Argumentation Study were provided in the prior annual report. As the UI team scored all the TARA and SARA assessments the UI team realized that the scoring rubric needed some modifications because the original rubric did not distinguish clearly enough between some scoring levels on a few items, resulting in scorer disagreements. After updating the rubric, the UI team rescored all the TARA assessments with the new rubric. This chapter provides the results of the TARA analyses with the updated rubric scores and a summary of how the results differed between the two rubrics. In addition to the original TARA research design, this chapter includes the pre and post results for the control teachers that participated in the LLAMA intervention in Years 3 and 4 after the completion of the original research study. The analysis includes the TARAs completed as a pretest at the conclusion of Year 2 (prior to their participation in the LLAMA professional development) and a posttest at the end of Year 4.

TARA Methods

Study Recruitment and Random Assignment

Study recruitment and random assignment is described within the chapter **Study 1: Student Achievement Study**.

Instrument Development & Interrater Reliability

> ***Target:*** *Treatment and control teachers complete the* ***Teacher Argument and Reasoning Assessment*** *beginning of Year 1 and end of Year 2.* ***Status:*** *Met*

The research team developed the Teacher Argument and Reasoning Assessment (TARA) to measure teachers' abilities to construct viable arguments and critique others' arguments. The TARA was originally developed and validated in the LAMP pilot study (NSF Award Number: 1317034). Instrument development and reliability is discussed in the SARA Study Chapter. The TARA scoring is shown in Exhibit 95. There are two types of ratings each TARA received during scoring. Total TARA scores range from 0-27; scores per item ranged from 0-3. Due to time constrains the PI of the project blindly scored all TARAs, i.e., the PI did not know which TARAs were pre or post or treatment or control. RMC Research prepared the blind data set for the PI.

Exhibit 95: TARA Ratings and Rating Scales

Rating Type	Rating Scale
Read Correctly: measures students' understanding of mathematical objects/definitions and of the format/structure/instructions of the task	0: No evidence of understanding 1: Some understanding 2: Demonstrates understanding
Viable Argumentation: measures students' demonstration of a viable argument	0: No elements of a viable argument 1: Limited elements of a viable argument 2: Elements of a viable argument 3: Viable argument

Data Collection

Target: *Treatment and control teachers complete the Teacher* ***Argument and Reasoning Assessment*** *beginning of Year 1 and end of Year 2.* ***Status:*** *Met.*

LLAMA teachers (in both the treatment and control groups) completed the Teacher Argument and Reasoning Assessment as a pretest, administered prior to the start of project activities (December 2016); they completed the same assessment as a posttest at the end of Year 2 (May 2018).

Modifications. The research team collected data as planned but included some additional data collection time points. The research team administered the TARA at the end of Year 3 as a second post for Cohort 1 and as an interim assessment for Cohort 2, and also administered the TARA to Cohort 2 again at the end of Year 4 as a posttest.

As shown in Exhibit 96, for the original study, 76% of the intent to treat treatment teachers and 58% of the control teachers submitted pre and post assessments. Of the teachers that were active by the end of Year 2, 100% submitted pre and post assessments. Exhibit 97 shows the TARA completion across the 4 years of LLAMA. For the intent to treat teachers, completion ranges from 39-85% and for the active teachers ranges from 90-100%.

Exhibit 96: TARA Completion for Primary RCT Study (Years 1 and 2)

	Intent to Treat		Active[a]	
Time Period	***Teachers***	**Completion**	***Teachers***	**Completion**
Treatment	*34*		*25*	
Pre	29	85%	25	100%
Post	26	76%	25	100%
Control	**31**		**16**	
Pre	26	84%	16	100%
Post	18	58%	16	100%

Note. One Cohort 1 teacher and two Cohort 2 teachers who dropped prior to the end of Year 2 submitted a posttest. They are included in the analytic sample for this report.
[a]Active as of end of 2017–2018 school year.

Exhibit 97: TARA Completion for Substudy (4 Time Points: Years 1, 2, 3, and 4)

	Intent to Treat		Active[a]	
Time Period	***Teachers***	**Completion**	***Teachers***	**Completion**
Cohort 1	*34*		*20*	
Pre	29	85%	20	100%
Post 1	26	76%	20	100%
Post 2	18	53%	18	90%
Cohort 2	**31**		**14**	
Pre 1	26	84%	14	100%
Pre 2	18	58%	14	100%
Post	18	58%	14	100%
Post 2	12	39%	12[b]	100%

[a]Active as of end of 2018–2019 school year.
[b]Active as of end of 2019-2020 school year

TARA Scoring

Due to time constraints, the TARA assessments were blindly scored in fall 2018 by one person, the Principal Investigator for LLAMA, using the SARA rubric. RMC Research prepared the TARA assessments for the Principal Investigator and ensured the Principal Investigator could not identify the treatment nor control teachers, nor if the TARA was a pre or post assessment. In Year 5, the Year 1 and Year 2 TARAs were rescored using the updated rubric and Year 3 and Year 4 TARAs were scored using the updated scoring rubric. TARAs were divided between the PI and 2 co-PIs. Because the team elected to score blank problems as a "0" on the rubric, there is no missing data to account for in these analyses.

Analytic Sample for Original Experimental Research Design

Of the 65 randomly assigned teachers, 44 submitted pretest and posttest TARA data for this study and are included in the analyses. Exhibit 98 shows the demographics for the entire analytic sample and also

for the treatment and control study groups. Inference tests were conducted to compare study groups on these characteristics (Chi-squared for categorical variables; Mann-Whitney U-tests or independent *t*-tests for continuous variables, depending on whether the frequency distributions appear to be Normal).

There are fewer teachers from Montana than from Washington or Idaho, but not significantly so. Teachers in the analytic sample have Bachelor's degrees (100%), are predominantly White (82%), and a majority are female (67%). Most teach in a middle or junior high school (89%), and many teach in a rural school (71%). The control group included more teachers with an advanced degree than the treatment group (67% and 46%, respectively). Though the difference in degree attainment was not significant between the study groups, the number of graduate credits in mathematics was significantly higher for the control group than the treatment group ($p = 0.004$).

Exhibit 98: Demographics of Analytic Sample

Item	All	Tx	Ct
Total in Analytic Sample	44	26	18
State (%)			
Idaho	46%	50%	39%
Montana	16%	15%	17%
Washington	39%	35%	44%
School setting (%)			
Rural	71%	69%	72%
Suburban	18%	19%	17%
Urban	11%	12%	11%
School type (%)			
K–8	2%	4%	0%
K–12	2%	0%	6%
Jr/sr high	5%	8%	0%
Middle/junior high	89%	88%	89%
High school	2%	0%	6%
Alternative	0%	0%	0%
Education and credentials (%)			
Bachelor's	100%	100%	100%
Master's	57%	46%	67%
Doctorate	2%	0%	6%

Item	All	Tx	Ct
Ethnicity[ab] (%)			
White	82%	81%	83%
Asian	2%	4%	0%
American Indian	2%	0%	6%
Gender[b] (%)			
Female	67%	62%	67%
Male	24%	23%	22%
Years of experience (*M*)			
Years teaching total	11.1	10.0	12.7
Years teaching mathematics	10.4	9.7	11.6
Highest level mathematics courses completed (%)			
100–199 (freshman)	9%	12%	6%
200–299 (sophomore)	9%	8%	11%
300–399 (junior)	21%	27%	11%
400–499 (senior)	23%	27%	17%
500+ (graduate)	37%	23%	56%
Course credits in mathematics (*M*)			
Undergraduate credits	19	22	15
Graduate credits**	7	3	12

Note. All = analytic sample. Tx = randomly assigned treatment teachers. Ct = randomly assigned control teachers.
[a]May have listed more than 1.
[b]Seven teachers did not provide their race/ethnicity; six teachers did not provide their gender.
**Significant at $p < 0.01$.

What Works Clearinghouse Guidelines

What Works Clearinghouse utilizes three steps for reviewing RCTs and QEDs that assign individual subjects to the intervention or comparison condition: [17]

- **Step 1:** Assess the study design,
- **Step 2:** Assess sample attrition, and
- **Step 3:** Assess equivalence of the intervention and comparison groups at baseline (prior to the intervention).

Step 1: Assess the study design

"To be eligible for the WWC's highest rating for group design studies, *Meets WWC Group Design Standards Without Reservations*, the study must be an RCT with low levels of sample attrition. A QED or high-attrition RCT is eligible for the rating *Meets WWC Group Design Standards With Reservations* if it satisfies the WWC's baseline equivalence requirement that the analytic intervention and comparison groups appear similar at baseline. A QED or high-attrition RCT that does not satisfy the baseline equivalence requirement receives the rating *Does Not Meet WWC Group Design Standards*."

This study is an RCT.

Step 2: Assess sample attrition

The What Works Clearinghouse (Institute of Education Sciences, 2008) definition of differential attrition is:

- **Differential Attrition:** "Differential attrition refers to the situation in which the percentage of the original study sample retained in the follow-up data collection is substantially different for the intervention and the control groups. Severe differential attrition makes the results of a study suspect, because it may compromise the comparability of the study groups."

The What Works Clearinghouse (Institute of Education Sciences, 2008) definition of overall attrition is:

- **Overall Attrition:** "Attrition is defined as a failure to measure the outcome variable on all the participants initially assigned to the intervention and control groups. High overall attrition generally makes the results of a study suspect, although there may be rare exceptions."

For the intent to treat sample included in this study, attrition was high. Of the 34 teachers assigned to the treatment group, 26 submitted both pre- and posttest data (76% submitted data; 24% attrition in the treatment group). Of the 31 teachers assigned to the control group, 18 submitted both pre- and posttest data (58% submitted data; 42% attrition in the control group). Thus, the differential attrition for the intent to treat sample is 18% and the overall attrition is 32% (44 of the 65 assigned teachers submitted data for this study).

In order to satisfy the WWC definition of "low attrition," differential attrition must fall below 11%, and the overall attrition must fall below the corresponding threshold listed in Table II.1.[18] Since the differential attrition falls above 11%, this study would be considered to have high attrition.

[17] Page 5; https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_handbook_v4.pdf
[18]Page 13; https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_handbook_v4.pdf

Due to high differential and overall attrition, this study does not meet the *WWC Group Design Standards Without Reservations*. This study may be considered for *WWC Group Design Standards With Reservations* if there is baseline equivalence.

Step 3: Assess equivalence of the intervention and comparison groups at baseline (prior to the intervention).

In an effort to avoid oversimplifying the analysis and to preserve the variance between scores for different problems, a multivariate approach was used in the analyses. Multivariate analyses allow us to conduct a holistic analysis with multiple response variables (i.e., comparing mean vectors of 9 problem scores), rather than reducing an assessment to a single sum, average, or scale score (i.e., comparing mean scores for a single response variable).

TARA pretest results were compared using Hotelling's T^2 test (a generalization of Student's *t*-test commonly used in univariate analyses[19]) to examine the baseline equivalence between the treatment and comparison groups. Similar to how an independent *t*-test compares the mean response of a single variable for two groups, Hotelling's T^2 test compares mean vectors of responses of multiple variables for two groups (in this case, the response variables are 9 separate rating scores for each teacher—one rating score for each problem in the TARA).

The intervention and comparison groups were not equivalent at baseline.

Box's M-test for homogeneity of covariance matrices suggests that the treatment and control groups have unequal variance ($p < 0.001$), so unequal variances are assumed and non-pooled variances (i.e., separate variances) were used in a Hotelling's T^2 test comparison of pretest scores for the treatment and control groups. Although Hotelling's T^2 test provided little evidence of a difference in baseline scores between the 2 groups (i.e., not statistically significant; $p = 0.268$; T^2 statistic = 11.1), Mahalanobis D^2 was used to estimate an effect size of 1.04, which is considered to be a large effect size[20] (comparable to Cohen's *d*[21] estimating a large effect size in univariate analyses). Even though the difference between the treatment and control group pretest scores were not statistically significant, due to the large effect size the research team opted for the conservative approach by using the TARA pretest scores as covariates in the analyses to account for any possible group differences at the outset of the study.

Descriptive statistics including measures of central tendency (means and standard deviations)[22] and frequency distributions for the entire analytic sample and for each study group are shown below in the descriptive summary tables (Exhibits 99-101). Overall on the pretest, teachers in both the treatment and control groups did well on Problems 1–6 and poorly on Problems 7–9. Ninety eight percent of teachers from both study groups scored a "3" on Problem 6; Eighty two percent scored a "0" on Problem 7.

[19]Hotelling, H. (1931). "The generalization of Student's ratio". Annals of Mathematical Statistics. 2 (3): 360–378. doi:10.1214/aoms/1177732979.

[20]Sapp, Marty & Obiakor, Festus & J. Gregas, Amanda & Scholze, Steffanie. (2007). Mahalanobis distance: A multivariate measure of effect in hypnosis research. Sleep and Hypnosis. 9. 67-70.

[21] Del Giudice, Marco. (2017). Heterogeneity Coefficients for Mahalanobis' D as a Multivariate Effect Size. Multivariate Behavioral Research. 52. 216-221. 10.1080/00273171.2016.1262237.

[22] The mean or average value is a measure of central tendency computed by adding a set of values and dividing the sum by the total number of values. The standard deviation (*SD*) is a measure of how spread out a set of values is. Higher standard deviations indicate greater variability in data across respondents.

Exhibit 99. Analytic Sample Pretest TARA Scores (Year 1)

	Rating Scale Score									
	0		1		2		3			
Problem	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***M***	***sd***
Problem 1	0	0%	1	2%	6	14%	37	84%	2.82	0.446
Problem 2	1	2%	2	5%	14	32%	27	61%	2.52	0.698
Problem 3	6	14%	7	16%	14	32%	17	39%	1.95	1.056
Problem 4	3	7%	3	7%	4	9%	34	77%	2.57	.900
Problem 5	0	0%	6	14%	18	41%	20	45%	2.32	0.708
Problem 6	1	2%	0	0%	0	0%	43	98%	2.93	0.452
Problem 7	36	82%	7	16%	0	0%	1	2%	0.23	0.565
Problem 8	10	23%	23	52%	2	5%	9	20%	1.23	1.031
Problem 9	12	27%	23	52%	3	7%	6	14%	1.07	.950

Note. n = 44. Rating Scale Scores: 0 = No elements of a viable argument. 1 = Limited elements of a viable argument. 2 = Elements of a viable argument. 3 = Viable argument

Exhibit 100. Analytic Sample Treatment Teacher Pretest TARA Scores (Year 1)

	Rating Scale Score									
	0		1		2		3			
Problem	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***M***	***sd***
Problem 1	0	0%	1	4%	2	8%	23	88%	2.85	0.464
Problem 2	0	0%	0	0%	8	31%	18	69%	2.69	0.471
Problem 3	4	15%	4	15%	10	38%	8	31%	1.85	1.047
Problem 4	2	8%	1	4%	2	8%	21	81%	2.62	.898
Problem 5	0	0%	3	12%	11	42%	12	46%	2.35	0.689
Problem 6	1	4%	0	0%	0	0%	25	96%	2.88	0.588
Problem 7	22	85%	3	12%	0	0%	1	4%	0.23	0.652
Problem 8	5	19%	14	54%	0	0%	7	27%	1.35	1.093
Problem 9	5	19%	14	54%	2	8%	5	19%	1.27	1.002

Note. n = 26. Rating Scale Scores: 0 = No elements of a viable argument. 1 = Limited elements of a viable argument. 2 = Elements of a viable argument. 3 = Viable argument

Exhibit 101. Analytic Sample Control Teacher Pretest TARA Scores (Year 1)

	Rating Scale Score									
	0		1		2		3			
Problem	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***M***	***sd***
Problem 1	0	0%	0	0%	4	22%	14	78%	2.78	0.428
Problem 2	1	6%	2	11%	6	33%	9	50%	2.28	0.895
Problem 3	2	11%	3	17%	4	22%	9	50%	2.11	1.079
Problem 4	1	6%	2	11%	2	11%	13	72%	2.50	.924
Problem 5	0	0%	3	17%	7	39%	8	44%	2.28	0.752
Problem 6	0	0%	0	0%	0	0%	18	100%	3.00	0.000
Problem 7	14	78%	4	22%	0	0%	0	0%	0.22	0..428
Problem 8	5	28%	9	50%	2	11%	2	11%	1.06	0.938
Problem 9	7	39%	9	50%	1	6%	1	6%	0.78	0.808

Note. n = 18. Rating Scale Scores: 0 = No elements of a viable argument. 1 = Limited elements of a viable argument. 2 = Elements of a viable argument. 3 = Viable argument

Findings for Posttest Comparisons of Treatment and Control Groups

> ***The hypothesis is that teachers in the treatment group will improve significantly more in argumentation skills than teachers in the control group.***

To compare the posttest scores for the treatment and control groups, RMC used a **Multivariate Analysis of Covariance (MANCOVA)** model: the 9 dependent variables are the 9 individual problem scores from the posttest; the independent variable is the study group (i.e., Treatment or Control); and the covariates are the 9 individual problem scores from the pretest. Using this model accounts for any potential baseline difference in pretest scores between groups by including the pretest scores as covariates. Using a multivariate approach preserves some of the complexity in a teacher's assessment performance that might otherwise be lost in summing, averaging, or otherwise aggregating the responses to a single score. The results of the MANCOVA analysis are shown in Exhibit 102 below.

There is strong evidence that the study group is significant in the model ($p < 0.001$), meaning that there is strong evidence to suggest a difference in posttest assessments between the Treatment and Control study groups. There is some evidence to suggest that the pretest score for Item 1 and Item 8 may also be significant as a covariate in the model ($p = 0.034$ and 0.019 respectively). With the previous version of the rubric this held true for only Item 8.

> *The research team should discuss the findings for Items 1 and 8; what this means theoretically and investigate it further. For example, what do items 1 and 8 have in common and why would high pretest knowledge of these items lead to higher overall TARA scores on the post?*

Exhibit 102: MANCOVA Results for Posttest Comparisons of Treatment and Control Groups

Variable	Coefficient[a]	*F*	*df1*	*df2*	*p*	Sig.
Treatment	0.320	5.900	9	25	< 0.001	***
Pretest Problem 1	0.526	2.500	9	25	0.034	*
Pretest Problem 2	0.620	1.704	9	25	0.141	
Pretest Problem 3	0.656	1.456	9	25	0.218	
Pretest Problem 4	0.611	1.766	9	25	0.126	
Pretest Problem 5	0.767	.843	9	25	.585	
Pretest Problem 6	0.732	1.019	9	25	0.452	
Pretest Problem 7	0.812	.642	9	25	0.751	
Pretest Problem 8	0.495	2.839	9	25	0.019	*
Pretest Problem 9	0.627	1.651	9	25	0.155	

Note. All: $N = 44$. Cohort 1: $n = 26$. Cohort 2: $n = 18$. Rating Scale Scores: 0 = No elements of a viable argument. 1 = Limited elements of a viable argument. 2 = Elements of a viable argument. 3 = Viable argument.
[a]The MANCOVA comparison used the Wilk's Lambda method for estimating the coefficients; however, the resulting *F* statistic and *p* values were the same among other methods (i.e., Pillai's Trace, Hotelling's Trace, Roy's Largest Root).
*Significant at $p < 0.05$.
**Significant at $p < 0.01$.
***Significant at $p < 0.001$.

Post Hoc Item Level Analyses

Because the MANCOVA results suggest significant differences between the posttest performance of the Treatment and Control study groups, the research team proceeded to conduct post-hoc analyses of the individual item scores. Analysis of Covariance (ANCOVA) was used for post-hoc comparisons: the dependent variable is the individual problem score from the posttest; the independent variable is the study group (i.e., Treatment or Control); and the covariates are the 9 individual problem scores from the pretest. A Tukey's Honest Significant Difference (Tukey's HSD) correction was applied to reduce the Type I error (i.e., "false positives"). Results of the post-hoc analyses are shown below in Exhibit 103.

> *Descriptively, treatment teachers outperformed control teachers on all items on the posttest. Differences were significant for Problems 1 and 4 at an alpha level of 0.038 and .033 respectively. There was stronger evidence to suggest significant differences for Problems 7, 8, and 9 ($p < 0.001$ for all three problems. The research team should discuss these results and consider what these findings may mean for future analyses. In terms of overall significance the findings were the same for the prior rubric.*

Exhibit 103. Post-Hoc Comparisons for Treatment and Control Posttest TARA Scores (Year 2)

Posttest Problem	Difference[a]	Lower Bound[b]	Upper Bound[b]	*p*[c]	Sig.
Problem 1	0.406	0.023	0.789	0.038	*
Problem 2	0.209	-0.181	0.600	0.283	
Problem 3	0.295	-0.324	0.914	0.340	
Problem 4	0.607	0.053	1.161	0.033	*
Problem 5	0.218	-0.250	0.686	0.350	
Problem 6	0.325	-0.258	0.908	0.265	
Problem 7	1.372	0.805	1.939	< 0.001	***
Problem 8	1.171	0.627	1.715	< 0.001	***
Problem 9	1.103	0.546	1.659	< 0.001	***

Note. n = 44. Rating Scale Scores: 0 = No elements of a viable argument. 1 = Limited elements of a viable argument. 2 = Elements of a viable argument. 3 = Viable argument
[a]Estimated difference between treatment and control groups' scores.
[b]95% confidence interval.
[c]*p*-value adjusted by Tukey's HSD correction.
*Significant at $p < 0.05$.
**Significant at $p < 0.01$.
***Significant at $p < 0.001$.

Descriptive statistics including measures of central tendency (means and standard deviations) and frequency distributions for the entire analytic sample and for each study group are shown below in the descriptive summary tables (Exhibits 104-106).

Exhibit 104. Analytic Sample Posttest TARA Scores (Year 2)

	Rating Scale Score									
	0		1		2		3			
Problem	***n***	**%**	***n***	**%**	***n***	**%**	***n***	**%**	***M***	***sd***
Problem 1	2	5%	1	2%	1	2%	40	91%	2.80	0.701
Problem 2	0	0%	4	9%	11	25%	29	66%	2.57	0.661
Problem 3	3	7%	6	14%	8	18%	27	61%	2.34	.963
Problem 4	4	9%	1	2%	2	5%	37	84%	2.64	.917
Problem 5	1	2%	6	14%	16	36%	21	48%	2.30	0.795
Problem 6	4	9%	2	5%	0	0%	38	86%	2.64	0.942
Problem 7	23	52%	8	18%	4	9%	9	20%	0.98	1.210
Problem 8	11	25%	12	27%	3	7%	18	41%	1.64	1.259
Problem 9	12	27%	15	34%	8	18%	9	20%	1.32	1.095

Note. n = 44. Rating Scale Scores: 0 = No elements of a viable argument. 1 = Limited elements of a viable argument. 2 = Elements of a viable argument. 3 = Viable argument

Exhibit 105. Analytic Sample Treatment Teacher Posttest TARA Scores (Year 2)

	Rating Scale Score									
	0		1		2		3			
Problem	*n*	%	*n*	%	*n*	%	*n*	%	*M*	*sd*
Problem 1	0	0%	0	0%	1	4%	25	96%	2.96	0.196
Problem 2	0	0%	1	4%	7	27%	18	69%	2.65	0.562
Problem 3	0	0%	6	23%	2	8%	18	69%	2.46	0.860
Problem 4	1	4%	0	0%	0	0%	25	96%	2.88	0.588
Problem 5	1	4%	2	8%	9	35%	14	54%	2.38	0.804
Problem 6	2	8%	0	0%	0	0%	24	92%	2.77	0.815
Problem 7	8	31%	5	19%	4	15%	9	35%	1.54	1.272
Problem 8	5	19%	3	12%	2	8%	16	62%	2.12	1.243
Problem 9	3	12%	9	35%	5	19%	9	35%	1.77	1.070

Note. n = 26. Rating Scale Scores: 0 = No elements of a viable argument. 1 = Limited elements of a viable argument. 2 = Elements of a viable argument. 3 = Viable argument

Exhibit 106. Analytic Sample Control Teacher Posttest TARA Scores (Year 2)

	Rating Scale Score									
	0		1		2		3			
Problem	*n*	%	*n*	%	*n*	%	*n*	%	*M*	*sd*
Problem 1	2	11%	1	6%	0	0%	15	83%	2.56	1.042
Problem 2	0	0%	3	17%	4	22%	11	61%	2.44	0.784
Problem 3	3	17%	0	0%	6	33%	9	50%	2.17	1.098
Problem 4	3	17%	1	6%	2	11%	12	67%	2.28	1.179
Problem 5	0	0%	4	22%	7	39%	7	39%	2.17	0.786
Problem 6	2	11%	2	11%	0	0%	14	78%	2.44	1.097
Problem 7	15	83%	3	17%	0	0%	0	0%	0.17	0.383
Problem 8	6	33%	9	50%	1	6%	2	11%	0.94	0.938
Problem 9	9	50%	6	33%	3	17%	0	0%	0.67	0.767

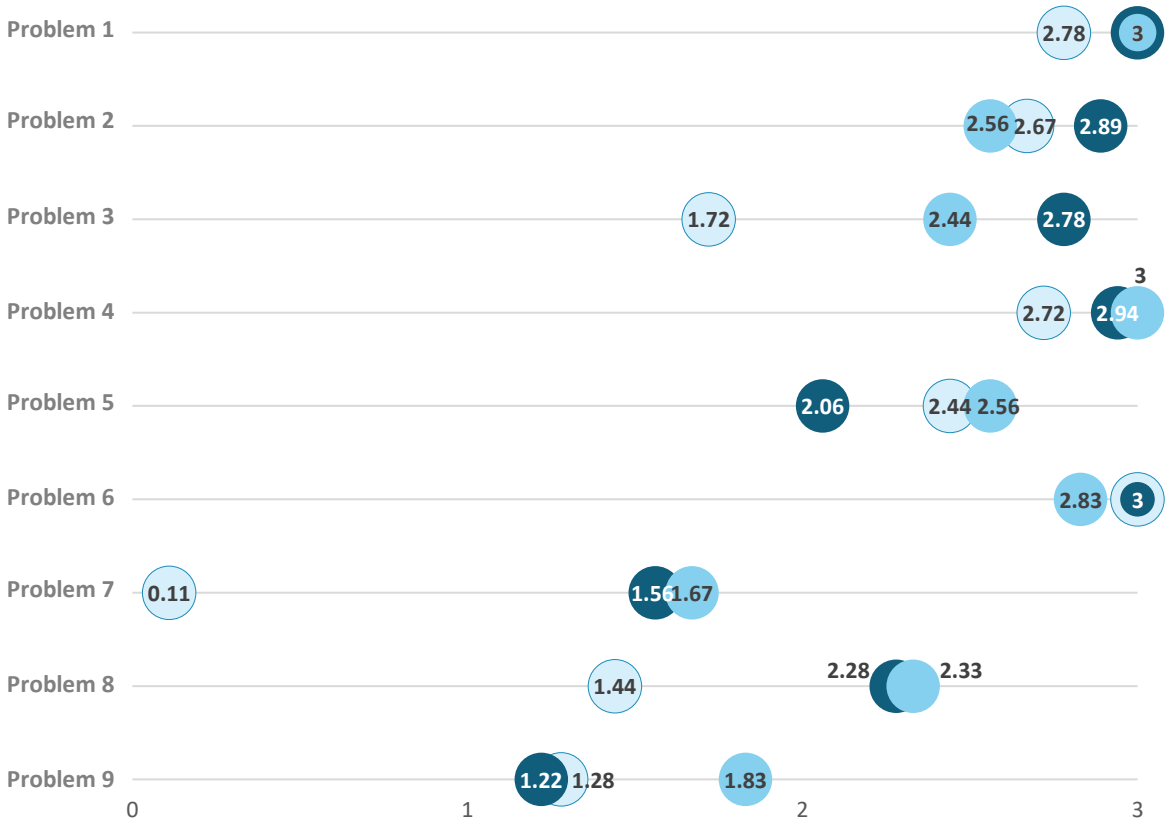
Note. n = 18. Rating Scale Scores: 0 = No elements of a viable argument. 1 = Limited elements of a viable argument. 2 = Elements of a viable argument. 3 = Viable argument

Cohort 1 TARA Scores Over Time

RMC Research used descriptive statistics to assess if there was an incremental increase in gains over all time points (from Pre to post 1 to Post 2, see Exhibit 107).

When looking at all three data points (pre, post 1, post 2) descriptively, there is no pattern in terms of gains over time. The expected pattern (increases in teacher argumentation incrementally between Pre, Post 1, and Post 2) is only seen for Problem 3.

Exhibit 107. Pre-Post Mean Comparisons of Cohort 1 Active Teachers



Note. **Pre** **Post 1** **Post 2** Cohort 1: *n* = 18. Rating Scale Scores: 0 = No elements of a viable argument. 1 = Limited elements of a viable argument. 2 = Elements of a viable argument. 3 = Viable argument.

Research Design 2: Descriptive Analyses of Cohort 2 (Control Teachers)

RMC Research conducted descriptive analyses of the subset of "active" Cohort 2 teachers (i.e., those who completed LLAMA professional development in Years 3 and 4 of the project). This includes 12 Cohort 2 teachers who were active as of the end of Year 4 of the project. To ensure alignment in terms of data points between this analysis and the prior analysis, RMC Research used Cohort 2's second Pre (at the end of Year 2) and second Post (at the end of Year 4) datapoints. See Exhibits 108-109.

Exhibit 108. Pre-Post Frequencies of Active Cohort 2 Teachers

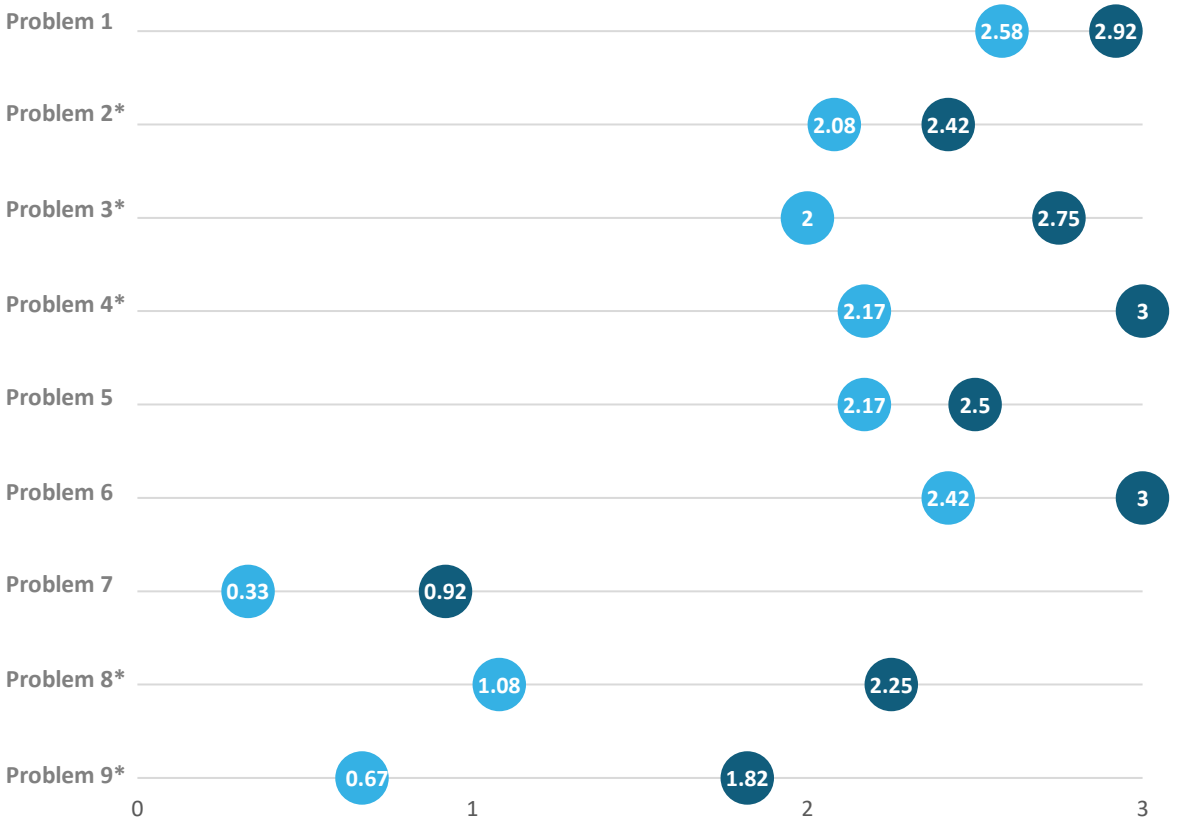
		Rating Scale Score									
		0		1		2		3			
Item	Problem	*n*	%	*n*	%	*n*	%	*n*	%	*M*	*sd*
1	Pre	1	8%	0	0%	2	17%	9	75%	2.58	0.900
	Post	0	0%	0	0%	1	8%	11	92%	2.92	0.289
2*	Pre	0	0%	3	25%	5	42%	4	33%	2.08	0.793
	Post	0	0%	1	8%	5	42%	6	50%	2.42	0.669
3*	Pre	2	17%	1	8%	4	33%	5	42%	2.00	1.128
	Post	0	0%	0	0%	3	25%	9	75%	2.75	0.452
4*	Pre	2	17%	1	8%	2	17%	7	58%	2.17	1.193
	Post	0	0%	0	0%	0	0%	3	100%	3.00	0.000
5	Pre	0	0%	2	17%	6	50%	4	33%	2.17	0.718
	Post	0	0%	0	0%	6	50%	6	50%	2.50	0.522
6	Pre	1	8%	2	17%	0	0%	9	75%	2.42	1.084
	Post	0	0%	0	0%	0	0%	12	100%	3.00	0.000
7	Pre	8	67%	4	33%	0	0%	0	0%	0.33	0.492
	Post	6	50%	2	17%	3	25%	1	8%	0.92	1.084
8*	Pre	2	17%	8	67%	1	8%	1	8%	1.08	0.793
	Post	1	8%	2	17%	2	17%	7	58%	2.25	1.055
9*	Pre	6	50%	4	33%	2	17%	0	0%	0.67	0.778
	Post	1	8%	4	33%	2	17%	4	33%	1.82	1.079

Note. $n = 12$. Rating Scale Scores: 0 = No elements of a viable argument. 1 = Limited elements of a viable argument. 2 = Elements of a viable argument. 3 = Viable argument

*Significant differences at $p < .05$. Differences assessed using paired t-tests. Results should be interpreted with caution due to the small sample size.

Exhibit 109. Pre-Post Mean Comparisons of Cohort 2 Active Teachers

Cohort 2 experienced increases from pre to post for all 9 items, with significant increases for Items 2, 3, 4, 7, 8, and 9. As with the original study, these data indicate that teachers that participate in the LLAMA intervention significantly increase their argumentation skills over time.



Note. ●**Pre** ●**Post** *n* = 12. Pre is the TARA taken just before beginning professional development and Post is the TARA taken after completion of professional development.
Rating Scale Scores: 0 = No elements of a viable argument. 1 = Limited elements of a viable argument. 2 = Elements of a viable argument. 3 = Viable argument.
*Significant differences at $p < .05$. Differences assessed using paired t-tests. Results should be interpreted with caution due to the small sample size.

Limitations and Considerations

There are a few **limitations** that should be taken into consideration. First, the TARA data is ordinal level data rather than interval level data which means the intervals between the TARA scores on the rubric are not necessarily evenly spaced (i.e., the space between a 0 and 1 score may not be the same as the space between a 2 and 3). Second, there is unequal variance in the data in these analyses and the data do not strictly adhere to a normal distribution. Third, although the scorers attained interrater reliability, it is always more rigorous to have at least two scorers score each TARA. Fourth, scorers used the same rubric for both student and teacher assessments, the UI team should consider whether a different scoring system should be used based on who took the assessment (teacher versus student)

Next Steps

Additional analyses the team is considering include an analysis of results by implementation category, incorporating “Read Correctly” scores into the analyses (only “Viable Argumentation” scores are considered in the analyses in this chapter), and comparisons of results with different analytic approaches (e.g., considering the data as categorical rather than continuous). For example, conducting tests to see if there is a significant increase in the percentage of teachers getting a ‘high’ score between pre and post. For this analysis the team would need to designate what qualifies as a high score.

Study 4: LLAMA Learning Progression Study Progress

Original design. There are 3 major components to the learning progression study design. In the first component, University of Idaho will gather classroom work or assessments from all treatment students. Treatment teachers will submit 13 pieces of student data from 13 time points from all students in Years 1, 2, and 3. The student work will address the 12 processes for students to master and 12 related conceptual pillars. In the second component, University of Idaho will draw a random sample of 10 treatment teachers to participate in an intensive case study. In Years 1, 2, and 3 these teachers will be observed and interviewed 3 times each year. Both research teams will complete a Classroom Argumentation Observation Protocol at each observation and videotape the observations. University of Idaho will interview the teachers using the Teacher Interview protocol and record the interviews. The recording and videotapes will allow for in-depth analysis. In the third component at the beginning of Years 1, 2, and 3, University of Idaho will draw a random sample of 10 treatment students each year. The students will each complete 12 Cognitive Task-Based Interviews (Ginsburg, 1997), which represents one interview for each of the processes/conceptual pillars expressed in the learning progression. Each interview is conducted immediately after the students' teacher implements a lesson associated with the process/conceptual pillar. The interviews will be videotaped and transcribed.

Utilizing all student data collected during the 3 components University of Idaho will use a methodology similar to Lobato et al. (2012) to address Research Question 4, "To what extent does treatment student learning align with that hypothesized in the LLAMA learning progression?" to assess the degree to which students' learning aligns with that hypothesized in the learning progression. Lobato, Hohensee, Rhodelhamel, & Diamond (2012) assert that learners might have rudimentary ways of coming to know and reason that are important for their development that have been forgotten by experts. These 12 conceptions become pivotal intermediate conceptions when they can be leveraged toward more sophisticated ways of reasoning. The majority of studies highlight differences between novices' ways of reasoning and proving and that of experts. To address Research Question 5, "What pivotal intermediate conceptions are important for Grade 8 students in developing viable argumentation conceptions and practices?" University of Idaho will use retrospective analysis of all teacher and student data collected during the 3 components to develop models of student conceptions at various time points, based on the methods of Miles, Huberman, & Saldaña (2013). This analysis draws upon frameworks for student thinking developed from previous iterations of the intervention and will be used to develop learning trajectories (Ellis, Weber, & Lockwood, 2014) that describe plausible paths through which students acquire more sophisticated thinking.

Year 1 Major Modifications. The LLAMA project was funded in September 2016 at the onset of the 2016–2017 school year, which meant the project could not begin the treatment with teachers in September. Rather, recruitment for the project began in September. Due to the arrival of the NSF funding 3 months later than proposed, it was not feasible to collect all Year 1 data for this study.

The original plan specified that 10 treatment teachers would be randomly selected to be case study teachers. Nine were randomly selected so that each coach worked closely with 3 teachers each. Due to the late project start, these teachers were not interviewed 3 times, nor were all observed 3 times in Year 1. Since 5 of the 9 case study teachers' circumstances changed (moving, reassignment, health issues), a new subset of treatment teachers was selected for follow-up in Year 3. These teachers were observed at least once.

Due to the late project start, case study students were not selected in Year 1: no Year 1 data collection related to this study occurred with these students. As noted in other chapters, student data were collected for other studies.

Year 2 Major Modifications. In Year 2 rather than randomly select case study students, 10 students from a single case study teacher's class were selected. This change was made to construct a richer data set. The team selected a teacher who was known to implement LLAMA with fidelity and whose location allowed the students to be interviewed by all of the UI PIs. Choosing students from one teacher known to be implementing the program with fidelity allowed the team to focus on the learning of students who had all received the treatment.

The proposal stated that UI would interview the case study students once for each conceptual pillar; however, to reduce the teacher burden, the 10 case students were interviewed 6 times during the school year. Multiple conceptual pillars were addressed in each interview.

Year 3 Major Modifications. No case study students were selected in Year 3.

The data collection plan specified in the proposal for all of the treatment students (13 pieces of data from every student corresponding to each of the original 13 pillars) was not feasible. Data collection was modified to a monthly submission with teachers submitting 3 student samples each month: 2 of the samples were student work demonstrating argumentation and the third represented a student's work on a pillar. These data were collected in Years 1 and

2. Student samples were collected in Year 3 using a new protocol developed to address logistical concerns and comments from the NAB.

Year 4 Major Modifications. Based on the success of the single case study teacher approach used in Year 2, this approach was used again in Year 4 instead of randomly selected case study students. Extensive coaching and planning was done with the case study teacher. Six students were selected as participants from his Grade 8 classes. The teacher chose them purposefully to provide what he thought would be a range of prior knowledge and ability. These students were interviewed over the course of the year-six of them were interviewed 7 times, two were interviewed 5 times. The interviews were designed to span all 12 conceptual pillars.

Details about study recruitment, power analyses, and attrition are described in the **LLAMA Participants** chapter.

Instrument Development

This study has 5 instruments: Student Work Sample Scoring Form, Classroom Argumentation Observation Protocol, Teacher Interview Protocol I, Teacher Interview Protocol II, and Cognitive Task-Based Interview Protocol. This section describes the instrument development process.

Student Work Samples and Work Sample Scoring Form

Though the monthly survey data are self-reported, the student work samples provide evidence of how well the treatment and control teachers understand argumentation, and how well the treatment teachers understand the LLAMA conceptual pillars.

Years 1-2

Each month the treatment teachers (Cohort 1) were asked via the monthly survey to electronically submit 3 student work samples:

- Sample 1: One picture of student work that shows a **rich understanding of the argument** practices that you focused on this month.
- Sample 2: One picture of student work that shows **partial understanding of the argument** practices you focused on this month.
- Sample 3: One picture of student work that illustrates the **primary conceptual pillar(s)** that you focused on this month.

Each month the control group teachers (Cohort 2) were asked via the monthly survey to electronically submit 2 student work samples:

- Sample 1: One picture of student work that shows a **rich understanding of the argument** practices that you focused on this month.
- Sample 2: One picture of student work that shows **partial understanding of the argument** practices you focused on this month.

Years 3–4

At three points during the Year 3 school year (October, January, and May) treatment and control teachers were asked to select an item from a quiz, test, etc. that addresses argumentation, identify the type of argument, and submit 3 student work samples:

- Sample 1: One picture of student work that shows a **limited understanding of the argument** in question.
- Sample 2: One picture of student work that shows **moderate understanding of the argument** in question.
- Sample 3: One picture of student work that illustrates the **strong understanding of the argument** in question.

Treatment and control teachers were also asked to write feedback that they would provide to the three students with regard to their use of argumentation. Year 3 changes were made to decrease teacher burden in a way that still enabled the research team to gain a rich understanding of teachers' comprehension of the different argument types and how they interact with their students. Student samples were collected from control teachers only in Year 4. Coaches used the work samples, and the way in which the teachers categorized them, to inform their coaching sessions with the teachers.

Classroom Argumentation Observation Protocol

> ***Target:*** *Develop a Classroom Argumentation Protocol in Year 1.* ***Status****: Met*

The Classroom Argumentation Observation Protocol developed in LAMP measures teachers' pedagogical practices in terms of teachers providing opportunities for students to engage in the mathematics learning experiences specified in the logic model. The protocol provides quantified scores for the types of claims a teacher uses, the explicitness of claims, the sophistication of the warranting, and the use of warrants and data. Open-ended questions ask for the extent to which the observed lessons address LLAMA lesson objectives. LAMP established content validity of this protocol through an expert panel. At the onset of the LLAMA project, the protocol was revised. The primary revisions include (a) the inclusion of our recent understanding of generic example arguments (see Yopp and Ely, 2016 and Yopp, Ely, and Johnson-Leung, 2016) and (b) asking about the percentage of students engaged in the classroom argumentation episode. The protocol was further revised during weekly Principal Investigator meetings and was piloted in Year 1. During Year 2, the research teams participated in an observation training to ensure interrater reliability among observers and maintain a codebook of decision rules pertaining to the coding of the observations. Minor modifications were made to the protocol during this training period. The team watched videos and scored the videos during multiple sessions and modified the rubric and wording accordingly, following those sessions.

Teacher Interview Protocols

> ***Target:*** *Develop a Teacher Interview Protocol in Year 1.* ***Status****: Met*

LLAMA provides support to grade 8 mathematics teachers in using an enhanced pedagogy to teach mathematics via viable argumentation, with the aim of making viable argumentation a daily feature of grade 8 mathematics instruction. In encouraging teachers to engage students in making claims about their solutions to problems, to explicitly support their claims using prior mathematics results, and to

communicate both their support and their mathematical insights to teachers and classmates, this enhanced pedagogy represents a departure from traditional mathematics pedagogies of demonstrating procedures and problem solving and asking students to practice processes they have observed. Such changes in practice are not easy to implement and some teachers do so more readily than others. With the aim of monitoring and understanding barriers and affordances to the practices of teaching through viable argumentation, LLAMA has planned to interview a collection of case study teachers periodically throughout years 1, 2, and 3 of the LLAMA project. During years 1 and 2, the case study teachers received training and coaching support. During year 3 the case study teachers received neither training nor coaching support; however coaches were available to answer questions and direct teachers towards resources upon request.

In Year 1 University of Idaho developed a **Teacher Interview Protocol I** to ascertain teachers' perspectives about why and how their teaching practices change, and barriers to change, with a focus on implementing teaching mathematics through argumentation. The interview questions were developed to address the research questions and align with the logic model. The Teacher Interview Protocol has scripted text at the beginning and end of the interview, and consent is obtained from interviewees. The questions are open-ended and expansive, and take about 30 minutes.

In Year 3 University of Idaho developed a second **Teacher Interview Protocol II** to measure teachers' perspectives regarding facilitators and barriers to LLAMA implementation [teaching mathematics through viable argument]. The Teacher Interview Protocol has scripted text at the beginning and end of the interview, and consent is obtained from interviewees prior to the interview. The questions are open-ended and expansive, and takes about 60 minutes.

Student Cognitive Task-Based Interviews

> ***Target:*** *Develop Cognitive Task-Based Interview Protocol in Year 1.* ***Status****: Met*

Original Proposed Activity

The Cognitive Task-Based Interviews developed in LAMP measure student thinking and behaviors related to the LLAMA learning progression. During the LAMP project, the Cognitive Task-Based Student Interview Protocol was developed, piloted, field tested, and refined. A clinical interview protocol will be developed specifically to capture thinking related to student growth trajectories in their ability to know about and use viable argumentation and content knowledge as specified in the proposal. Clinical interviews akin to the task-based interviews described in Goldin (1997, 2000) will be conducted. Participants will be interviewed individually or in pairs immediately after (within 1 week) each lesson in the learning trajectory. The interviews will be used to validate classroom observation data and student work assessment and to assess the nature and growth of students' argumentation understanding and practices as described in the learning trajectory. Data on student understanding and emerging practices while interacting with the teacher/researcher and project materials and whether or not lesson objectives are met will be collected. Ultimately, these data will assess the plausibility of the learning trajectory as a model for how students' progress in argumentation knowledge, skills, and practices in the context of early algebra instruction.

Two types of task-based clinical interviews will be used in each episode with students. Type 1 will focus on the student work developed in class projects or in the online environments. Students will be given a sample of their work on the previous episode and asked scripted questions based on the analysis of this work. Once the scripted questions are complete, open-ended questions based on the student responses will be asked (see Szilágyi, Clements, & Sarama, 2013), for a similar approach). Data from this part of the

interview will triangulate observation and student work data and will offer the researcher opportunities to note unanticipated conceptions and misconceptions.

The second type of task-based interview (Type 2) will present a similar, parallel task to the one used in the lesson and in the first part of the interview. This interview will assess the efficacy of the lesson on similar tasks addressed by the student individually. Because much of the lessons involves group work, this second task will add validity to assertions about the degree to which the learning progress accurately describes a plausible model for students' growth in creating viable arguments and critiquing the arguments of others in the context of early algebra. This interview will be a "think aloud" as the participant addresses the task and will begin with a script, followed by enhancement akin to Piaget's method of clinical interviewing (Ginsburg, 1997). To ensure that open-ended interactions do not influence the participant's responses to the tasks, the open-ended probes will be administered only after the scripted interview is complete. Such methods have been employed by Clements et al. (2004) and Szilágyi, Clements, and Sarama (2013).

Year 2 Modifications

In Year 2, the proposed activities were modified so that only Type 2 task-based interviews were used. This decision was made to collect research data on students' argument practices. The team was successful in creating a sequence of rich tasks that were able to draw out students' extant argument knowledge and practices. The Type 1 interviews occurred naturally during the Type 2 interviews, because students were asked to reflect on the arguments they produced during the Type 2 activities and to discuss their problem-solving and argumentation approaches and thinking involved in producing their responses. No modifications were made in Year 3 and no interviews were conducted in Year 3.

Year 4 Modifications

In Year 4, the proposed activities were modified so that only Type 2 task-based interviews were used. The sequence of interview protocols from Year 2 was modified for Year 4, in light of the classroom lessons taught by the case study teacher. As was true in Year 2, the Type 1 interviews occurred naturally during the Type 2 interviews, because students were asked to reflect on the arguments they produced during the Type 2 activities and to discuss their problem-solving and argumentation approaches and thinking involved in producing their responses.

Data Collection

Student Work Samples Completion

> ***Target:*** *University of Idaho will work with treatment teachers to submit 12 pieces of* ***student data*** *from 12 time points from all students in the treatment teachers' classroom in Years 1, 2, and 3. Modified to 3* ***Student Work Samples*** *each month and at least one sample per conceptual pillar in Years 1 and 2. Year 3 and 4 was further modified to 3 student work samples at 3 points during the school year (October, January, May).* ***Status:*** *Partially Met*

Exhibit 110 shows Year 1 completion rates for student work samples by month as of May 31, 2017. Exhibit 111 shows completion rates for student work samples by month as of May 31, 2018. Completion rates represent those teachers who either submitted a complete data set or who indicated they did not have student samples to upload due to not addressing argumentation in their mathematics classes. December completion rates were low due to teachers' lack of familiarity with the process of uploading

student samples and inclement weather which resulted in school cancellations. Exhibit 112 shows the completion rates for Year 3 and Exhibit 113 shows the completion rates for Year 4.

Exhibit 110: Year 1 Completion Rates: Student Samples

	Cohort 1			Cohort 2		
Survey Month	***n***	**Samples**	**Completion**	***n***	**Samples**	**Completion**
December 2016	29	11	38%	27	16	59%
January 2017	29	19	66%	27	22	81%
February 2017	29	21	72%	26	25	96%
March 2017	29	23	79%	26	23	88%
April 2017	28	21	75%	27[a]	24	89%
May 2017	28	19	68%	27	26	96%

Note. n = the total number of active participants at the time of the survey administration. Non-RCT teachers are included in this table.
[a]One non-RCT teacher joined the project in April 2017.

Exhibit 111: Year 2 Completion Rates: Student Samples

	Cohort 1			Cohort 2		
Survey Month	***n***	**Samples**	**Completion**	***n***	**Samples**	**Completion**
September 2017	25	21	84%	20	18	90%
October 2017[a]	25	21	84%	19	16	84%
November 2017[b]	25	20	80%	18	18	100%
December 2017[c]	25	21	84%	21	20	95%
January 2018	25	22	88%	21	19	90%
February 2018	25	20	80%	21	20	95%
March 2018	25	23	92%	21	20	95%
April 2018	25	21	84%	21	19	90%
May 2018	25	21	84%	21	19	90%

Note. n = the total number of active participants at the time of the survey administration. Non-RCT teachers are included in this table.
[a]One Cohort 2 teacher left the project in October 2017.
[b]One Cohort 2 teacher left the project in November 2017.
[c]As of January 30, there are 5 non-RCT teachers who are active in the project.

Exhibit 112: Year 3 Completion Rates: Student Samples

	Cohort 1			Cohort 2		
Survey Month	***n***	**Samples**	**Completion**	***n***	**Samples**	**Completion**
October 2017[a]	22	4	18%	21	15	71%
January 2018	22	6	27%	20	13	65%
April 2018	22	9	41%	20	16	80%

Note. n = the total number of active participants at the time of the survey administration. Non-RCT teachers are included in this table.

Exhibit 113: Year 4 Completion Rates: Student Samples

	Cohort 2		
Survey Month	***n***	**Samples**	**Completion**
October 2019[a]	12	8	67%
January 2020	12	9	75%

Note. n = the total number of active participants at the time of the survey administration. Non-RCT teachers are included in this table. A third set of samples were not requested from teachers due to COVID-19 closures.

Classroom Argumentation Observation Protocol Completion

> ***Target:*** *University of Idaho and RMC Research will conduct observations on all treatment teachers* ***twice a year*** *in Years 1, 2, and 3. University of Idaho will observe case study teachers 3 times each year.* ***Status:*** *Partially Met*

As shown in Exhibit 114, the research team did not observe each treatment teacher twice in Year 1 and did not observe each case study teacher 3 times in Year 1: 59% of the intent-to-treat teachers were observed twice and 71% of the active teachers were observed twice. One of the case study teachers was observed 3 times in Year 1. The LLAMA coaches conducted a total of 53 observations of the Cohort 1 teachers during Year 1. All of the active teachers were observed at least once during the 2016–2017 school year. Only 1 of the teachers who dropped before the end of Year 1 was observed.

Exhibit 114: Observation Completion: Year 1 RCT Intent-to-Treat Completion Rates

Observation	*Intent to Treat*	*Active Cohort 1 Teachers*	*Case Study Teachers*
0 observations	5	0	0
1 observation	7	6	2
2 observations	20	20	6
3 observations	2	2	1
Total teachers	**34**	**28**	**9**

Note. All Cohort 1 teachers: *n* = 34. Case Study teachers: *n* = 9. Only 28 teachers were active as of May 31, 2017.

As shown in Exhibit 115, the research team observed most RCT treatment teachers at least twice in Year 2: 74% of the intent-to-treat teachers were observed at least twice, and 100% of the active teachers were observed at least twice. All 9 case study teachers were observed 3 times in Year 2. The LLAMA coaches conducted a total of 262 observations of the Cohort 1 teachers during Year 2.

Exhibit 115: Observation Completion: Year 2 RCT Intent-to-Treat Completion Rates

Observation	*Intent to Treat*	*Active Cohort 1 Teachers*	*Case Study Teachers*
0 observations	9	0	0
1 observation	0	0	0
2 observations	3	3	0
3 observations	1	1	1
4 observations	1	1	0
5 observations	2	2	2
6 observations	2	2	1
7 observations	0	0	0
8 observations	2	2	1
9 observations	3	3	1
10 or more observations	11	11	3
Total teachers	**34**	**25**	**9**

Note. All Cohort 1 teachers: *n* = 34. Case Study teachers: *n* = 9. Only 25 teachers were active as of May 31, 2018.

Exhibit 116 shows the observations completed in Years 3 and 4. Nine treatment teachers were observed in Year 3: 2 of the nine teachers were observed twice and the rest were only observed once.

Observations for control teachers in Year 3 are shown in Exhibit 116 below. The research team observed 14 of the 31 RCT control teachers at least twice (45%). The LLAMA coaches conducted a total of 115 observations of the Cohort 2 teachers during Year 3 and 84 observations during Year 4. Ninety-two percent of active Cohort 2 teachers were observed at least twice.

Exhibit 116: Observation Completion:
Year 3 and 4 RCT and Non-RCT Control Teacher Completion Rates

	Year 3		Year 4	
Observation	**Intent to Treat[a]**	**All Active[b] Cohort 2 Teachers**	**Intent to Treat[a]**	**All Active[c] Cohort 2 Teachers**
0 observations	16	0	0	1
1 observation	1	1	0	0
2 observations	1	1	0	0
3 observations	1	2	0	0
4 observations	0	1	0	0
5 observations	2	3	2	4
6 observations	1	1	0	0
7 observations	1	1	2	2
8 observations	2	3	0	0
9 observations	1	1	0	0
10 or more observations	5	5	5	5
Total teachers	**31**	**19**	**9**	**12**

Note. All Cohort 1 teachers: *n* = 34. Case Study teachers: *n* = 9.
[a]"Intent to Treat" only includes RCT control teachers.
[b]Active as of May 31, 2019. Includes both RCT and non-RCT control teachers (14 RCT; 5 non-RCT).
[c]Active as of May 31, 2020. Includes both RCT and non-RCT control teachers (9 RCT, 3 non-RCT)

Teacher Interview Completion

Target: *In Years 1, 2, and 3 University of Idaho will interview 10 case study teachers 3 times each year using the* ***Teacher Interview Protocol****.*

Modified Target Year 2: *In Years 1, 2, and 3 University of Idaho will interview 9 case study (3 interviewed by each lead coach) teachers 3 times each year using the* ***Teacher Interview Protocol****.* ***Status:*** ***Partially Met***

Modified Target Year 3: Because several of the case study teachers have become inactive, reporting less use of teaching via viable argument during Year 3 than in the previous year when coaching support was provided, the LLAMA research team created a modified interview

protocol Teacher Interview II to address teachers' perspective on affordances and barriers to teaching with viable argument. Teachers to interview were selected with differing categories of implementation, based on self-report and coach rating and with different levels of mathematics knowledge, as measured by MKT.

The research team did not interview any teachers in Year 1. All 9 case study teachers were interviewed by coaches 3 times in Year 2. RMC Research conducted Year 3 interviews with 12 teachers July-October 2019. These teachers were selected so that there was variation in terms of LLAMA implementation—both low and high implementers were selected to get a sense of facilitators and barriers to implementation. All interviews will be recorded and transcribed at UI for analyses in Year 5. No teacher interviews were conducted in Year 4.

Student Cognitive Task-Based Interview Completion

Target: *University of Idaho will conduct* ***Cognitive Task-Based Interviews*** *with 30 treatment students:* ***10 in Years 1, 2, and 3****. Each year 13 interviews will be conducted with each student, one interview for each process/conceptual pillar expressed in the learning progression. Each interview is conducted immediately after their teacher implements the signature lesson associated with the process/conceptual pillar. Interviews will be videotaped and transcribed.*
Status: ***Partially Met***

Due to the late start of the grant these interviews did not occur in Year 1. To reduce teacher burden, multiple conceptual pillars were included in each student interview. Student cognitive task-based interviews were conducted 6 times during the 2017–2018 school year. All 10 case study students were interviewed 6 times, except 1 student who was only interviewed 5 times (Exhibit 117). No students were interviewed in Year 3; this is because it was judged that, as Cohort 2 was just beginning, the implementation category in the classroom of the signature lessons associated with each conceptual pillar was not sufficiently reliable. In Year 4, 6 students were interviewed 7 times, except 2 students who were interviewed 5 times (Exhibit 118).

Exhibit 117: Student Cognitive Task-Based Interview Year 2 Completion Rates

Interview	*Number of Students Interviewed*
September 2017	10
Interview 7 (spring 2020)	10
February 2018	10
March 2018	10
April 2018	9
May 2018	10

Exhibit 118: Student Cognitive Task-Based Interview Year 4 Completion Rates

Interview	*Number of Students Interviewed*
October 1, 2019	6
October 17, 2019	6
November 13, 2019	6
December 5, 2019	6
January 28, 2020	6
May 2020 (various dates due to COVID-19)	4
June 2020 (various dates due to COVID-19)	4

Analysis and Findings: Student Work Samples

In general, the student work samples were not systematically analyzed, but were instead used as formative assessment to inform coaching with the teachers. Examples were drawn from some of these student work samples and analyzed to illustrate types of student reasoning related to the conceptual pillars. These data have been used in several project publications.

Analysis and Findings: Observations

Classroom observation is one component of the LLAMA Learning Progression Study (Study 4). This section describes the findings from observations conducted in Years 1 through 4. Details about study recruitment and attrition are described in the **Student Achievement Chapter**. In the initial proposal, observations were to be conducted twice per year for Cohort 1 teachers in Years 1, 2, and 3, with an additional third observation for randomly-selected case study teachers. However, due to many early changes in the research studies' data collection plans, the team agreed that conducting an observation with each coaching visit would provide context and additional information to the primary analyses. Because Cohort 1 teachers began the professional development in Year 1, and Cohort 2 teachers delayed entry into the professional development until Year 3, Cohort 1 teachers were observed in Years 1, 2, and 3, whereas Cohort 2 teachers were only observed in Years 3 and 4. The results for Years 1-3 were originally reported in the prior year's report. This report chapter includes the results for Year 1-4. Sample sizes were too small to conduct analyses to account for teacher variance of LLAMA implementation fidelity.

Analytic Sample

The analytic sample for the observation analyses includes observations from all teachers (both RCT and non-RCT) who participated in the LLAMA professional development and have at least one classroom observation, summarized below in Exhibit 119. Teachers were observed the least during Year 1 due to the late start of the project. Observations per teacher averaged 12 over the course of the project.

Exhibit 119: Analytic Sample for Observation Analyses

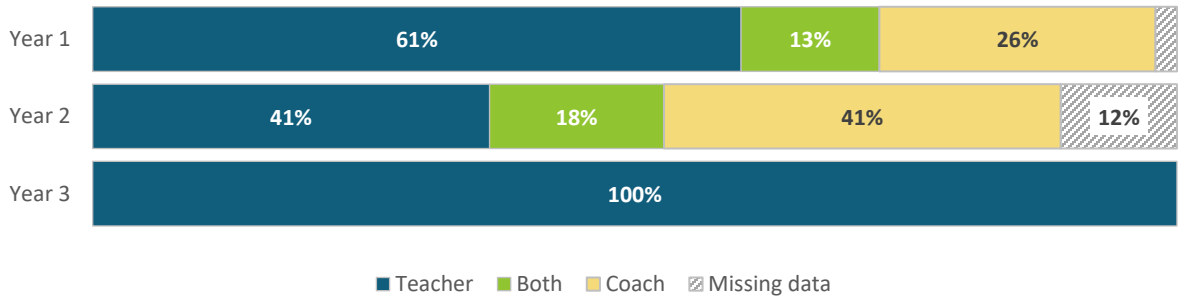
Project Year	Project Year Observations (*n*)	Unique Teachers Observed (*n*)	Observations per Teacher (*M*)
Year 1: 2016-2017	47	28	2
Year 2: 2017-2018	275	25	11
Year 3: 2018-2019	202	30	7
Year 3: Cohort 1	11	9	1
Year 3: Cohort 2	191	21	9
Year 4: 2019-2020	100	11	9
Total	**624**	**51**	**12**

Note. Year 3 is the only year in which both cohorts were observed. Only Cohort 1 was observed during Years 1 and 2, and only Cohort 3 was observed during Year 4.

Descriptive Summary of Observed Classes

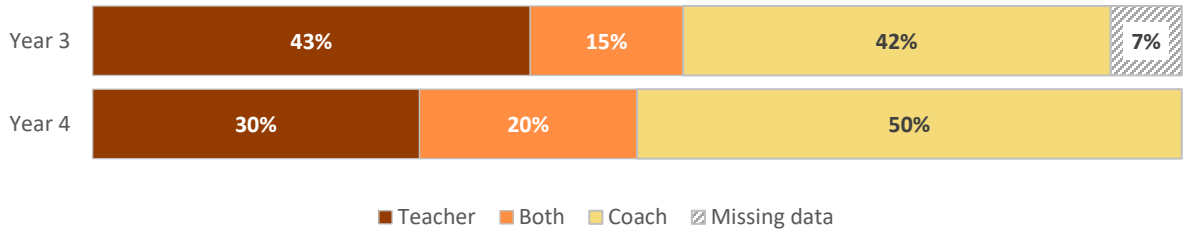
The average class size for observed classes was around 20 students. Most observed classes were labeled Grade 8 Math, although a few Pre-Algebra, Algebra I, Geometry, Grade 7 Math, and Intervention classes were also observed. For each active teacher, coaches were instructed to conduct at least one fall and one spring observation of a class where only the teacher taught the class (i.e., the coach did not assist in instruction). However, in practice data coded as "fall" or "spring" observations include a mix of teacher taught, coach taught, and combination classes. Exhibit 120 shows the Cohort 1 observations and Exhibit 121 shows the Cohort 2 observations. Frequencies in who taught the class varied by year for both cohorts; teacher-led observations ranged from 30% to 100%.

Exhibit 120: Observations by Who Taught the Class, Cohort 1



Note. The *n*'s count the number of observations included in the exhibit. **Cohort 1: Year 1:** *n* = 46 (*n* = 1 missing, 2%); **Year 2:** *n* = 241 (*n* = 34 missing, 12%); **Year 3** *n* = 11 (*n* = 0 missing, 0%). Percentages may not total 100%, due to rounding.

Exhibit 121: Observations by Who Taught the Class, Cohort 2



Note. The *n*'s count the number of observations included in the exhibit. **Cohort 2: Year 3:** *n* = 177 (*n* = 14 missing, 7%) **Cohort 2: Year 3:** *n* = 100. Percentages may not total 100%, due to rounding.

RMC planned on conducting a descriptive analysis for only teacher-led classes at a fall and spring time-point, however as Exhibit 122 shows, the sample size is low and there are only four teachers who were observed teaching at both the fall and spring time-point.

Exhibit 122: Fall and Spring Teacher-Led Observations by Cohort and Year

	Fall	Spring	No. of Teachers Observed both Fall and Spring
Cohort 1			
Year 1	0	4	0
Year 2	14	13	1
Year 3	6	2	1
Cohort 2			
Year 3	7	6	2
Year 4	2	1	0

Conceptual Pillars

Observations captured the LLAMA Conceptual Pillars observed during the class (see Exhibit 123).

Conceptual Pillars were more consistently implemented during each cohort's second implementation year. During the second implementation year at least 88% of the observations addressed a conceptual pillar. Conceptual Pillars 1, 2, 3, 4 and 8 were each observed for at least one class in each cohort and year. Conceptual Pillar 1 was the most frequently observed pillar, while Conceptual Pillar 12 was the least frequently observed.

Exhibit 123: Conceptual Pillars Observed, by Year and Cohort

	Cohort 1			Cohort 2	
Conceptual Pillar	**Year 1**	**Year 2**	**Year 3**	**Year 3**	**Year 4**
Pillar 1	**23%**	**31%**	**46%**	**45%**	**32%**
Pillar 2	**11%**	**28%**	**18%**	**30%**	**36%**
Pillar 3	**9%**	**30%**	**9%**	**26%**	**17%**
Pillar 4	**2%**	**11%**	**9%**	**9%**	**15%**
Pillar 5	0%	**10%**	**27%**	**9%**	**10%**
Pillar 6	0%	**11%**	0%	**3%**	**15%**
Pillar 7	0%	**10%**	0%	**6%**	**17%**
Pillar 8	**4%**	**16%**	**18%**	**15%**	**26%**
Pillar 9	0%	**12%**	0%	**2%**	**14%**
Pillar 10	**2%**	**29%**	0%	**12%**	**18%**
Pillar 11	0%	**18%**	0%	**7%**	**7%**
Pillar 12	0%	**6%**	0%	**4%**	**2%**
None addressed	**53%**	**11%**	**36%**	**9%**	**10%**

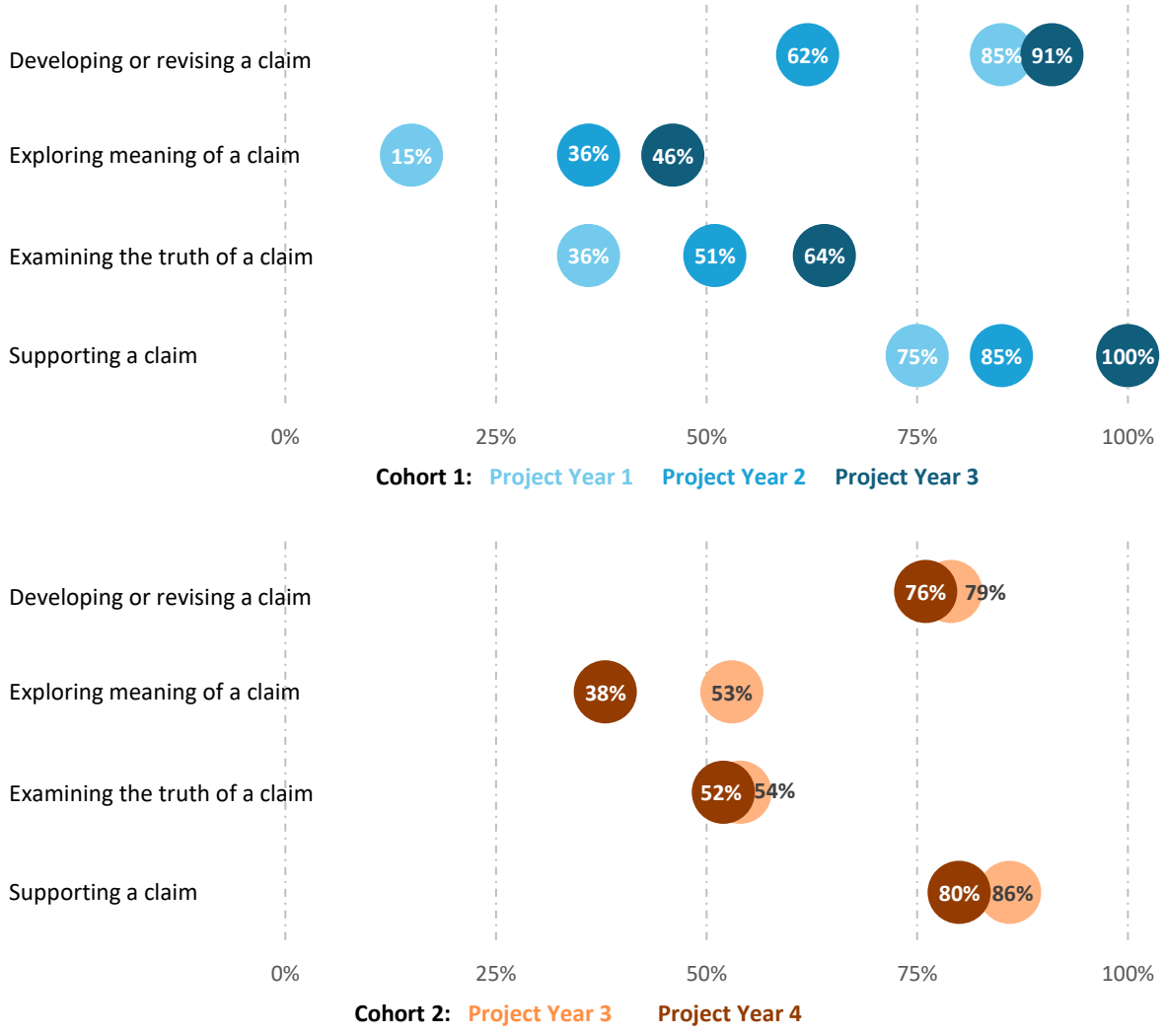
Note. The *n*'s count the number of observations included in the exhibit. **Cohort 1: Year 1:** *n* = 47; **Year 2:** *n* = 275; **Year 3:** *n* = 11. **Cohort 2: Year 3:** *n* = 191. **Cohort 2: Year 4:** *n* = 100

Claims for Argument Episodes

The Observation Protocol instructs the rater to focus on one argumentation episode (e.g., overarching reasoning type) observed during the class. The first 4 items in this section describe the observed claim: the **nature** of the claim, the **type** of claim, the **explicitness** of the claim, and the **clarity** of the claim.

In terms of the nature of the claim, percentages increased over time for each nature for Cohort 1, while percentages decreased for Cohort 2. The LLAMA team should consider possible reasons for this difference between cohorts. "supporting a claim" was observed for most argument episodes (75% or more of episodes in any given cohort and year).

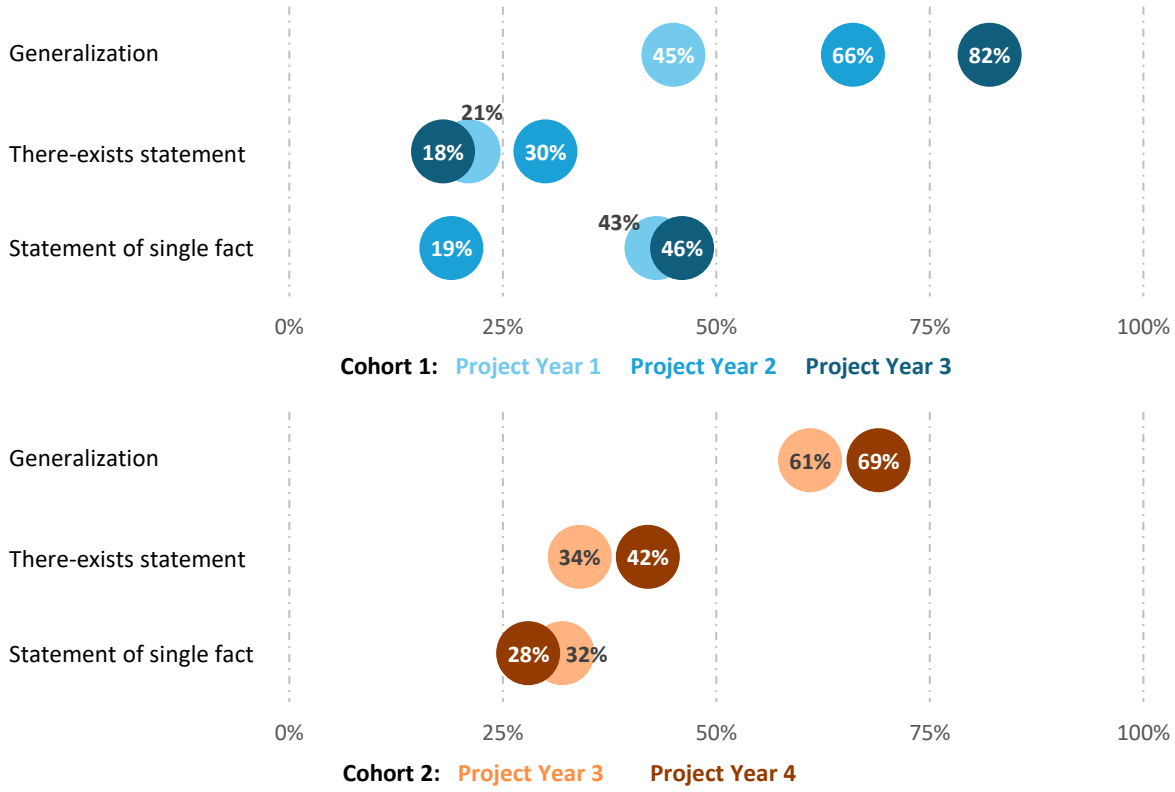
Exhibit 124: Nature of the Claim Observed in the Argument Episode



Note. The *n*'s count the number of observations included in the exhibit. **Cohort 1: Year 1:** *n* = 47; **Year 2:** *n* = 275; **Year 3:** *n* = 11. **Cohort 2: Year 3:** *n* = 191 Cohort 2: **Year 4:** *n* = 100.

There were no discernable patterns across time for either cohort in terms of type of claim observed. Generalization was the most frequently observed type of claim in all years for both cohorts.

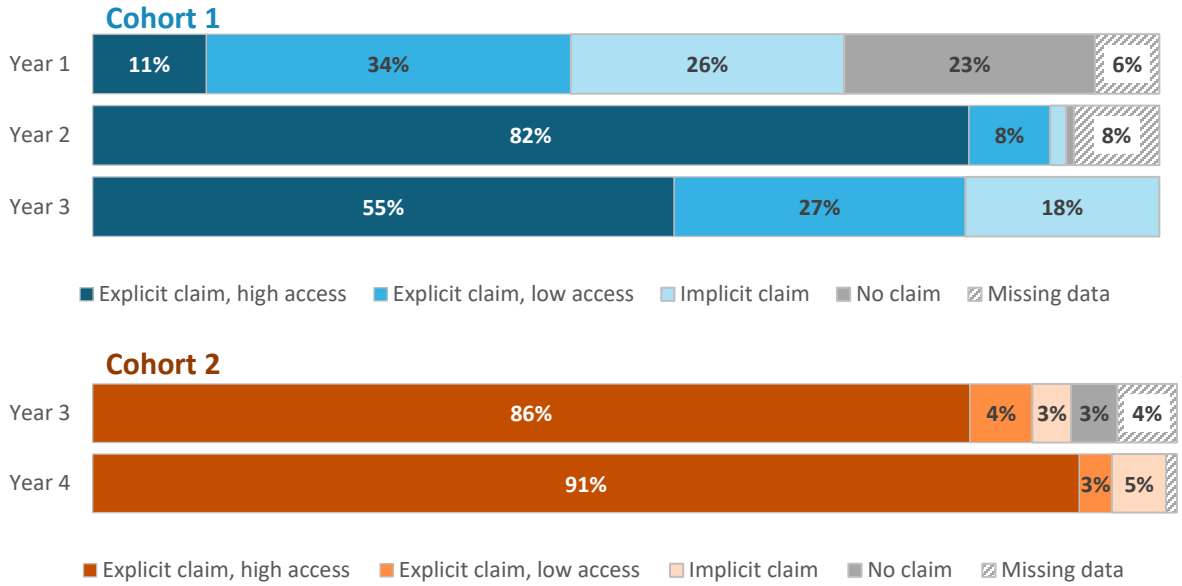
Exhibit 125: Type of Claim



Note. The *n*'s count the number of observations included in the exhibit. **Cohort 1: Year 1:** *n* = 47; **Year 2:** *n* = 275; **Year 3:** *n* = 11. **Cohort 2: Year 3:** *n* = 191 Cohort 2: **Year 4:** *n* = 100.

In terms of Explicitness of Claim, Cohort 1 had the highest percentage of high-access explicit claims during Year 2 (82%). The vast majority of Cohort 2 observations were rated high-access explicit claims for both Years 3 (86%) and Year 4 (91%).

Exhibit 126: Explicitness of Claim

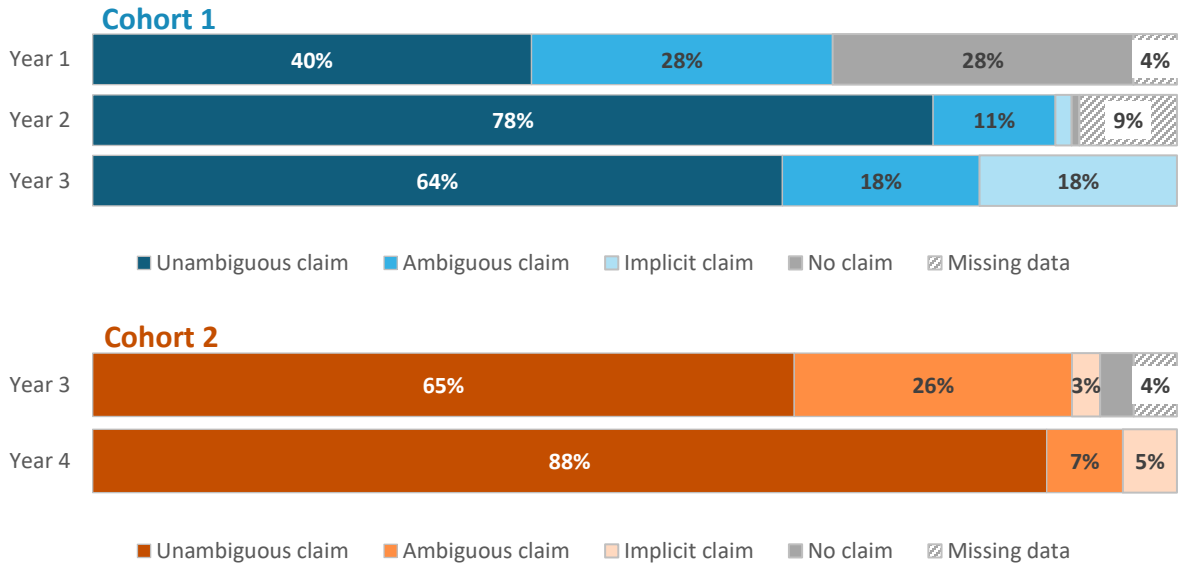


Note. The *n*'s count the number of observations included in the exhibit. **Cohort 1: Year 1:** *n* = 44 (*n* = 3 missing, 6%); **Year 2:** *n* = 252 (*n* = 23 missing, 8%); **Year 3:** *n* = 11 (*n* = 0 missing, 0%). **Cohort 2: Year 3:** *n* = 183 (*n* = 8 missing, 4%). **Year 4:** *n* = 99 (*n* = 1 missing, 1%). Percentages may not total 100%, due to rounding.

In Year 2 "Clarity of Claim" was recoded from a 3-point rubric (0, 1, 2) into a 4-point rubric (0, 1, 2, 3): "ambiguous claim" (originally scored as "1") was divided into "ambiguous claim" (new score of "2") and "implicit claim" (new score of "1").

The clarity of observed claims nearly doubled from Year 1 to Year 2 for Cohort 1 (40% in Year 1, 78% in Year 2). Unambiguous claims were observed in 65% of observations for Cohort 2 in Year 3, which increased to 88% in Year 4.

Exhibit 127: Clarity of Claim



Note. The *n*'s count the number of observations included in the exhibit. **Cohort 1: Year 1:** *n* = 45 (*n* = 2 missing, 4%); **Year 2:** *n* = 250 (*n* = 25 missing, 9%); **Year 3:** *n* = 11 (*n* = 0 missing, 0%). **Cohort 2: Year 3:** *n* = 184 (*n* = 7 missing, 4%); **Year 4**: n = 100. Percentages may not total 100%, due to rounding.

Argument Type and Support for Observed Argument Episodes

The last section of the Observation Protocol asks coaches to select the argument type for the observed argument episode and to rate support accompanying the selected argument type: the rater first circles the argument type(s) observed and then rates the support for the corresponding argument type, on a scale of 0 to 3, with 0 being low and 3 being high. The exact rubric criteria for scores of 0, 1, 2, or 3 vary by argument type. It should also be noted that (a) there is no rubric to score support for argument type b: non-constructive argument for existence on the protocol, (b) Argument type h: argument for claim of specific fact was added to the protocol in Year 2, and (c) Support scores for argument type d: direct argument are broken out into two subcategories (generic example and other direct argument). For this analysis, rubric responses were recoded as a dichotomous variable of '3' or 'less than 3' to provide a clearer picture of the support scores.

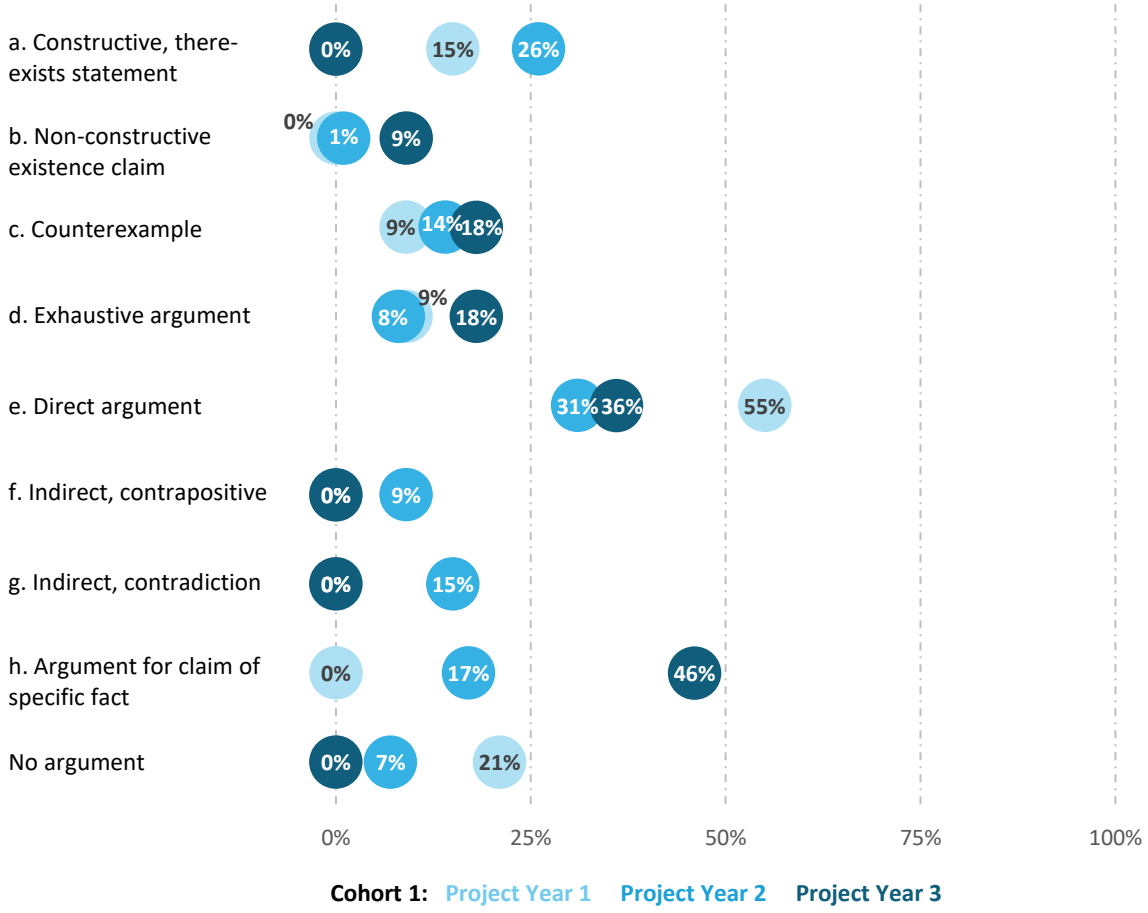
For analyses based on the raw rubric scores, see Appendix A.

Cohort 1

Exhibit 128 provides the percentage of observations by argument type by year for Cohort 1. It should be noted only 11 observations were conducted for Cohort 1 during Year 3. The Year 3 observations for Cohort 1 capture a glimpse of the classroom after completing all of the LLAMA professional development. Cohort 1 teachers did not receive any professional development in Year 3, and in some cases, the quality of argumentation was higher with coach support in Year 2 than without coach support in Year 3. It is important to consider that the sample size for Cohort 1 in Year 3 is extremely small--only 2 teachers in the analytic sample were observed more than once.

In Year 2 every type of argument was observed. Across all years, "direct argument" was the most common type of argument observed.

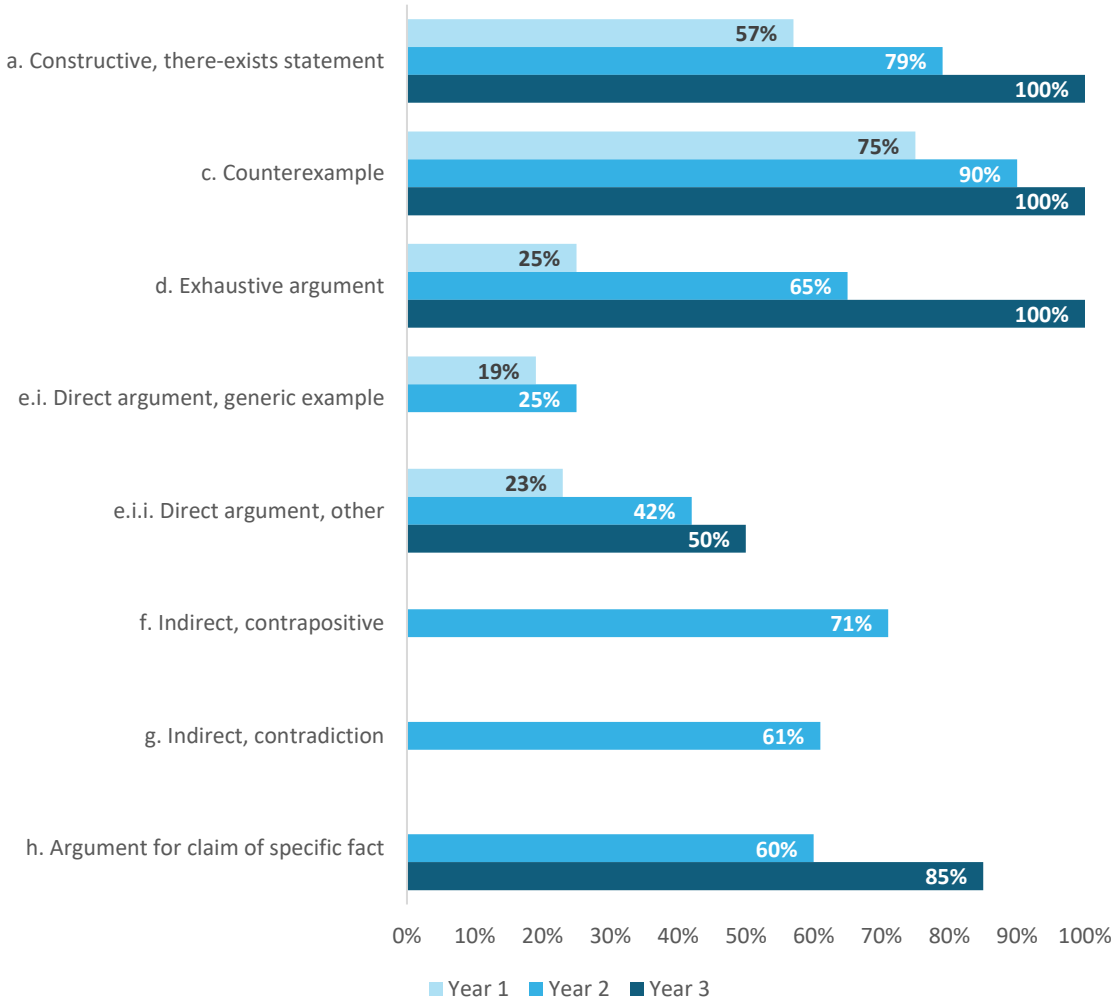
Exhibit 128: Argument Type(s) for Observed Argument Episode, Cohort 1



Note. The *n*'s count the number of observations included in the exhibit. **Cohort 1: Year 1:** *n* = 47; **Year 2:** *n* = 275; **Year 3:** *n* = 11.

Exhibit 129: Percentage of Observations Receiving a High Support Score, Cohort 1

High support ratings increased for all argument types over time for Cohort 1. Results should be interpreted with caution due to the small sample size in Year 3 (i.e., no argument type was observed more than 5 times).

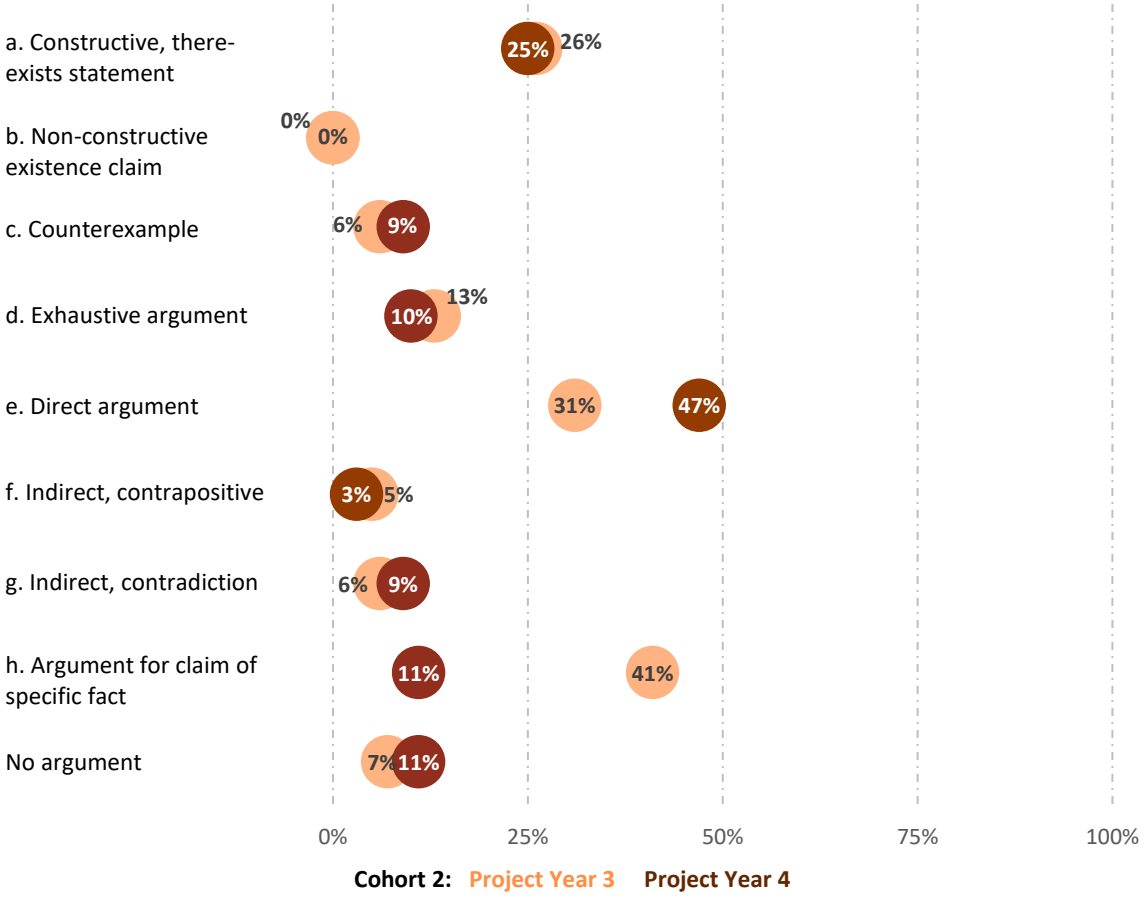


Note. Support for the corresponding argument type, was rated on a scale of 0 to 3, with 0 being low and 3 being high. Percentages in exhibit reflect the percentage of observations that received a high score of '3' for each argument type. The denominator for rubric score percentages is the number of observations that included the selected argument type. For robust analyses of these items, at least 5 observations for each rating for each type are needed, and the only type to exceed 20 observations is "direct argument," which is further divided into 2 support scores. Percentages may not total 100%, due to rounding.
Year 1: **a:** *n* = 7; **b:** *n* = 0; **c:** *n* = 4; **d:** *n* = 4; **e:** *n* = 26; **f:** *n* = 0; **g:** *n* = 0; **h:** *n* = 0.
Year 2: **a:** *n* = 72; **b:** *n* = 3; **c:** *n* = 39; **d:** *n* = 23; **e:** *n* = 85; **f:** *n* = 24; **g:** *n* = 41; **h:** *n* = 17.
Year 3: **a:** *n* = 1; **b:** *n* = 0; **c:** *n* = 2; **d:** *n* = 2; **e:** *n* = 4; **f:** *n* = 0; **g:** *n* = 0; **h:** *n* = 5.

Cohort 2

Each argument type was rated for at least one observation with the exception of "non-constructive existence claims" for Cohort 2. "Argument for claim of specific fact" was most frequently observed in Year 3 while "direct argument" was most frequently observed in Year 4. It is important to note that in Year 4 observations were not conducted in the final months of the school year due to the pandemic; yet that is the time period when teachers in prior years often felt ready to try indirect arguments

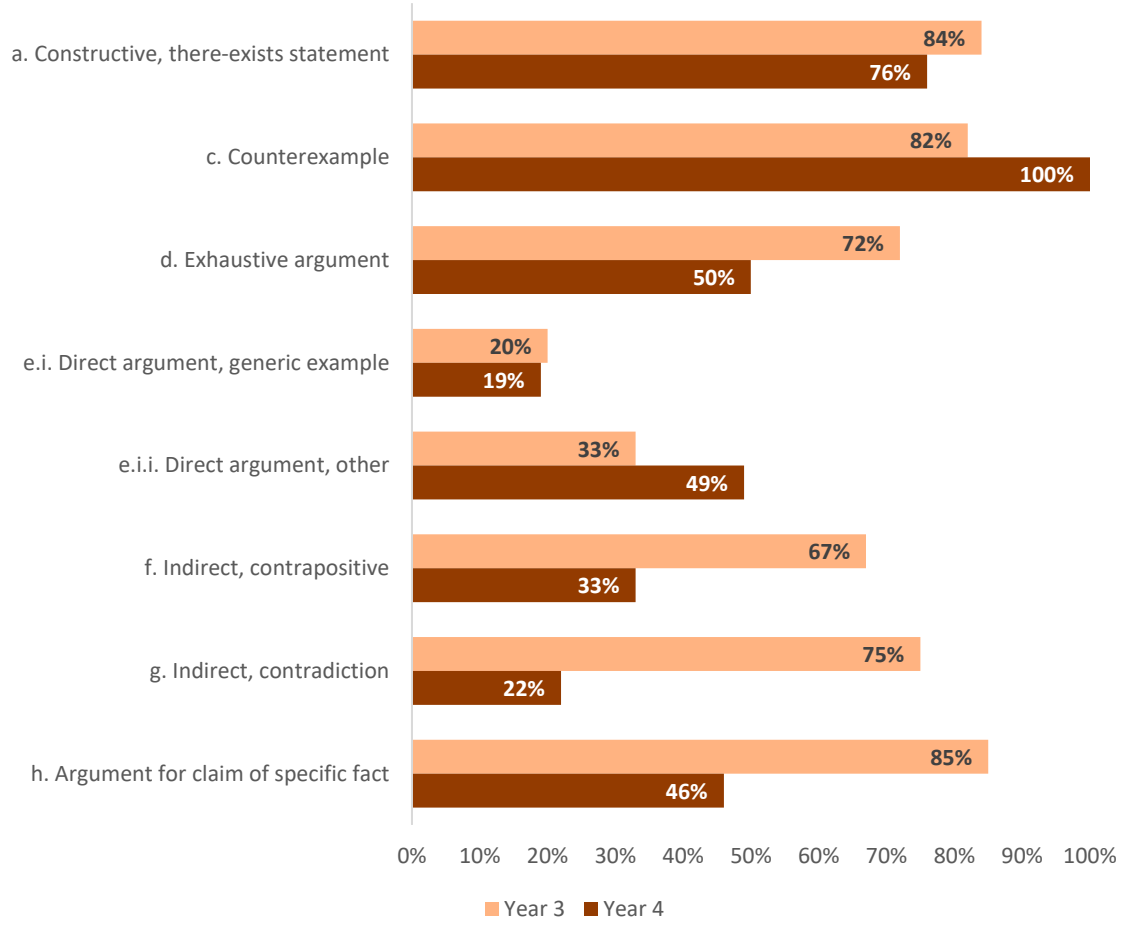
Exhibit 130: Argument Type(s) for Observed Argument Episode, Cohort 2



Note. The *n*'s count the number of observations included in the exhibit. **Cohort 2: Year 3:** *n* = 191; **Year 4:** *n* = 100.

The percentage of observations receiving a high score decreased for 6 of the 8 arguments over time. Results should be interpreted with caution due to the small sample size. It is important to note that observations did not occur during the later part of the school year during Year 4 due to the pandemic.

Exhibit 131: Percentage of Observations Receiving a High Support Score, Cohort 2



Note. Support for the corresponding argument type, was rated on a scale of 0 to 3, with 0 being low and 3 being high. Percentages in exhibit reflect the percentage of observations that received a high score of '3' for each argument type. The denominator for rubric score percentages is the number of observations that included the selected argument type. Percentages may not total 100%, due to rounding.
Year 3: **a:** *n* = 49; **b:** *n* = 0; **c:** *n* = 11; **d:** *n* = 25; **e:** *n* = 60; **f:** *n* = 9; **g:** *n* = 12; **h:** *n* = 40.
Year 4: **a:** *n* = 25; **b:** *n* = 0; **c:** *n* = 9; **d:** *n* = 10; **e:** *n* = 47; **f:** *n* = 3; **g:** *n* = 9; **h:** *n* = 11.

Limitations and Recommendations

There are a number of limitations to these analyses:

- Overall, observations received high ratings; however, for robust analyses, either more variation in responses (at least 5 observations per response option) or a larger sample size would be needed. Although there were 54 teachers in the sample, the sample size was very small for the various aspects of the rubric . Ideally, each argument type should have at least 5 responses for

each score on the rubric (e.g., at least 20 counterexamples, at least 5 of which were scored 0, 5 were scored 1, 5 were scored 2, and 5 were scored 3).

- Some teachers were observed many times, while others were observed only a few times; therefore, analyses may be biased toward teachers who received more support.
- The analyses in this chapter include observations for teacher-led instruction, coach-led instruction, and classes taught by both the coach and the teacher, which may not provide an accurate picture of the class when the coach is not assisting in instruction. The sample size for teacher-led instruction was too small to conduct statistical analyses.
- The analyses in this chapter take time into account but again, due to the small sample size statistical analyses over time could not be conducted.

Analysis and Findings: Teacher Interviews

The two sets of interviews are currently being analyzed and will be used to determine barriers and affordances to implementing LLAMA argumentation practices in instruction. Preliminary results indicate that teachers found competing demands on classroom time, increased teacher workload [for planning to teach with argumentation], and curricula that are not compatible with implementing classroom argumentation to be barriers to implementing mathematical argumentation instruction. Teachers' lack of comfort with their own argumentation skills proved to be a barrier for some. A few teachers reported that teaching with argumentation was incompatible with their teaching approach. These teachers largely had a traditional teaching practice. Teachers also said that teaching mathematics through argumentation affords deeper thinking and understanding; increases student discourse, engagement, mathematical connections, and student achievement; and lays the foundation for more difficult mathematics topics in the future.

Analysis and Findings: Student Cognitive Task-Based

As previously reported in the Year 2 Annual Report, the team conducted task-based interviews with students of a particular, high-implementing Cohort 1 teacher to better understand how students acquire the viable argument conceptions described in our proposal (a.k.a., conceptual pillars) and to better understand other argument conceptions students might acquire as a teacher implemented the LLAMA framework and lessons. We interviewed 10 students at 6 time points spread across the year. Each interview occurred shortly after the teacher implemented lessons targeting the acquisition of particular conceptual pillars, though often more than one week after the lesson, in the order lain out in our proposal.

During Year 3, the analyses of these data occurred. This led to one paper now accepted with minor modifications for publication in the Journal of Mathematical Behavior and another paper under review. These papers report on a significant finding from our research: That students who experience our intervention develop sophisticated notions of viable argument and proof that coalesce around a notion of proof of a generalization as eliminating counterexamples. As our intervention progressed, we found that students exhibited behaviors associated with our framework including but not limited to:

1. Skepticism of the truth of claims that were not proved, wondering if a counterexample might exist.
2. Skepticism of proofs provided to them that did not understand, wondering if the proof actually eliminated the possibility of counterexamples.

3. Skepticism of proofs provided to them that they felt were flawed, such as using a result they did not use as prior or making an illogical inference, wondering if the proof actually eliminated the possibility of counterexamples.
4. Conceptions of domain appropriate argumentation, meaning students evaluated arguments based on their impressions of whether or not the argument was sufficiently general to addresses every case in the domain of a claim.
5. Criticisms of arguments as viable or not viable based on criterion such as whether or not the logical inferences were correct, the results to show that every case of the condition must have the conclusion were indeed prior results.
6. Criticism of arguments as viable or not viable based on whether or not counterexamples to general claims were impossible. This occurred in both direct and indirect contexts. In direct contexts, students discussed whether cases of the conditions and not the conclusion are shown to be impossible, and in indirect contexts students discussed whether cases where the conclusion was not met, the conditions could not have occurred (contrapositive) or discussed whether the argument demonstrated that the supposition of counterexample lead to a contradiction or false statement.

Despite the positive findings among some students, we uncovered numerous conceptions that arose among other students that served as barriers to understanding our framework/conceptual pillars as we envisioned them such as a tendency to be overly skeptical, that counterexamples are impossible to eliminate except with exhaustive argument as that any conditional claim might someday be falsified by some strange counterexample yet to be constructed.

Year 3 Findings that Modified and Improved Our Understandings of Teaching and Learning Viable Argument

One substudy of our case study focused on student learning of CP 4-5 based on the LLAMA-based classroom intervention. We are currently preparing an article manuscript reporting about the findings of this substudy and their significance.

Framing and significance.

To describe the understandings targeted by CPs 4-5, skepticism of empirical arguments and knowledge of exhaustion as a secure mode of argumentation, we are now using the term "domain appropriateness" of the argument. Domain appropriateness is the degree to which an argument is appropriate to the claim, based on the relationship between the argument type and the claim's domain. By "appropriate" we mean treated as appropriate in the broader mathematical community. A person who understands domain appropriateness understands the following:

- Empirically checking a proper subset of the cases in a claim's domain does not guarantee the claim is true, although checking *all* the cases does guarantee this.
- Empirically checking a proper subset of the cases in a claim's domain provides an inappropriate argument for the claim, although checking all the cases provides an appropriate argument.

One reason this work is important is that it presents an account that is alternative to existing accounts of student learning about skepticism. In much existing literature, students are seen as using empirical arguments for general claims because they themselves are convinced by such arguments that such claims are true (e.g., Harel & Sowder, 1998). This literature thus seeks to unseat students' empirical proof schemes by getting students to doubt that general claims are true even in the face of confirming

evidence. The idea is that once a student has sufficient skepticism, they will turn to something more secure than empirical evidence in order to convince themselves that a general claim is true. In contrast, now we seek to development students' understanding of the limitations of empirical argumentation in a different way. Rather than viewing this student understanding as deriving from the degree of skepticism students have toward general mathematical claims, we treat this student understanding as deriving from the degree of knowledge students have about the norms of argumentation practiced in the broader mathematical community. Thus it is an understanding of domain appropriateness, not skepticism, that our model proposes as prerequisite to the sought understanding of the limitations of empirical arguments.

Domain appropriateness is an understanding that operates at two levels. In one level the student has experienced, internalized, and generalized how a claim can be false despite supporting empirical evidence. The other level is at a meta-level—the student understands a norm of the community of practice, the rationale for which is provided by their understanding at the first level. This helps students to see why this meta-level norm is different from an arbitrary mathematical convention, such as the choice of the letter *m* for a line's slope.

This new view arose from one of our task-based interviews that involved the task below:

Thomas makes the following argument:

Claim: For every whole number value of *n*, if you compute $7n - 1$ you will *not* get a perfect square. (A perfect square is a number like 36, because it is 6^2)

Foundation:

n	$7n - 1$	$\sqrt{7n-1}$ (approximately)
1	6	2.45
2	13	3.61
3	20	4.47
4	27	5.2
5	34	5.83
6	41	6.4
7	48	6.93

Narrative Link: I tested 7n – 1 in the foundation and it is not a perfect square. The claim is true and my argument is viable because I provide evidence.

Is Thomas' argument viable?

The task was chosen to express a generalization that students have not previously addressed. It had an infinite domain and was actually true, but was one for which students were judged unlikely to find a conceptual insight (e.g., the pertinent structure linking the conditions to the conclusion) that can be used to develop a viable general argument. This task allowed us to assess whether students recognize the limitations of empirical evidence as a viable argument, even when no conceptual insight is present and a large number cases and extreme cases are tested. The follow-up questions allowed for the pursuit of an exhaustion argument (an exhaustion of cases actually tested). The week prior to the interviews, students had received instruction on our existing LLAMA lessons on skepticism and on the method of exhaustion.

We found that Seven of the ten students displayed understanding of domain appropriateness in the interviews, and five of them displayed this understanding entirely consistently. They stated that Thomas' argument was not viable, providing the reason that the argument did not account for all of the cases in the claim's domain. Six of these students proposed ways to make Thomas' argument viable, either by restricting the domain of the claim or by finding some sort of general reason that it works for all values of n. Two students explicitly volunteered that they thought the claim was true, even though they also understood that the empirical evidence provided was insufficient for a viable argument. Thus supported our distinction between domain appropriateness and skepticism.

Several students still displayed limited understandings of domain appropriateness. Two thought an empirical argument was appropriate for the claim, although Thomas' argument would be improved by checking more cases. One student also believed that there is no way for a claim with an infinite domain to be established as definitely true, because it is impossible to check all cases. This way of thinking accords with Karl Popper's idea of falsifiability as a criterion for scientific theories. The student doubted that a viable argument had the power to establish a mathematical claim as true.

Year 4 Implications

Our findings suggest that the LLAMA intervention develops student understanding of domain appropriateness of argumentation. Previous studies have developed similar understandings with undergraduates; this substudy shows it to be possible with eighth-graders, by using an approach that is theoretically novel and does not require the developing of skepticism about a claim's truth. Rather it effectively provides students with a rationale for the practices of proof and viable argumentation used in the broader mathematical community.

Along with currently preparing an article manuscript with these findings, we are also informing next year's implementation of LLAMA CP 4-5 activities in light of them. The theoretical distinctions highlighted by this study are important to our summer PD workshop and ongoing coaching with Cohort 2. The findings are of theoretical and empirical significance to our ongoing work, in developing our model of helping students develop an understanding of the limitations of empirical argumentation for general claims.

Year 4 Planned Activities for Addressing our Student Learning Trajectories Question

While our student interviews were very productive for our research on student learning, the interviews did not enable us to track student learning relative to our learning progression as precisely as we hoped. One issue is that even though we keep records of what the teacher reported covering and when, we did not collect sufficient data to compare student learning to what was actually taught. For that to happen, much more explicit record keeping such as teacher journaling, more frequent observations, and video recording of lesson would be needed.

Thus, we have developed an agreement with one of our Cohort 2 teachers to perform a more careful study of student learning in Year 4. This teacher has agreed to use the LLAMA materials for Grade 8 geometry as his primary curriculum throughout the fall of 2019. His efforts will be supported by four members of the LLAMA implementation team, who will provide weekly coaching sessions and lesson planning. Data will collected will be as follows:

1. Daily teacher journaling on what was covered—content and LLAMA CPs—and student reactions, such as student activities, comments, and responses to questions and tasks. The journal will contain an ongoing summary of students' learning of the LLAMA CPs from the teacher's perspective.
2. Daily work samples from six Grade 8 students selected as follows: 2 students who scored advanced on the previous year's annual achievement assessments (SBAC), 2 students who scored proficient, and 2 students who scored near proficient.
3. Classroom observations. On a biweekly basis, a member of the LLAMA research team will observe implementation lessons. These lessons will be video recorded and transcribed. Also, the LLAMA observation protocol will be completed for each lesson.
4. Student task-based interviews. The six students discussed in Data Source 2 will be interviewed at 6 time points during Fall 2019. Interview one will use a task-based design to assess students' knowledge, understanding, and use of CPs 1-3. Interview two, three, four, five, and six will do the same for CPs 4-6, CP 7, CP 9-10, CP 11, and CP 12, respectively. These interviews will be video recorded and transcribed. The tasks for these interviews were developed and field tested in Year 2 of the project.

All of the data described above will be used to develop a comparison between the learning progression theorized and presented during instruction and models for actual students learning trajectories. In consequence, we will have a more complete picture of how students acquire, and in what order, our CPs, and the barriers they face, in comparison to our hypothetical model for learning.

Year 4 Planned Activities for Addressing Teacher Implementation of Viable Argumentation: Affordances, Barriers, and Other Perspectives.

The team identified **12 Cohort 1 teachers of interest** with differing categories of implementation and MKT results. They were interviewed in Years 3 and 4. The interviews have been transcribed and are currently being analyzed. The team will triangulate the interview results with the LLAMA monthly survey, MKT results, and coaches' ratings of teachers to develop a qualitative summary of teacher factors that associate with implementation of viable argumentation or lack thereof. This will be used to determine barriers and affordances to implementing LLAMA argumentation practices in instruction.

Accountability

The accountability evaluation was conducted during Years 1 through 4. The results of the accountability evaluation are included in the Year 4 report.

Dissemination

LLAMA's multipronged communication strategy will reach a broad audience through the major dissemination efforts to distribute the LLAMA research results, curriculum materials (with lesson plans, pacing calendars, software), and valid and reliable instruments. See Exhibit 132 which shows the different target audiences and dissemination methods. Year 1 and Year 2 of the project were devoted to project implementation and data collection. In Year 3 the project team began focusing efforts on analyzing data and distributing research results. In Year 4 the project team will focus heavily on dissemination.

Exhibit 132: Audience and Dissemination Method

Researchers. Prepare manuscripts for publications in prominent math and math education journals.
NSF Community. Participate in the Community for Advancing Discovery Research in Education (CADRE) project network and the annual CADRE meeting.
Project Participants. UI will host a website and offer access to all project newsletters, presentation materials, and articles/conference papers that are developed through the LLAMA project. Provide materials to participating schools for use in their individual improvement plans and reports.
PD Providers. Findings and curriculum materials will be presented to groups involved in teacher PD, including both pre-service and in-service providers.
Math Teachers and Other Stakeholders. LLAMA has created a 6-stage social media plan to share project findings with a broad audience: (1) identify key social media platforms frequently used by educators, (2) encourage project teachers to share reflections with social media networks, (3) create documents to distribute on the identified platforms, (4) create short video testimonials from project participants, (5) identify educators with a social media presence and collaborate to disseminate findings, and (6) conduct social network analysis to measure the reach of the project.

Researchers

As shown in Exhibit 133 as of August 1, 2021 the project team has submitted 2 articles for publication, is currently writing 12 articles, and has delivered 1 conference presentation of a juried conference paper. Exhibit 134 details work that has been submitted and/or presented.

Exhibit 133: Dissemination Efforts

Type	Status
Publications	
Published	2
Awaiting Publication	2
Awaiting Review	1
Currently Writing	8
Juried Conference Papers	1
Conference Presentations	0
Total	16

.

Exhibit 134: Year 3 Dissemination Efforts

Publications	
Yopp, D. (2020). *Eliminating counterexamples: Indirect arguments for improving adolescents' contrapositive reasoning. Journal of Mathematical Behavior, 59. https://doi.org/10.1016/j.jmathb.2020.100794*	Published (LAMP data)
Yopp, D., Ely, R. Adams, A. E., Nielsen, A. W., & Corwine, E.C. (2020) **Eliminating counterexamples: A case study intervention for improving adolescents' ability to critique direct arguments**. *The Journal of Mathematical Behavior*, 57, 1-19. https://doi.org/10.1016/j.jmathb.2019.100751	Published
Ely, R., Yopp, D., & Adams, A. E. (2020). Domain appropriateness and skepticism in viable argumentation. In Sacristán, A.I., Cortés-Zavala, J.C. & Ruiz-Arias, P.M. (Eds.). *Mathematics Education Across Cultures: Proceedings of the 42nd Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education, Mexico.* Cinvestav / AMIUTEM / PME-NA. https:/doi.org/10.51272/pmena.42.2020	Accepted
Yopp, D. Ely, R., Adams, A.E., Nielson, A. (2020) Proof in the middle grades: Can we label middle grade arguments as proof with a capital P? In Bieda, K and et. al., ***Conceptions and Consequences of Argumentation, Justification and Proof*. In press.**	In Press
Yopp, A. D., Dawes, K., Frankie, K., Thummel, M. (2020). The Power of Making Claims: Content, Viable Argumentation, and other Practices. Submitted to NCTM's *Journal of Mathematics Teacher: Learning and Teaching PK-12*.	Submitted
Yopp. D., Adams, A. E., Nielsen, A., Ely, R., Thomas, M., & Barfuss, L. *A sidecar fosters daily viable arguments.*	Writing (Rejected and rewriting)
Yopp, D. *Item 3 and 4 (TARA): Responses to generalizations with finite domains.*	Writing
Yopp, D. *Item 3 and 4 (SARA): Responses to generalizations with finite domains.*	Writing
Ely, R. *Generic examples.*	Writing
Adams, A. E. & Ely, R. Domain Appropriateness.	Writing
Adams, A. E. *Barriers and Affordances to Teaching Mathematics with Viable Argument*	Writing
Yopp, D. *Why is rigid motion Geometry so ripe with viable argument opportunities?*	Writing
Nielsen, A. Choosing argument types	Writing
Nielsen, A. & Yopp, D. *Line of "Good" Fit in Grade 8 (Age 13) Classrooms.* Submitted to Journal of Mathematical Behavior, December 2018.	Rejected
Juried Conference Papers	
Nielsen, A., Adams, A., & Yopp, D. (2018). *Introducing Residual Criterion for Line of "Good" Fit in Grade 8 Classrooms.* Paper and presented at National Council of Teachers of Mathematics (NCTM) Research Conference, Washington, D.C. [https://engagefully.org/Sessions/Details/379537]	Presented

NSF Community

The LLAMA PI (Yopp) and CoPI (Lewis) have attended each of the CADRE PI national meetings and attended the virtual meetings. The LLAMA Research team views the CADRE newsletter and applies all applicable resources to the LLAMA work.

Project Participants

RMC Research created a website (https://sites.google.com/view/llama-project/home) that hosts all project materials for participants including newsletters that keep participants informed of what project activities are coming up, professional development videos for all 12 conceptual pillars, and a resource link that includes argumentation lesson plans and video lessons. Participants who attended summer professional development sessions were also provided binders that included hard copies of argumentation lesson plans. Additionally, an infographic was created for project participants that included data based on the success of a Cohort 1 teacher.

PD Providers

Findings and curriculum materials will be presented to professional development organizations in Idaho, Washington, and Montana such as regional math centers, educational service districts, and math teacher organizations.

Math Teachers and Other Stakeholders

LLAMA has created a 6-stage social media plan to share project findings with a broad audience: (1) identify key social media platforms frequently used by educators, (2) encourage project teachers to share reflections with social media networks, (3) create documents to distribute on the identified platforms, (4) create short video testimonials from project participants, (5) identify educators with a social media presence and collaborate to disseminate findings, and (6) conduct social network analysis to measure the reach of the project.

Reference List

Adams, A. E., Ely, R., & Yopp, D. (2017). Using generic examples to make arguments viable. *Teaching Children Mathematics, 23*(5), 293-300.

Anderson, T., Howe, C., & Tolmie, A. (1996). Interaction and mental models of physics phenomena: Evidence from dialogues between learners. In J. Oakhill & A. Garnham (Eds.), *Mental models in cognitive science* (pp. 247–273). Hove, United Kingdom: Psychology Press.

Antonini, S. (2004). A statement, the contrapositive and the inverse: Intuition and argumentation. In M. Johnsen Høines & A. Berit Fuglestad (Eds.), *Proceedings of the 28th conference of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 47–54). Norway: Bergen.

Antonini, S. & Mariotti, M. A. (2008). Indirect proof: What is specific to this way of proving? *ZDM Mathematics Education*, *40*(3), 401–412.

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

Ball, D. L. & Bass, H. (2000). Making believe: The collection construction of public mathematical knowledge in the elementary classrooms. In D. Phillips (Ed.), *Yearbook of the National Society for the Study of Education: Constructivism in education* (pp. 193–224). Chicago, IL: University of Chicago Press.

Bass, H. (2011). Proof in mathematics education: An endangered species? A review of teaching and learning proof across the grades: A K–16 perspective. *Journal for Research in Mathematics Education*, *42*(1), 98–103.

Bickman, L. (1987). The functions of program theory. *New Directions for Program Evaluation*, *1987*(33), 5–18.

Bieda, K. N. (2010). Enacting proof-related tasks. *Journal for Research in Mathematics Education*, *41*(4), 351–382.

Bieda, K., Holden, C., & Knuth, E. (2006). Does proof prove? Students' emerging beliefs about generality and proof in middle school. *Proceedings of the 28th annual meeting of the North American chapter of the International Group for the Psychology of Mathematics Education* (pp. 395–402).

Burton, L. (1999). The practices of mathematicians: What do they tell us about coming to know mathematics? *Educational Studies in Mathematics, 37*(2), 121–143.

Case, R. (1984). The process of stage transition: A neo-Piagetian view. In R. J. Sternberg (Ed.), *Mechanisms of cognitive development* (pp. 19–44). New York, NY: Freeman.

Cheng, P. W. & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, *17*(4), 391–416.

Cheng, P. W., Holyoak, K. J., Nisbett, R. E., & Oliver, L. M. (1986). Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology*, *18*(3), 293–328.

Del Giudice, Marco. (2017). Heterogeneity Coefficients for Mahalanobis' D as a Multivariate Effect Size. Multivariate Behavioral Research. 52. 216-221. 10.1080/00273171.2016.1262237.

Ellis, A. B. (2011). Generalizing-promoting actions: How classroom collaborations can support students' mathematical generalizations. *Journal of Research in Mathematics Education*, *4*(42), 308–341.

Ellis, A. B., Weber, E., & Lockwood, E. (2014). The case for learning trajectories research. In S. Oesterle, C. Nicol, P. Liljedahl, & D. Allan. (Eds.), *Proceedings of the 38th conference of the International Group for the Psychology of Mathematics Education and the 36th conference of the North American chapter of the Psychology of Mathematics Education* (Vol. 6). Vancouver, Canada: PME.

Emshoff, J. G., Blakely, C., Gottschalk, R., Mayer, J., Davidson, W. S., & Erickson, S. (1987). Innovation in education and criminal justice: measuring fidelity of implementation and program effectiveness. *Educational Evaluation and Policy Analysis*, *9*(4), 300–311.

Evans, J. St. B. T. (1982). *The psychology of deductive reasoning*. London: Routledge & Kegan Paul.

Fischbein, E. (1982). Intuition and proof. *For the Learning of Mathematics, 3*(2), 9–18.

Galbraith, P. L. (1981). Aspects of proving: A clinical investigation of process. *Educational Studies in Mathematics, 12*(1), 1–28.

Ginsburg, H. (1997). *Entering the child's mind*. Cambridge, United Kingdom: Cambridge University Press.

Girotto, V., Light, P. H., & Colbourn, C. J. (1988). Pragmatic schemas and conditional reasoning in children. *Quarterly Journal of Experimental Psychology*, *40*(3), 469–482.

Hanna, G. (1990). Some pedagogical aspects of proof. *Interchange, 21*, 6–13.

Hanna, G. (2000). Proof, explanation, and exploration: An overview. *Educational Studies in Mathematics*, *44*(1–2), 5–23.

Hanna, G. & Jahnke, H. N. (1996). Proof and proving. In A. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *International handbook of mathematics education* (pp. 877–908). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Harachi, T. W., Abbott, R. D., Catalano, R. F., Haggerty, K. P., & Fleming, C. B. (1999). Opening the black box: using process evaluation measures to assess implementation and theory building. *American Journal of Community Psychology*, *27*(5), 711–731.

Healy, L. & Hoyles, C. (2000). A Study of Proof Conceptions in Algebra. *Journal for Research in Mathematics Education*, 31(4), 396-428.

Hersh, R. (1993). Proving is convincing and explaining. *Educational Studies in Mathematics*, *24*(4), 389–399.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal*, *105*(1), 11–30.

Hotelling, H. (1931). "The generalization of Student's ratio". Annals of Mathematical Statistics. 2 (3): 360–378. doi:10.1214/aoms/1177732979.

Institute of Education Sciences (IES). What Works Clearinghouse Evidence Standards for Reviewing Studies, Version 1.0. Revised May 2008. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_version1_standards.pdf.

Institute of Education Sciences (IES). What Works Clearinghouse Standards Handbook, Version 4.0. Revised October 2017. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Cambridge, United Kingdom: Cambridge University Press.

Johnson-Laird, P. N. (1990). The development of reasoning ability. In G. Butterworth & P. Bryant (Eds.), *Causes of development* (pp. 85–110). Hillsdale, NJ: Laurence Erlbaum Associates.

Johnson-Laird, P. N. & Byrne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Laurence Erlbaum Associates.

Johnson-Laird, P. N., Oakhill, J. V., & Bull, D. (1986). Children's syllogistic reasoning. *Quarterly Journal of Experimental Psychology*, *38*(1), 35–58.

Knuth, E. J. (2002). Proof as a tool for learning mathematics. *Mathematics Teacher*, *95*(7), 486–491.

Krummheuer, G. (1995). The ethnography of argumentation. In P. Cobb & H. Bauersfeld (Eds.), *The emergence of mathematical meaning: Interaction in classroom cultures* (pp. 229–269). Hillsdale, NJ: Lawrence Erlbaum.

Küchemann, D. & Hoyles, C. (2009). From empirical to structural reasoning: Tracking changes over time. In D. A. Stylianou, M. L. Blanton, & E. J. Knuth (Eds.), *Teaching and learning proof across the grades: A K–16 perspective* (pp. 171–190). New York, NY: Routledge.

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

Leithwood, K. A., & Montgomery, D. J. (1980). Evaluating program implementation. *Evaluation Review*, *4*(2), 193–214.

Leron, U. (1985). A direct approach to indirect proofs. *Educational Studies in Mathematics*, *16*(3), 321–325.

Lobato, J., Clarke, D., & Ellis, A. B. (2005). Initiating and eliciting in teaching: A reformulation of telling. *Journal for Research in Mathematics Education*, *36*(2), 101–136.

Lobato, J., Hohensee, C., Rhodelhamel, D., & Diamond, J. (2012). Using student reasoning to inform the development of conceptual learning goals: The case of quadratic functions. *Mathematical Thinking and Learning*, *14*(2), 85–119.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2013). *Qualitative data analysis: An expanded sourcebook* (3rd ed). Thousand Oaks, CA: Sage Publications.

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation*, *24*, 315–340.

Munter. C., Wilhelm A. G., Cobb, P., & Cordray, D.S., (2014) Assessing fidelity of implementation of an unprescribed, diagnostic mathematics intervention, *Journal of Research on Educational Effectiveness*, *7*(1), 83-113, DOI: 10.1080/19345747.2013.809177/. To link to this article: https://doi.org/10.1080/19345747.2013.809177

National Science Foundation. (n.d.). *CAREER: Proof in Secondary Classrooms: Decomposing a Central Mathematical Practice* [DRL1453493]. Retrieved November 27, 2015, from http://nsf.gov/awardsearch/showAward?AWD_ID=1453493&HistoricalAwards=false

National Science Foundation. (n.d.). *Preparing Urban Middle Grades Mathematics Teachers to Teach Argumentation Throughout the School Year* [1417895]. Retrieved November 27, 2015, from http://nsf.gov/awardsearch/showAward?AWD_ID=1417895

Nelson, M. C., Cordray, D. S., Hulleman, C. S., Darrow, C. L., & Sommer, E. C. (2010, March). *A procedure for assessing fidelity of implementation in experiments testing educational interventions*. Paper presented at the annual meeting of the Society for Research on Educational Effectiveness, Washington, DC.

Nelson, M. C., Cordray, D. S., Hulleman, C. S., Darrow, C. L., & Sommer, E. C. (2012). A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. *Journal of Behavioral Health Services and Research*, *39*, 374–396.

New York State Education Department, EngageNY. *Grade 8 Mathematics*. Retrieved September 10, 2013, from https://www.engageny.org/resource/grade-8-mathematics

O'Donnell, C. L., (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum interventions. *Review of Educational Research*. *78*(1), 33–84 DOI: 10.3102/0034654307313793 © 2008 AERA. http://rer.aera.net.

Pedemonte, B. (2008). Argumentation and algebraic proof. *ZDM Mathematics Education*, *40*(3), 385–400.

Porteous, K. (1990). What do children really believe? *Educational Studies in Mathematics*, *21*(6), 589–598.

The Regents of the University of Michigan. (2006). *Learning Mathematics for Teaching (LMT) Project Mathematical Knowledge of Teaching Assessment (Patterns, Functions and Algebra Version)*. [For information, questions, or permission requests please contact Merrie Blunk, Learning Mathematics for Teaching, 734-615-7632.]

Reid, D. & Dobbin, J. (1998). Why is proof by contradiction difficult? In A. Olivier & K. Newstead (Eds.), *Proceedings of the 22nd conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 41–48). Stellenbosch, South Africa: University of Stellenbosch.

Rowland, T. (2002). Generic proofs in number theory. In S. R. Campbell & R. Zazkis (Eds.), *Learning and teaching number theory: Research in cognition and instruction* (pp. 157–184). Westport, CT: Ablex Publishing.

Sandefur, J., Mason, J., Stylianides, G. J., & Watson, A. (2013). Generating and using examples in the proving process. *Educational Studies in Mathematics*, *83*(3), 323–340.

Sapp, Marty & Obiakor, Festus & J. Gregas, Amanda & Scholze, Steffanie. (2007). Mahalanobis distance: A multivariate measure of effect in hypnosis research. Sleep and Hypnosis. 9. 67-70.

Schleppegrell, M. J. (2007). The linguistic challenges of mathematics teaching and learning: A research review. *Reading & Writing Quarterly*, *23*(2), 139–159.

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

Smarter Balanced Assessment Consortium. (2012a). *Claims for the mathematics summative assessment*. Retrieved from http://www.smarterbalanced.org/sample-items-and-performance-tasks/

Smarter Balanced Assessment Consortium. (2012b). *Smarter Balanced assessment: 8th grade mathematics*. Retrieved from http://www.smarterbalanced.org/smarter-balanced-assessments/

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S.W. (2011). *Optimal design plus empirical evidence: Documentation for the "optimal design" software version 3.0*.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research*. Thousand Oaks: Sage Publications.

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

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

Stylianides, G. J. (2009). Reasoning-and-proving in school mathematics textbooks. *Mathematics Thinking and Learning*, *11*(4), 258–288.

Stylianides, G. J. & Stylianides, A. J. (2008). Proof in school mathematics: Insights from psychological research into students' ability for deductive reasoning. *Mathematical Thinking and Learning*, *10*(2), 103–133.

Stylianides, G. J. & Stylianides, A. J. (2009). Facilitating the transition from empirical arguments to proof. *Journal for Research in Mathematics Education*, *40*(3), 314–352.

Stylianides, G. J. & Stylianides, A. J. (2015). Research-based interventions in mathematics classrooms: Enhancing students' learning of proving. *Educational Studies in Mathematics, 89*, 149–150.

Thompson, D. R. (1996). Learning and teaching indirect proof. *The Mathematics Teacher*, *89*(6), 474–82.

Toulmin, S. (1958/2003). *The uses of argument.* Cambridge, United Kingdom: Cambridge University Press.

Wason, P. C. (1966). Reasoning. In B. M. Foss (Ed.), *New horizons in psychology* (Vol. 1, pp. 135–151). Harmondsworth, United Kingdom: Penguin.

Weber, K. (2014). *What is a proof? A linguistic answer to an educational question*. Paper presented at the 17th annual conference of the Special Interest Group of the Mathematical Association of America on Research in Undergraduate Mathematics Education, Denver, CO. Retrieved from http://sigmaa.maa.org/rume/crume2014/Schedule/Papers.htm

West, L., & Staub, F. C. (2003). *Content-focused coaching: Transforming mathematics lessons*. Portsmouth, NH: Heinemann.

Wu, H. (1996). The role of Euclidean geometry in high school. *Journal of Mathematical Behavior*, *15*(3), 221–237.

Yopp, D. A. (2009). From inductive reasoning to formal proof. *Mathematics Teaching in the Middle School*, *15*(5), 286–291.

Yopp, D. A. (2010). Developing an understanding of logical necessity. *Teaching Children Mathematics*, *16*(7), 410–422.

Yopp, D. A. (2011a). How some research mathematicians and statisticians use proof in undergraduate mathematics. *Journal of Mathematical Behavior*, *30*(2), 115–130.

Yopp, D. A. (2011b). Valuing informal arguments and empirical investigations during collective argumentation. *Primus*, *22*(8), 643–663.

Yopp, D. A. (2013). Counterexamples as starting points for reasoning and sense making. *The Mathematics Teacher*, *106*(9), 674–679.

Yopp, D. A. (2014). Viable arguments, conceptual insights, and technical handles. In S. Oesterle, C. Nicol, P. Liljedahl, & D. Allan (Eds.), *Proceedings of the 38th conference of the International Group for the Psychology of Mathematics Education and the 36th conference of the North American chapter of the Psychology of Mathematics Education* (Vol. 5, pp. 401–408). Vancouver, Canada: PME.

Yopp, D. A. (2015). Prospective elementary teachers' arguing and claiming in responses to false generalizations. *Journal of Mathematical Behavior*, *39*, 79–99.

Yopp, D. A. (in press). Using dilations and similarity to derive the equation of a line. *Teaching Mathematics in the Middle School*.

Yopp, D.A. & Ely, R. (2015). When Does an Argument Use a Generic Example? *Educational Studies in Mathematics,* 1–17. doi: 10.1007/s10649-015-9633-z

Yopp, D. A., Ely, R., & Johnson-Leung, J. (2015). Generic example proving criteria for all. *For the Learning of Mathematics*, *5*(3), 8-13.

Yopp, D. A., Sutton, J. T., Espel, E., & Wang, X. (2015). *Learning Algebra and Methods for Proving (LAMP) Annual National Science Foundation Report*. Report in progress.

Appendix A
Observation Analysis Using Raw Scores

Analysis and Findings: Observations

Classroom observation is one component of the LLAMA Learning Progression Study (Study 4). This section describes the findings from observations conducted in Years 1 through 4. Details about study recruitment and attrition are described in the **Student Achievement Chapter**. In the initial proposal, observations were to be conducted twice per year for Cohort 1 teachers in Years 1, 2, and 3, with an additional third observation for randomly-selected case study teachers. However, due to many early changes in the research studies' data collection plans, the team agreed that conducting an observation with each coaching visit would provide context and additional information to the primary analyses. Because Cohort 1 teachers began the professional development in Year 1, and Cohort 2 teachers delayed entry into the professional development until Year 3, Cohort 1 teachers were observed in Years 1, 2, and 3, whereas Cohort 2 teachers were only observed in Years 3 and 4. Observations were intended to be recorded by coaches at each in-person or remote coaching visit. For each active teacher, coaches were instructed to conduct at least one fall and one spring observation of a class where only the teacher taught the class (i.e., the coach did not assist in instruction). However, in practice data coded as "fall" or "spring" observations include a mix of teacher taught, coach taught, and combination classes, so these variables were not included in the analysis. Observations in Year 1 were conducted by 4 of the 5 coaches; observations in Years 2 through 4 were conducted by all 5 coaches. When more than one coach observed a class, the senior team member's observation was entered, and the junior team member's observation was excluded from the data set. Observations occurred throughout the school year, beginning in August for teachers who started school earlier and ending in June for teachers whose school year ended later.

Analytic Sample

The analytic sample for the observation analyses includes observations from all teachers (both RCT and non-RCT) who participated in the LLAMA professional development and have at least one classroom observation, summarized below in Exhibit A.1. Teachers were observed the least during Year 1 due to the late start of the project. Observations per teacher averaged 12 over the course of the project.

Exhibit A.1: Analytic Sample for Observation Analyses

Project Year	Project Year Observations (*n*)	Unique Teachers Observed (*n*)	Observations per Teacher (*M*)
Year 1: 2016-2017	47	28	2
Year 2: 2017-2018	275	25	11
Year 3: 2018-2019	202	30	7
Year 3: Cohort 1	11	9	1
Year 3: Cohort 2	191	21	9
Year 4: 2019-2020	100	11	9
Total	**624**	**51**	**12**

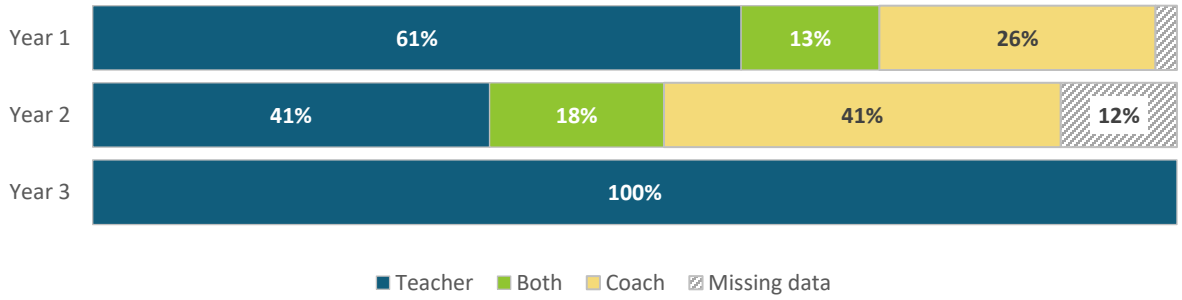
Note. Year 3 is the only year in which both cohorts were observed. Only Cohort 1 was observed

during Years 1 and 2, and only Cohort 3 was observed during Year 4.

Descriptive Summary of Observed Classes

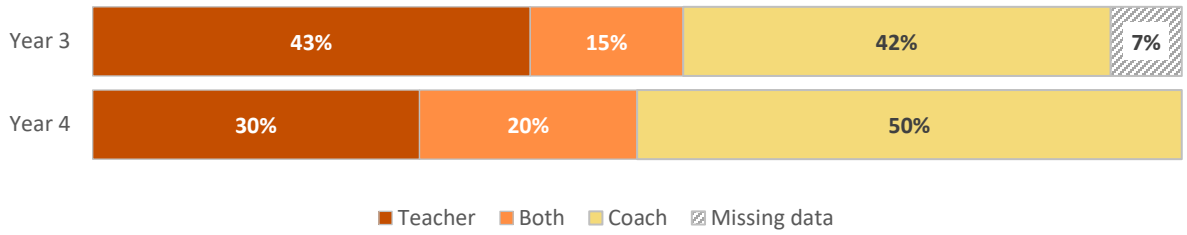
The average class size for observed classes was around 20 students. Most observed classes were labeled Grade 8 Math, although a few Pre-Algebra, Algebra I, Geometry, Grade 7 Math, and Intervention classes were also observed. Additionally, observations were coded by who taught the observed class: the teacher, the coach, or a combination of both. Exhibit A.2 shows the Cohort 1 observations and Exhibit A.3 shows the Cohort 2 observations. Frequencies in who taught the class varied by year for both cohorts; teacher-led observations ranged from 30% to 100%.

Exhibit A.2: Observations by Who Taught the Class, Cohort 1



Note. The *n*'s count the number of observations included in the exhibit. **Cohort 1: Year 1:** *n* = 46 (*n* = 1 missing, 2%); **Year 2:** *n* = 241 (*n* = 34 missing, 12%); **Year 3** *n* = 11 (*n* = 0 missing, 0%). Percentages may not total 100%, due to rounding.

Exhibit A.3: Observations by Who Taught the Class, Cohort 2



Note. The *n*'s count the number of observations included in the exhibit. **Cohort 2: Year 3:** *n* = 177 (*n* = 14 missing, 7%) **Cohort 2: Year 3:** *n* = 100. Percentages may not total 100%, due to rounding.

Conceptual Pillars

Observations captured the LLAMA Conceptual Pillars observed during the class(see Exhibit A.4).

Conceptual Pillars were more consistently implemented in Year 2 for Cohort 1 and Year 4 for Cohort 2. Conceptual Pillars 1, 2, 3, 4 and 8 were each observed for at least one class in each cohort and year. More than a quarter of Cohort 2 observations were more heavily focused on Conceptual Pillars 1, 2, and 3 (45%, 30%, and 26%, respectively) during Year 3, while during Year 4 there was an increase in focus on Pillars 4-10.

Exhibit A.4: Conceptual Pillars Observed, by Year and Cohort

	Cohort 1			Cohort 2	
Conceptual Pillar	**Year 1**	**Year 2**	**Year 3**	**Year 3**	**Year 4**
Pillar 1	**23%**	**31%**	**46%**	**45%**	**32%**
Pillar 2	**11%**	**28%**	**18%**	**30%**	**36%**
Pillar 3	**9%**	**30%**	**9%**	**26%**	**17%**
Pillar 4	**2%**	**11%**	**9%**	**9%**	**15%**
Pillar 5	0%	**10%**	**27%**	**9%**	**10%**
Pillar 6	0%	**11%**	0%	**3%**	**15%**
Pillar 7	0%	**10%**	0%	**6%**	**17%**
Pillar 8	**4%**	**16%**	**18%**	**15%**	**26%**
Pillar 9	0%	**12%**	0%	**2%**	**14%**
Pillar 10	**2%**	**29%**	0%	**12%**	**18%**
Pillar 11	0%	**18%**	0%	**7%**	**7%**
Pillar 12	0%	**6%**	0%	**4%**	**2%**
None addressed	**53%**	**11%**	**36%**	**9%**	**10%**

Note. The *n*'s count the number of observations included in the exhibit. **Cohort 1: Year 1:** *n* = 47; **Year 2:** *n* = 275; **Year 3:** *n* = 11. **Cohort 2: Year 3:** *n* = 191. **Cohort 2: Year 4:** *n* = 100

Student Participation

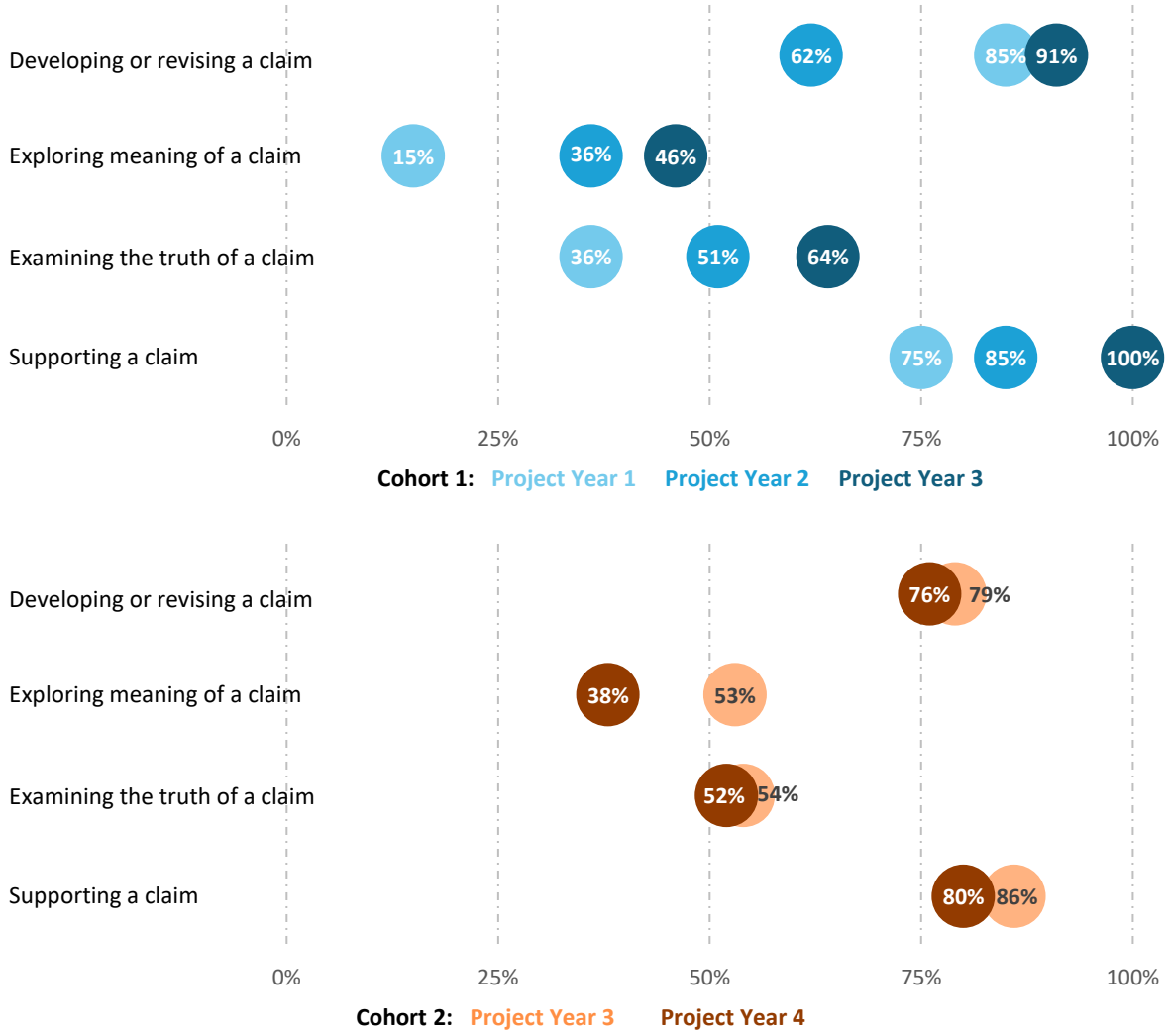
In Year 1, the Observation Protocol asked, "Explicitly quantify and describe the approximate number of students and percentage of the students in the class who were significantly involved in the argumentation episode (this means constructing their own argument or exploring a relevant claim on paper, computer, etc. or activity participating in the class discussion)?" In Year 2, this was broken into 2 items: (1) "record the approximate percentage of the class who were actively involved in writing or developing arguments at some point during the class (this includes constructing their own argument or exploring a relevant claim on paper, computer, etc. or actively participating in the class discussion);" and (2) "record the approximate percentage of the class who had access to the particular argumentation episode you chose to focus on for prompts 3-10 below (in other words, students who were present and attentive or active, and not doing something entirely different during the argumentation episode)." For the analysis these qualitative responses were recoded as numeric percentages. In Years 2 through 4 nearly all the observations report 90% or more of students were significantly involved in the argumentation episode and have access to the argument. These percentages are very high and may indicate that observers were recording student participation in general, rather than the percentage of students were significantly involved the argumentation episode. The LLAMA leadership team should discuss these data. It should be noted that several observations are missing data for these items.

Claims for Argument Episodes

The Observation Protocol instructs the rater to focus on one argumentation episode (e.g., overarching reasoning type) observed during the class. The first 4 items in this section describe the observed claim: the **nature** of the claim, the **type** of claim, the **explicitness** of the claim, and the **clarity** of the claim.

"supporting a claim" was observed for most argument episodes (75% or more of episodes in any given cohort and year). "Developing or revising a claim" was observed often in Years 1, 3, and 4 but was less frequent in Year 2.

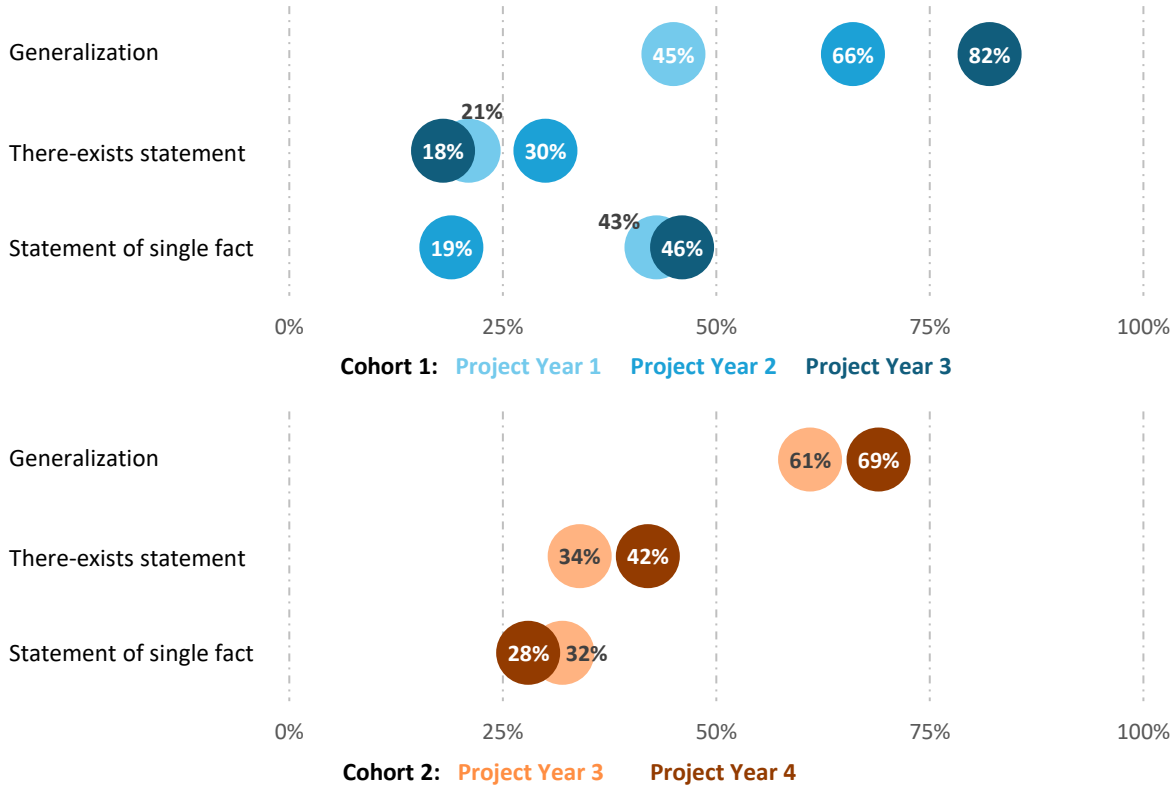
Exhibit A.4: Nature of the Claim Observed in the Argument Episode



Note. The *n*'s count the number of observations included in the exhibit. **Cohort 1:** Year 1: *n* = 47; Year 2: *n* = 275; Year 3: *n* = 11. **Cohort 2:** Year 3: *n* = 191 Cohort 2: Year 4: n = 100.

Generalization was the most frequently observed type of claim in all years for both cohorts.

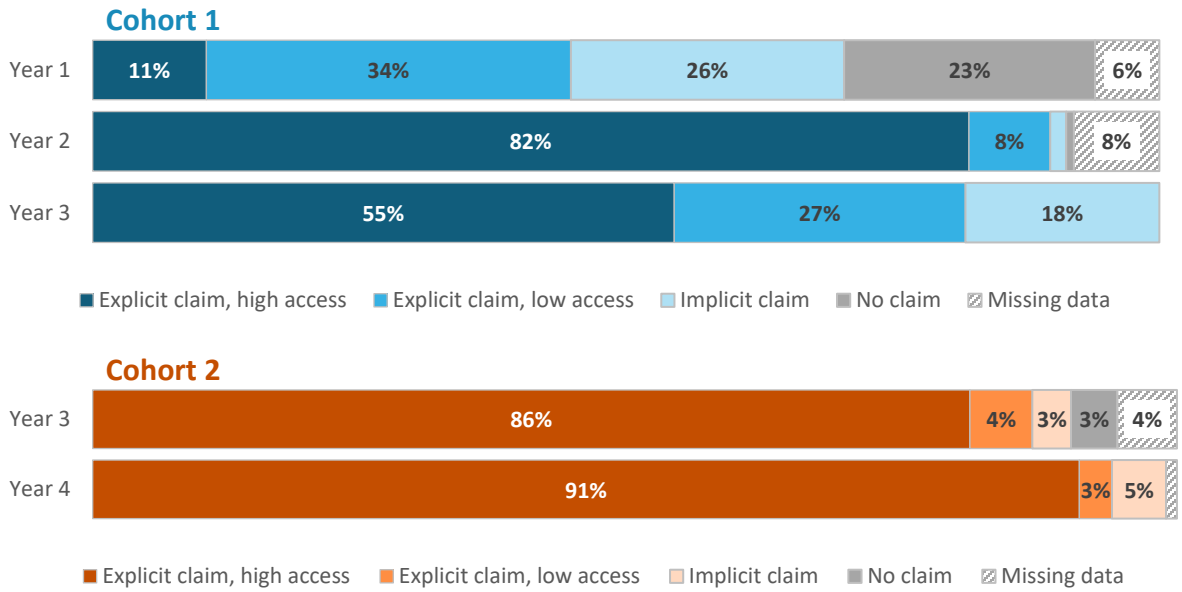
Exhibit A.5: Type of Claim



Note. The *n*'s count the number of observations included in the exhibit. **Cohort 1:** Year 1: $n = 47$; Year 2: $n = 275$; Year 3: $n = 11$. **Cohort 2:** Year 3: $n = 191$ Cohort 2: Year 4: $n = 100$.

In terms of Explicitness of Claim, Cohort 1 had the highest percentage of explicit claims during Year 2 (82%). High percentages of Cohort 2 observations were rated explicit for both Years 3 (86%) and Year 4 (91%).

Exhibit A.6: Explicitness of Claim

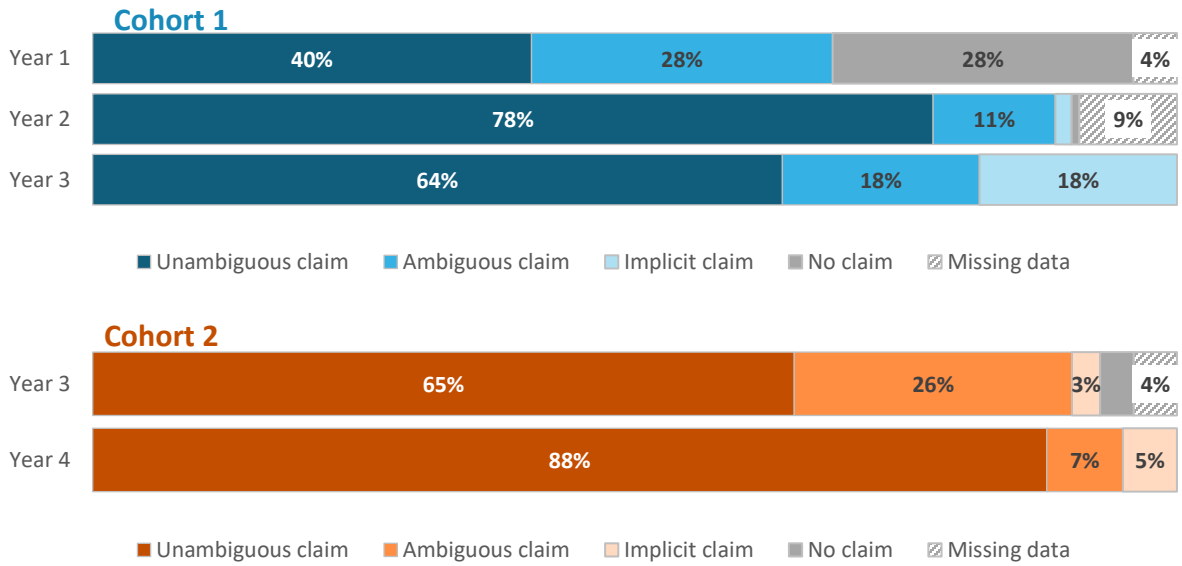


Note. The *n*'s count the number of observations included in the exhibit. **Cohort 1: Year 1:** *n* = 44 (*n* = 3 missing, 6%); **Year 2:** *n* = 252 (*n* = 23 missing, 8%); **Year 3:** *n* = 11 (*n* = 0 missing, 0%). **Cohort 2: Year 3:** *n* = 183 (*n* = 8 missing, 4%). **Year 4:** *n* = 99 (*n* = 1 missing, 1%). Percentages may not total 100%, due to rounding.

In Year 2 "Clarity of Claim" was recoded from a 3-point rubric (0, 1, 2) into a 4-point rubric (0, 1, 2, 3): "ambiguous claim" (originally scores as "1") was divided into "ambiguous claim" (new score of "2") and "implicit claim" (new score of "1").

The clarity of observed claims increased from Year 1 to Year 2 for Cohort 1: the percentage of unambiguous claims observed nearly doubled (from 40% in Year 1 to 78% in Year 2). Unambiguous claims were observed in 65% of observations for Cohort 2 in Year 3, which increased to 88% in Year 4.

Exhibit A7: Clarity of Claim



Note. The *n*'s count the number of observations included in the exhibit. **Cohort 1: Year 1:** *n* = 45 (*n* = 2 missing, 4%); **Year 2:** *n* = 250 (*n* = 25 missing, 9%); **Year 3:** *n* = 11 (*n* = 0 missing, 0%). **Cohort 2: Year 3:** *n* = 184 (*n* = 7 missing, 4%); **Year 4**: n = 100. Percentages may not total 100%, due to rounding.

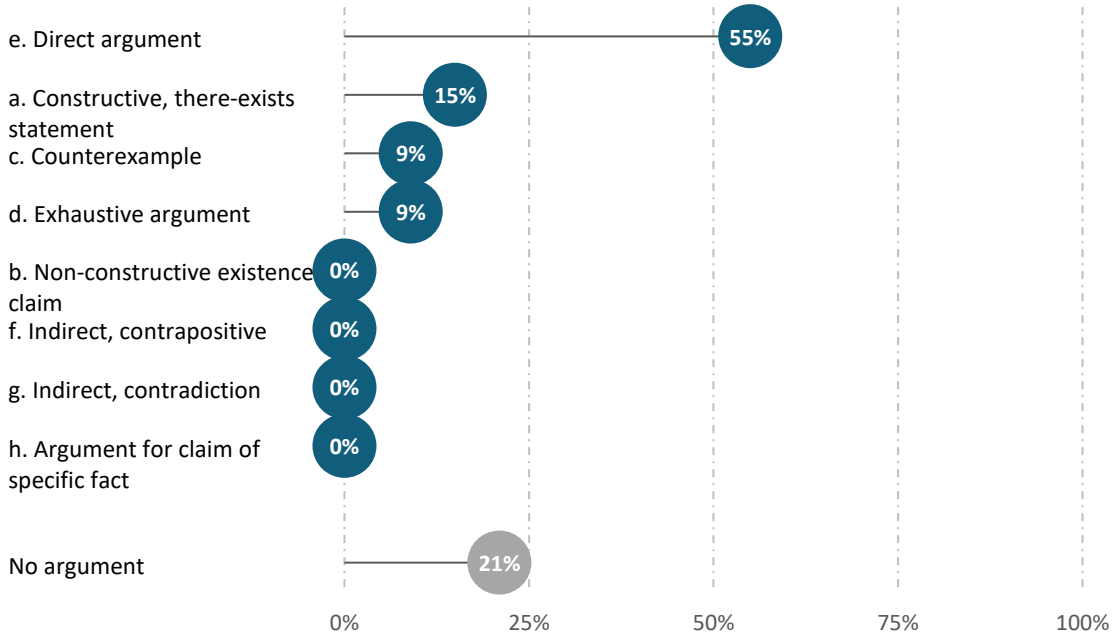
Argument Type and Support for Observed Argument Episodes

The last section of the Observation Protocol asks coaches to select the argument type for the observed argument episode and to rate support accompanying the selected argument type: the rater first circles the argument type(s) observed and then rates the support for the corresponding argument type, on a scale of 0 to 3, with 0 being low and 3 being high. The exact rubric criteria for scores of 0, 1, 2, or 3 vary by argument type. On the Observation Protocol used in Year 1, support was scored for both the teacher and the students; however, after discussion the research team agreed that the student scores are more meaningful. Therefore, the teacher scores from Year 1 are omitted from this report. Note also that on the Observation Protocol, **there is no rubric to score support for argument type b: non-constructive argument for existence**. Argument type h: argument for claim of specific fact was added to the protocol in Year 2.

Year 1, Cohort 1

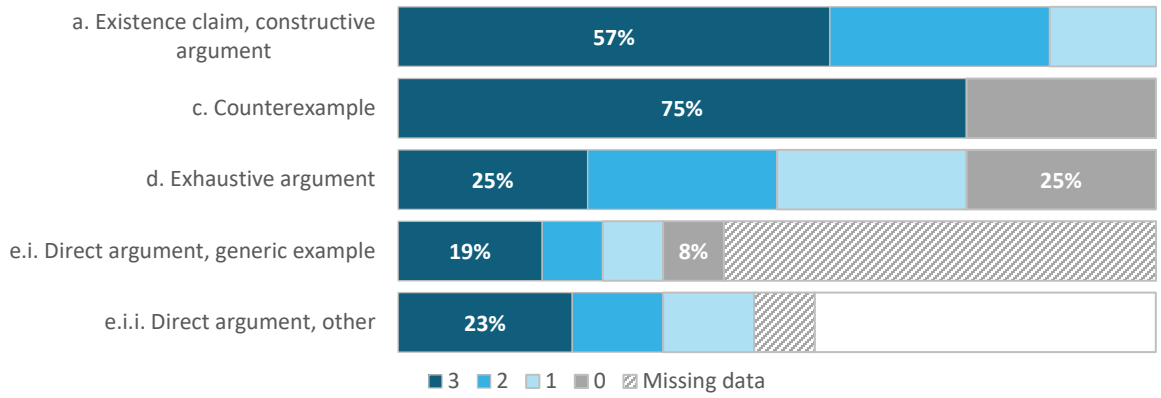
Argument episodes were scored for nearly 80% of Cohort 1 observations in Year 1. Over half of the observations recorded the use of a "direct argument" (55%), with fewer observations noting the use of a "constructive, there-exists statement," "counterexample," or "exhaustive argument"(9%-15%). Overall support ratings (Exhibit A.9) were high for these items.

Exhibit A.8: Argument Type(s) for Observed Argument Episode, Year 1, Cohort 1



Note. The *n*'s count the number of observations included in the exhibit. **Cohort 1: Year 1:** *n* = 47.

Exhibit A.9: Support for Observed Argument Episode, Year 1, Cohort 1

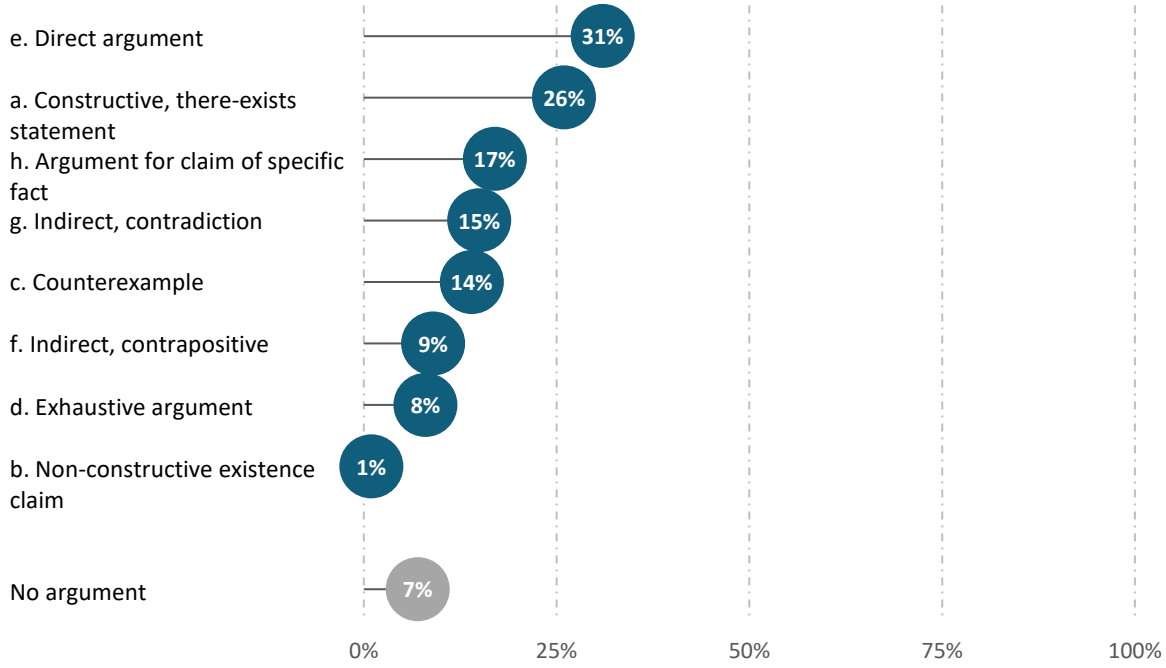


Note. **a:** *n* = 7; **b:** *n* = 0; **c:** *n* = 4; **d:** *n* = 4; **e:** *n* = 26; **f:** *n* = 0; **g:** *n* = 0; **h:** *n* = 0. The denominator for rubric score percentages is the number of observations that included the selected argument type. The *n*'s count the number of observations included in the exhibit. for robust analyses of these items, at least 5 observations for each rating for each type are needed, and the only type to exceed 20 observations is "direct argument," which is further divided into 2 support scores. Percentages may not total 100%, due to rounding.

Year 2, Cohort 1

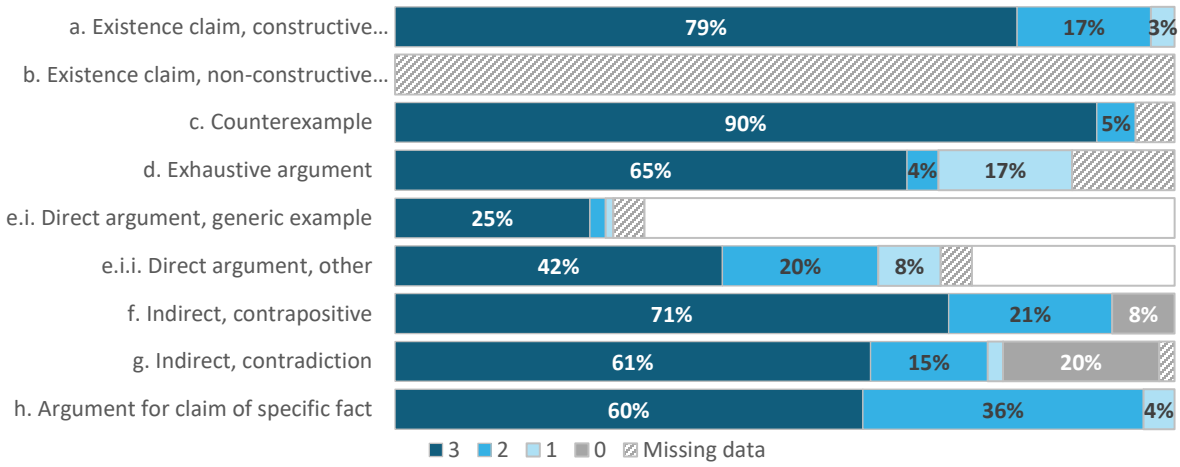
Argument episodes were scored for nearly all Cohort 1 observations in Year 2 (93%). Each argument type was rated for at least one observation, though "direct argument" was most frequently observed **(31% of observations). At least 60% of observations for each argument type received a score of "3."**

Exhibit A.10: Argument Type(s) for Observed Argument Episode, Year 2, Cohort 1



Note. The *n*'s count the number of observations included in the exhibit. **Cohort 1: Year 2:** *n* = 275.

Exhibit A.11: Support for Observed Argument Episode, Year 2, Cohort 1



Note. **a:** *n* = 72; **b:** *n* = 3; **c:** *n* = 39; **d:** *n* = 23; **e:** *n* = 85; **f:** *n* = 24; **g:** *n* = 41; **h:** *n* = 17. The denominator for rubric score percentages is the number of observations that included the selected argument type. The *n*'s count the number of

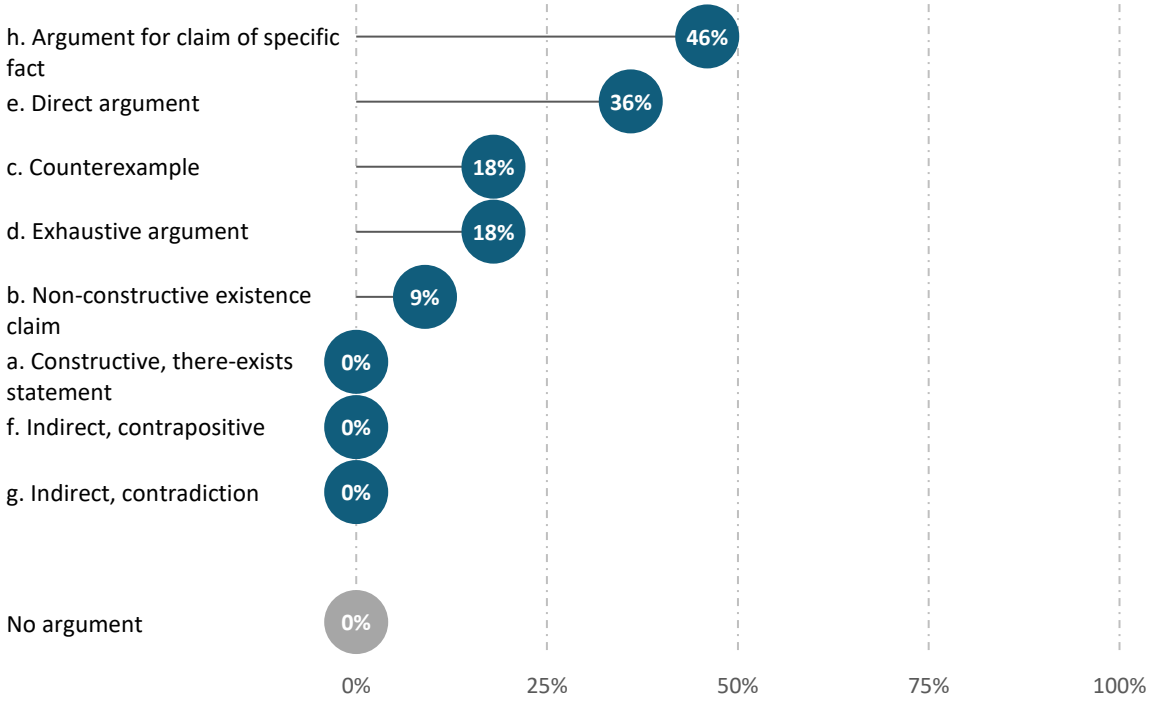
observations included in the exhibit. Percentages may not total 100%, due to rounding.

Year 3, Cohort 1

Argument episodes were scored for all Cohort 1 observations in Year 3; however, only 11 observations were conducted for this cohort and year. The Year 3 observations for Cohort 1 capture a glimpse of the classroom after completing all of the LLAMA professional development. Cohort 1 teachers did not receive any professional development in Year 3, and in some cases, the quality of argumentation was higher with coach support in Year 2 than without coach support in Year 3.It is important to consider that the sample size for Cohort 1 in Year 3 is extremely small--only 2 teachers in the analytic sample were observed more than once.

Nearly half of the observations recorded the use of a "argument for claim of specific fact" (46%). "Constructive, there-exists statement" and "indirect arguments" (neither contrapositive nor contradiction) were observed.

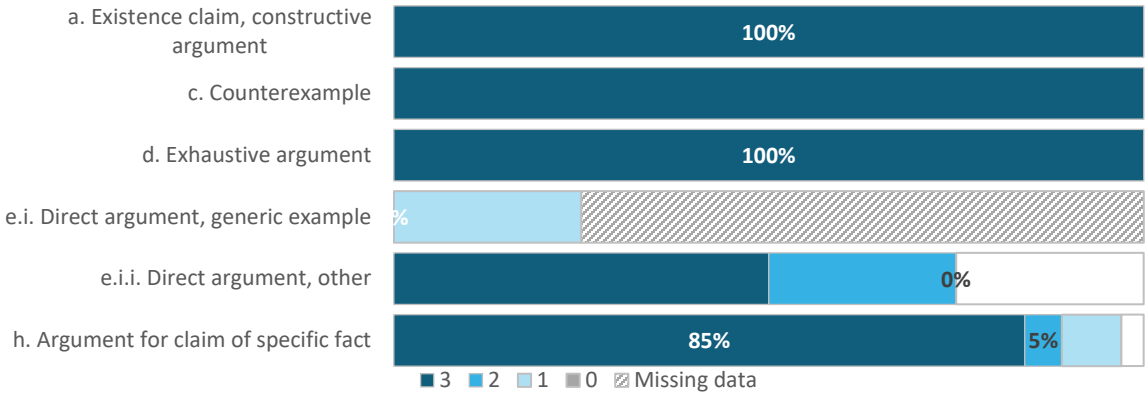
Exhibit A.12: Argument Type(s) for Observed Argument Episode, Year 3, Cohort 1



Note. The *n*'s count the number of observations included in the exhibit. **Cohort 1: Year 3:** *n* = 11.

Although support ratings were generally high, note that no argument type was observed more than 5 times for Cohort 1 in Year 3.

Exhibit A.13: Support for Observed Argument Episode, Year 3, Cohort 1

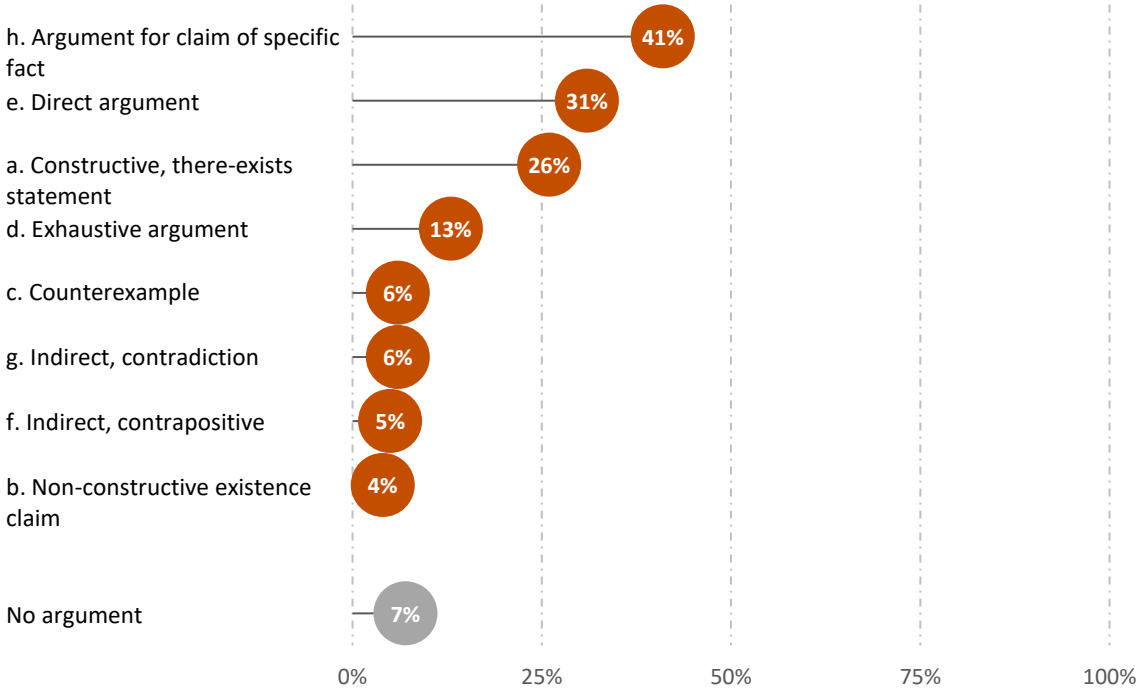


Note. **a:** *n* = 1; **b:** *n* = 0; **c:** *n* = 2; **d:** *n* = 2; **e:** *n* = 4; **f:** *n* = 0; **g:** *n* = 0; **h:** *n* = 5. The denominator for rubric score percentages is the number of observations that included the selected argument type. The *n*'s count the number of observations included in the exhibit. Percentages may not total 100%, due to rounding.

Year 3, Cohort 2

Argument episodes were scored for nearly all Cohort 2 observations in Year 3 (93%). Each argument type was rated for at least one observation, though "argument for claim of specific fact" was most frequently observed (41% of observations).

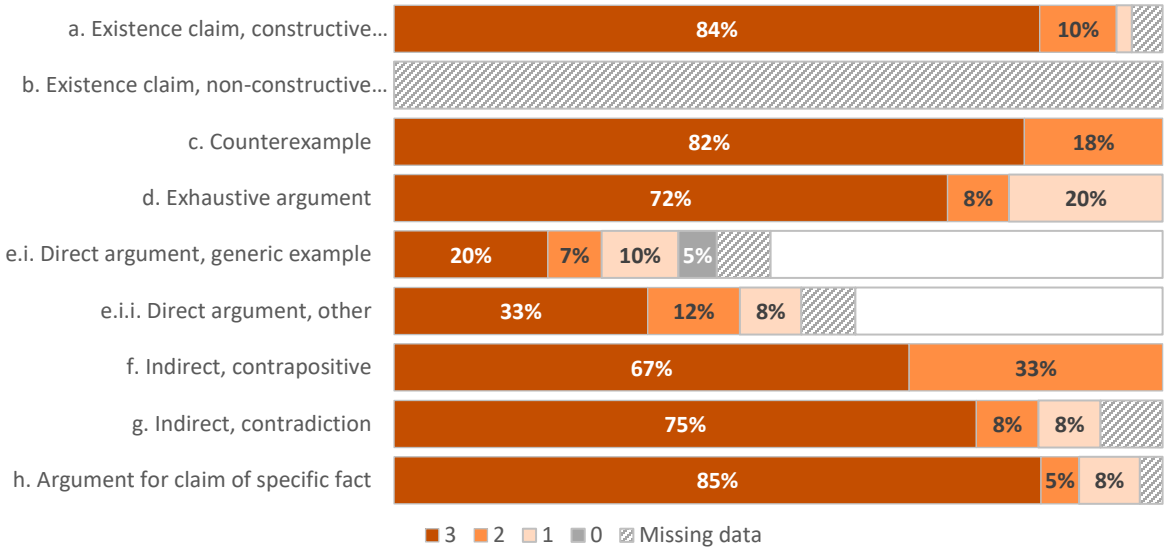
Exhibit A.14: Argument Type(s) for Observed Argument Episode, Year 3, Cohort 2



Note. The *n*'s count the number of observations included in the exhibit. **Cohort 2: Year 3:** *n* = 191.

At least 67% of observations for each argument type received a score of "3," with the exception of "direct argument" (only approximately 53% of observations scored a "3"). The highest rated argument type in terms of support was "argument for claim of specific fact" (85% rated as "3").

Exhibit A.15: Support for Observed Argument Episode, Year 3, Cohort 2

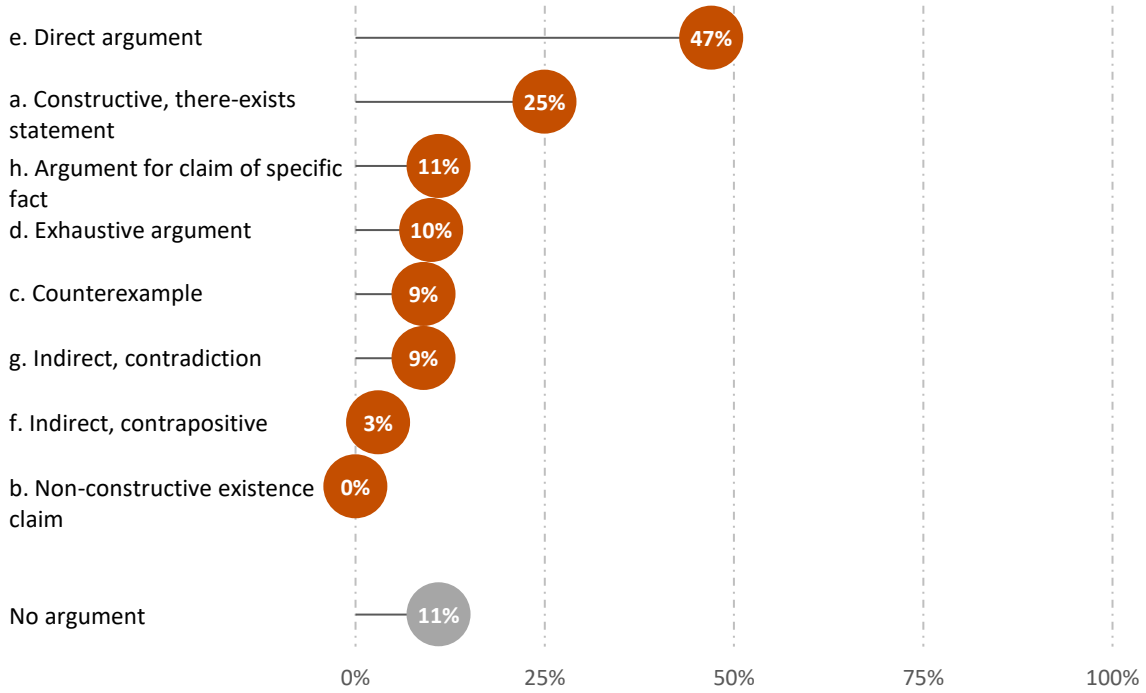


Note. **a:** *n* = 49; **b:** *n* = 0; **c:** *n* = 11; **d:** *n* = 25; **e:** *n* = 60; **f:** *n* = 9; **g:** *n* = 12; **h:** *n* = 40. The denominator for rubric score percentages is the number of observations that included the selected argument type. The *n*'s count the number of observations included in the exhibit. Percentages may not total 100%, due to rounding. Although 3 observations noted use of argument type b: existence claim, non constructive argument, no support score was provided. Additionally, 5 observations had a score for Support e.i. or Support e.i.i. but were not coded as Argument Type e. and 1 observation observations had a score for Support h. but was not coded as Argument Type h; these observations are excluded from the analysis.

Year 4, Cohort 2

Argument episodes were scored for nearly all Cohort 2 observations in Year 3 (89%). Each argument type was rated for at least one observation with the exception of "non-constructive existence claims". "direct argument" was most frequently observed (47% of observations), followed by "constructive, there-exists statement" (25%).

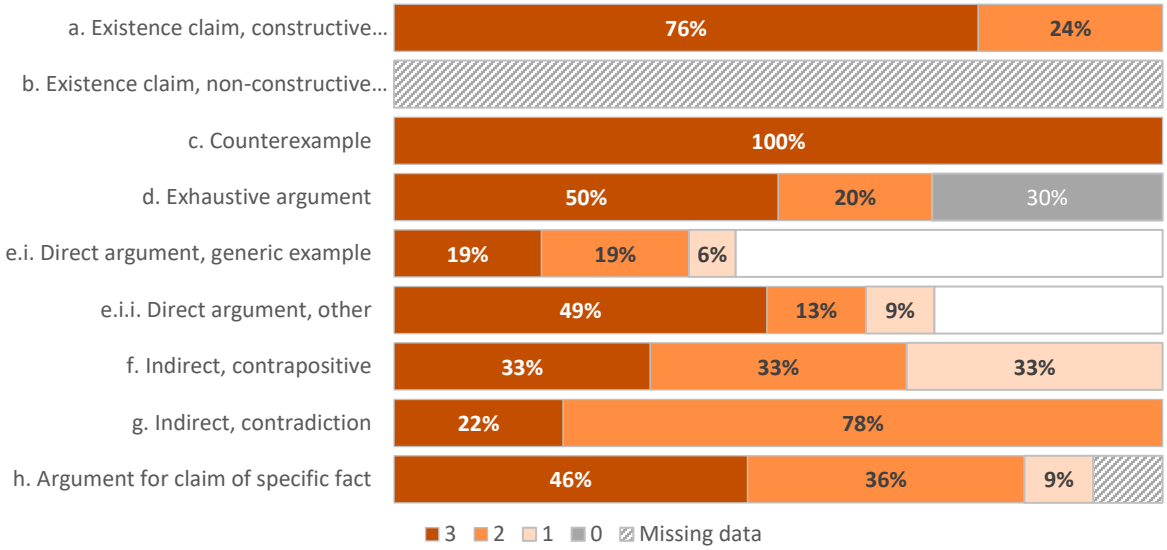
Exhibit A.16: Argument Type(s) for Observed Argument Episode, Year 4, Cohort 2



Note. The *n*'s count the number of observations included in the exhibit. **Cohort 2: Year 4:** *n* = 100.

About half or more of the observations for all argument types received a score of "3" with the exception of "indirect contrapositive" and "indirect contradiction."

Exhibit A.17: Support for Observed Argument Episode, Year 4, Cohort 2



Note. **a:** *n* = 25; **b:** *n* = 0; **c:** *n* = 9; **d:** *n* = 10; **e:** *n* = 47; **f:** *n* = 3; **g:** *n* = 9; **h:** *n* = 11. The denominator for rubric score percentages is the number of observations that included the selected argument type. The *n*'s count the number of observations included in the exhibit. Percentages may not total 100%, due to rounding.

Exhibit A1. Control Group Scores (Year 1 and Year 2)

Item	0		1		2		3		N	median	IQR	mean	sd	95% CI: lower	95% CI: upper
	n	%	n	%	n	%	n	%							
Pretest															
Item 1	215	66%	101	31%	6	2%	6	2%	328	0	1	0.40	0.622	0.332	0.467
Item 2	175	53%	118	36%	26	8%	9	3%	328	0	1	0.60	0.752	0.519	0.682
Item 3	187	57%	61	19%	28	9%	52	16%	328	0	1	0.83	1.125	0.711	0.954
Item 4	266	81%	35	11%	9	3%	18	5%	328	0	0	0.33	0.778	0.242	0.410
Item 5	264	80%	32	10%	26	8%	6	2%	328	0	0	0.31	0.696	0.236	0.386
Posttest															
Item 1	211	64%	101	31%	6	2%	10	3%	328	0	1	0.44	0.683	0.362	0.510
Item 2	154	47%	119	37%	33	10%	19	6%	325	1	1	0.75	0.868	0.659	0.847
Item 3	145	44%	65	20%	47	14%	71	22%	328	1	2	1.13	1.199	1.004	1.264
Item 4	214	65%	68	21%	18	5%	28	9%	328	0	1	0.57	0.932	0.472	0.674
Item 5	235	72%	46	14%	35	11%	12	4%	328	0	1	0.46	0.827	0.374	0.553
Item 6	299	91%	10	3%	6	2%	12	4%	327	0	0	0.18	0.636	0.108	0.246
Item 7	318	97%	9	3%	0	0%	0	0%	327	0	0	0.03	0.164	0.010	0.045
Item 8	282	86%	45	14%	0	0%	0	0%	327	0	0	0.14	0.345	0.100	0.175
Item 9	276	84%	39	12%	11	3%	1	0%	327	0	0	0.20	0.494	0.142	0.249

Exhibit A2. Treatment Group Scores (Year 3)

Item	0		1		2		3		N	median	IQR	mean	sd	95% CI: lower	95% CI: upper
	n	%	n	%	n	%	n	%							
Pretest															
Item 1	283	64%	143	32%	15	3%	4	1%	445	0	1	0.42	0.604	0.360	0.472
Item 2	259	58%	159	36%	17	4%	10	2%	445	0	1	0.50	0.680	0.438	0.564
Item 3	221	50%	83	19%	52	12%	89	20%	445	1	2	1.02	1.190	0.910	1.131
Item 4	338	76%	74	17%	10	2%	23	5%	445	0	0	0.37	0.767	0.295	0.438
Item 5	357	80%	51	11%	24	5%	13	3%	445	0	0	0.31	0.706	0.245	0.376
Posttest															
Item 1	274	62%	119	27%	22	5%	30	7%	445	0	1	0.57	0.866	0.488	0.649
Item 2	181	41%	165	37%	49	11%	50	11%	445	1	1	0.93	0.981	0.837	1.019
Item 3	148	33%	72	16%	73	16%	152	34%	445	2	3	1.51	1.266	1.397	1.632
Item 4	239	54%	73	16%	44	10%	89	20%	445	0	2	0.96	1.199	0.850	1.073
Item 5	290	65%	81	18%	43	10%	31	7%	445	0	1	0.58	0.925	0.498	0.670
Item 6	405	91%	12	3%	6	1%	22	5%	445	0	0	0.20	0.697	0.137	0.267
Item 7	434	98%	11	2%	0	0%	0	0%	445	0	0	0.02	0.155	0.010	0.039
Item 8	389	87%	56	13%	0	0%	0	0%	445	0	0	0.13	0.332	0.095	0.157
Item 9	367	82%	65	15%	11	2%	2	0%	445	0	0	0.21	0.492	0.163	0.255