Embodied truths: How dynamic gestures and speech contribute to mathematical proof practices

Elizabeth L. Pier[a], Candace Walkington[b,*], Virginia Clinton[c], Rebecca Boncoddo[d], Caroline Williams-Pierce[e], Martha W. Alibali[f], Mitchell J. Nathan[a]

[a] Dept. of Educational Psychology, University of Wisconsin – Madison, 1025 W. Johnson St., Madison, WI 53706, USA
[b] Dept. of Teaching and Learning, Southern Methodist University, 6401 Airline Rd., Suite 301, Dallas, TX 75205, USA
[c] Dept. of Educational Foundations and Research, University of North Dakota, 231 Centennial Dr., Grand Forks, ND 58202, USA
[d] Dept. of Psychological Science, Central Connecticut State University, 1615 Stanley St., New Britain, CT 06050, USA
[e] Dept. of Educational Theory and Practice, University at Albany – State University of New York, 1400 Washington Ave., Albany, NY 12222, USA
[f] Dept. of Psychology, University of Wisconsin – Madison, 1202 W. Johnson St., Madison, WI 53706, USA

ARTICLE INFO

ABSTRACT

Grounded and embodied theories of cognition suggest that both language and the body play crucial roles in grounding higher-order thought. This paper investigates how particular forms of speech and gesture function together to support abstract thought in mathematical proof construction. We use computerized text analysis software to evaluate how speech patterns support valid proof construction for two different tasks, and we use gesture analysis to investigate how *dynamic gestures*—those gestures that depict and transform mathematical objects—further support proof practices above and beyond speech patterns. We also evaluate the degree to which speech and gesture convey distinct information about mathematical reasoning during proving. Dynamic gestures and speech indicating logical inference support valid proof construction, and both dynamic gestures and speech uniquely predict variance in valid proof construction. Thus, dynamic gestures and speech each make separate and important contributions to the formulation of mathematical arguments, and both modalities can convey elements of students' understanding to teachers and researchers.

1. Introduction

A central question within the learning sciences is the symbol grounding problem—the question of how arbitrary and abstract symbols, such as those used in mathematics and science, come to have meaning (Harnad, 1990; Searle, 1980). Recently, some scholars have argued that this problem is "solved" (Steels, 2008, p. 223), since theories of embodied cognition (e.g., Glenberg, 1997; Wilson, 2002) and grounded cognition (e.g., Barsalou, 2008) have established *embodiment* as a key means for grounding the meaning of symbols (De Vega, Glenberg, & Graesser, 2012; Havas, Glenberg, Gutowski, Lucarelli & Davidson, 2010; Kaschak, Jones, Carranza, & Fox, 2014; Pulvermüller, 2005). The basic idea is that learners' cognitive representations become grounded through the environment, their bodies, and their brains' modal systems (Barsalou, 2010).

Although these theories of grounded and embodied cognition provide a theoretical account for how symbols are grounded to support

human thought, there is still work to be done to establish how symbol-grounding mechanisms actually operate to support cognition within learning environments that focus on academically relevant knowledge. Scholars posit that speech and bodily action are two such mechanisms by which humans ground higher-order thought (e.g., Kelly et al., 2002; Nathan, 2014), which makes both modalities potentially powerful tools for examining and understanding academic learning and performance. The current study aims to investigate how speech and a specific form of bodily action—gesture—function together to support abstract thought within one particular domain: mathematical proof.

Mathematical proofs are statements of general truth about the properties of and relations among mathematical entities, and they are a key means by which knowledge is built in mathematics (Marghetis, Edwards, & Núñez, 2014; Schoenfeld, 1994). However, in educational settings, students demonstrate great difficulty constructing and understanding proofs, often verifying statements based only on salient perceptual features or specific concrete examples (Chazan, 1993; Healy &

Hoyles, 2000; Knuth, 2002). Thus, the study of mathematical proof practices is an important domain from a pedagogical perspective, and it may also be especially suitable for exploring the embodied and grounded nature of abstract thought more broadly.

In this paper, we examine students' speech and gestures as they engage in constructing proofs for two different mathematical conjectures. We focus on a class of gestures that is particularly relevant for mathematical proof: *dynamic gestures*, which are gestures that depict the progressive transformation of objects or entities. We also focus on characteristics of student speech during proof that possess structural elements of deductive reasoning. We then examine whether learners convey distinct information in speech and gesture as they construct mathematical proofs. This paper contributes to a growing body of research on speech and gesture as grounding mechanisms during mathematical reasoning; with this work, we seek to expand understanding of how people construct and express mathematical arguments or proofs using language and action.

2. Theoretical framework

In the following sections we discuss justification and proof, then examine theories of embodied and grounded cognition, and finally review research on the roles of gesture and language in thinking and learning.

2.1. Justification and proof

Mathematical proofs are a central mode of doing and communicating mathematics (Stylianides, 2007). We follow Harel and Sowder's (1998, 2007) definition of proving, which acknowledges that proof is context dependent and reliant on established standards within the community. To provide an analytic perspective on students' proof practices, we use a proof taxonomy developed by Harel and Sowder (2007) that classifies various known methods of proving. Of greatest relevance to our study are *transformational proof schemes*, part of a broader category of *deductive proof schemes*, which are centrally important to mathematical proof activities. Transformational proof schemes involve the prover's operating upon mathematical objects, observing the result, and constructing the proof accordingly. Harel and Sowder focus on transformations expressed through spoken or written language, but given the body of work on the grounded and multimodal nature of mathematical reasoning (for a review, see Nathan, 2014), we extend their definition to include physical actions and gestures that accompany speech. Specifically, we hypothesize that such movements can represent and transform mathematical objects in ways that can ground the mental transformations that provers are simulating.

The transformational proof scheme has three necessary characteristics: *generality*, or the need for the prover to consider that the proof accounts for all possible cases; *operational thought,* or the prover's "application of mental operations that are goal oriented and anticipatory" (Harel & Sowder, 1998, p. 261); and *logical inference,* or the need for the prover to accurately develop and follow a logical chain of reasoning. We expect that these characteristics of transformational proofs will be evident in the speech and gestures of successful provers, and that an analysis of gestures and speech will highlight the multimodal, embodied nature of proof.

Indeed, examinations of expert mathematicians' proving practices have demonstrated that proof is "a richly embodied practice that involves inscribing and manipulating notations, interacting with those notations through speech and gesture, and using the body to enact the meanings of mathematical ideas" (Marghetis et al., 2014, p. 243). The multimodal nature of proof is also evident among novice students within classroom settings, as students' proofs often take spontaneous verbal and gestural forms, as opposed to formal, written ones (Healy & Hoyles, 2000). Both K-12 teachers and students use gestures as a way to track the development of key ideas when exploring mathematical

conjectures (Nathan et al., 2013). Thus, both speech and gesture can serve as important grounding mechanisms for proof-related reasoning in mathematics classrooms.

2.2. Grounded and embodied cognition

Traditional, symbolic accounts of reasoning propose an overarching cognitive organization that allows arbitrary, abstract, and amodal symbols (so called "AAA symbols," Glenberg et al., 2004) to stand for objects, ideas, events, and relations (Fodor & Pylyshyn, 1988; Newell & Simon, 1972). This perspective has many advantages from a computational point of view: Arbitrary symbols support flexible representations; amodal symbols are easily implemented in current general-purpose digital computers that are often used to model cognition; and operations for manipulating abstract symbols enable the cognitive system to use a single computational mechanism to address both individual instances (tokens) and categories (types).

Yet, there are serious shortcomings to this view. The "symbol grounding problem" (Harnad, 1990) is perhaps best illustrated by Searle's Chinese Room conundrum (1980), a thought experiment in which Searle, who does not know Chinese, is in a closed room where he receives slips of paper with Chinese ideographs, looks up the associated ideographs, and returns slips of paper with the new Chinese ideographs as a response. Searle argues that, although this may give the appearance to people outside the room that they are conversing with a Chinese speaker, Searle derives no meaning from the exchanges. According to traditional symbolic accounts, to know something is to manipulate symbol structures according to syntactic rules. There is no meaning inherent in the symbolic system—save for reference to other symbols—and no systematic way to learn the meaning. The formal symbol account of the exchange among people inside and outside the Chinese Room suffers from what Harnad (1990) describes as "dictionary-go-round" (p. 43), in which unknown terms referenced in one place in the book reference other unknown terms elsewhere in the book, which in turn reference other terms *ad infinitum*.

Grounded and embodied theories of cognition address the symbol grounding problem by restoring meaning to the core of what it is to know something. Such theories posit that there are non-arbitrary, modally rich mechanisms that *ground* the meaning of our thinking, physical actions, sensations, language, and social interactions (Nathan, 2014). These grounding mechanisms can refer not only to actual objects, interactions, and events, but also to *mental simulations* of objects, interactions, and events (Barsalou, 2009), such as mathematical objects.

Grounded and embodied cognitive theories further embrace the idea that mathematical reasoning *itself* is both grounded and embodied, as basic mathematical concepts arise from our physical interactions with our environment, and those interactions, in turn, serve as *grounding metaphors* (Lakoff & Nuñez, 2000) for more complex and abstract ideas. For example, the conceptual metaphor of *Arithmetic is Motion Along a Path* can serve to ground the abstract mathematical concepts of *greater than* or *less than*.

The current study investigates how, in the course of their mathematical reasoning and proof production, people ground the meaning of abstract mathematical objects and operations with language and action. This research offers an embodied account for the ways that people naturally solve the symbol grounding problem in an educationally relevant area. As such, this work can inform the design of learning environments and alert practitioners to the ways gestures, along with speech, reveal insights about students' higher-order thinking.

2.3. The role of gesture in thinking and learning

According to theories of grounded and embodied cognition, the actions that our bodies engage in directly affect our thinking. One particular type of physical action that researchers have explored as a

mechanism for both conveying and affecting thinking is *gesture*. Gesture does not typically involve acting upon the environment or manipulating objects; instead, gestures are produced in order to express ideas or meanings (see Alibali, Boncoddo & Hostetter, 2014, for discussion). Here, we screen the gesture data to focus only on gestures that accompany speech, called *co-speech gestures*. Within co-speech gestures, we examine *representational gestures* (Kita, 2000)—those gestures for which there is a transparent relationship between form and meaning, such as twisting one's cupped hand to depict a gear turning.

A large body of research has established that gestures convey important information about speakers' cognitive processes (e.g., Goldin-Meadow, 2003; Goldin-Meadow & Alibali, 2013; Hostetter, 2011; Hostetter & Alibali, 2008; McNeill, 1992, 2005). Gestures are a particularly rich source of information about student thinking in mathematics, in part because even when gestures accompany speech, they can convey complementary (i.e., nonredundant) information that reveals a great deal about the speaker's thought process (Alibali & Goldin-Meadow, 1993; Alibali & Nathan, 2012; Church & Goldin-Meadow, 1986). An example of this is provided in Williams et al. (2012), in which a learner justifying a conjecture about a triangle uses a high-pitched verbal sound effect ("Zhoop"), in conjunction with a flat-hands gesture that shows three non-connecting sides of an "impossible" triangle, to explain why the sides would not connect. She did not articulate in her speech her reasoning as to why the triangle would not form—she showed it only with her hands. To date, little research has investigated the gestures that people produce when generating proofs, and no research has explored how dynamic gestures and speech are integrated in proof construction.

2.3.1. Gestures as simulated action

A number of theoretical accounts of the processes that give rise to gestures have been proposed (e.g., Kita & Ozyürek, 2003; McNeill, 2005). One account, the *Gesture as Simulated Action (GSA)* framework (Hostetter & Alibali, 2008, 2018), makes explicit ties to theories of grounded and embodied cognition. According to this framework, gestures derive from simulated actions and perceptual states that people activate while thinking or speaking. When simulating or imagining actions or perceptual states, people activate the same motor areas in the brain that they activate when they are actually producing actions and perceiving stimuli (e.g., Jeannerod, 2001). When this motor activation exceeds an individual's *gesture threshold*—an activation level that depends on individual and social factors, as well as task demands—that individual will produce an overt movement that is commonly recognized as a gesture.

From the perspective of the GSA framework, gestures manifest the embodied nature of reasoning. As such, speakers' gestures provide evidence about the nature of the simulated actions and perceptual states that speakers activate in reasoning. To express such simulations verbally, speakers must "package" those ideas in the linear structure of speech (Kita, 2000). Features of those mental simulations that are successfully packaged in verbal form are expressed in gestures that convey information that is redundant with the co-expressive speech. Features of the simulations that are not selected for verbalization, or that are not successfully packaged in verbal form, may be expressed in gestures that are not redundant with speech (Alibali, Yeo, Hostetter, & Kita, 2017; Hostetter & Alibali, 2008). Thus, although speech and gesture are typically co-expressive, information can be conveyed in one modality but not the other.

2.3.2. Dynamic gestures

Our focus in this work is on a particular subset of representational gestures that we hypothesize may support successful proof construction: *dynamic gestures*, which we define as gestures that represent the progressive transformation of a mathematical object through bodily movement (see Garcia & Infante, 2012). For example, a dynamic gesture might fluidly depict a single triangle dilating or contracting, or two

gears rotating in the same or in opposite directions. *Non-dynamic gestures*, on the other hand, represent objects without directly representing or implying a transformation or manipulation—for example, a single, unmoving triangle or a single, rotating gear. Importantly, the distinction between dynamic and non-dynamic gestures is not between gestures that "move" versus those that "stay still," because moving gestures can depict static, unmoving objects, such as when one traces a triangle with a finger or depicts a gear by tracing a circle. Instead, dynamic gestures represent the progressive *transformation* or *manipulation,* either of a single mathematical object or of multiple mathematical objects related to one another. Given the importance that Harel and Sowder (2007) place on the transformational proof scheme, we hypothesize that dynamic gestures depicting such transformations will be associated with valid proof construction.

Our operationalization of dynamic gestures focuses on the transformative nature of the *representation* depicted, as opposed to the movement of the hand, which aligns with Garcia and Infante's (2012) original use of the term in characterizing the gestures students produce when discussing calculus problems; these authors define dynamic gestures as hand movements that describe mathematical actions or concepts. This differs from how some other scholars have defined dynamic gestures. For example, Marghetis et al. (2014) define dynamic gestures in terms of whether gestural movements are smooth and unbroken. Researchers studying mental rotation tasks (Göksun, Goldin-Meadow, Newcombe, & Shipley, 2013; Newcombe & Shipley, 2015; Uttal et al., 2013) classify dynamic gestures in terms of whether the gesture captures the *intrinsic* components of an object, or the *extrinsic* relationship between the object and other objects. Here, we do not consider depicting a single gear turning via gesture to be dynamic—the gear must be affecting another part of the system (e.g., turning another gear).

The role of dynamic gestures in mathematical reasoning may be different for different types of mathematical tasks. For example, Nathan et al. (2014) found that pedagogical language that alerted learners to the relevance of directed motions was beneficial for solving the triangle task explored here, but detrimental for the gear task. They attributed this difference to the characteristics of the gear task, and particularly the abstract nature of the final gesture people tend to use when they solve this task: tapping back and forth to represent parity (Boncoddo, Dixon, & Kelley, 2010; Schwartz & Black, 1996), rather than physically representing the turning gears. The triangle task, on the other hand, often elicits spatial, relational hand gestures that concretely resemble the variations of a triangle. For this reason, dynamic gestures may play a more important role in promoting reasoning for concrete geometric tasks whose actions correspond to spatial relations, compared to more abstract tasks with no such correspondence, such as those related to the underlying structure of the number system.

2.4. The role of language in thinking and learning

Many theories of grounded and embodied cognition (e.g., Barsalou, 2010; Louwerse & Jeuniaux, 2008) posit that language does not simply express or transmit mental simulations, but that it also plays a part in creating those simulations (Glenberg, 1997). To this end, psycholinguistic research has investigated the degree to which specific semantic features of language may correlate with or even predict mental states. According to the *cognitive reflection model*, the way in which words are used can convey information beyond the words' dictionary meanings that is indicative of psychological characteristics of the speaker (Chung & Pennebaker, 2007). For example, a speaker who is focusing on himself or herself may use more first-person singular pronouns (Raskin & Shaw, 1988; Rude, Gortner, & Pennebaker, 2004). Likewise, a speaker who is engaging in complex thinking may use connective words, such as "and," "because," or "so", because speakers use these words when they join ideas together (Duggleby, Tang, & Kuo-Newhouse, 2016). Thus, the use of connectives may index complex thinking (Clinton, Carlson, & Seipel, 2016).

In addition to words themselves, variations in the *manner* in which words are used may be indicative of variations in underlying mental states (Pennebaker, Mehl, & Niederhoffer, 2003). For example, researchers have found that deceptive speech involves more complex language, likely because expressing dishonest statements is more complicated than expressing honest ones (Duran, McCarthy, Hall, & McNamara, 2010). Thus, both the words a speaker uses and the manner in which those words are used are indicative of mental states. However, very little research has investigated how characteristics of word use shed light on mental states within an educational domain, such as mathematical proof construction (for a notable exception, see González & Herbst, 2013).

It is difficult and time-intensive for human analysts to systematically discern such changes in linguistic features, given their subtlety. One approach to addressing this challenge is to use computerized analytical tools leveraging natural language processing algorithms, which yield various linguistic measures (e.g., McNamara, Graesser, McCarthy, & Cai, 2014). In recent years, some researchers have used such programs in educational contexts, such as for evaluating students' reading comprehension skills (Allen, Snow, & McNamara, 2014) and scientific understanding (Williams & D'Mello, 2010).

2.4.1. Language in proof

An embodied cognition perspective suggests that language plays an integral role in grounding human thinking. Consequently, just as gesture has been shown to affect mathematical thinking and problem solving, certain features or characteristics of speech might help ground people's thinking as they engage in mathematical argument.

The current investigation examines whether certain speech patterns emerge as important for students' successful proving practices. In particular, we explore whether patterns such as "if…then" statements (González & Herbst, 2013) and repetitive speech (McNamara, Graesser, Cai, & Kulikowich, 2011) may be *predictive* of students' ability to verbalize a valid mathematical proof. However, we also explore whether there are other speech patterns that are important for mathematical justification and proof that have not yet been identified. The recent rise of automated text analysis tools presents a unique opportunity to explore students' speech patterns across a variety of dimensions, in order to generate hypotheses about language patterns associated with proof processes.

3. Research questions

To examine the relationship between mathematical proof, learners' speech, and learners' gestures, we posed the following research questions:

1. Is dynamic gesture production associated with constructing valid mathematical proofs?

Although the literature suggests that dynamic gestures may be related to learners' proof practices (given that these gestures transform and manipulate mathematical objects), little research has examined the relations between producing dynamic gestures and engaging in mathematically valid proof activities.

2. Are certain speech patterns associated with constructing valid mathematical proofs?

Past research suggests that "if… then" statements and repetition may be associated with successful mathematical reasoning (González & Herbst, 2013). Recent advances in text-mining software offer the opportunity to explore how a variety of other language patterns might be related to formulating mathematical arguments and proofs.

3. Do dynamic gestures and speech patterns each uniquely predict whether learners verbalize valid mathematical proofs?

It is possible that speech and gesture convey primarily overlapping information as students communicate mathematical arguments; that is, gesture and speech might be largely *redundant*. Alternatively, gestures might capture information about mathematical reasoning not expressed in speech; that is, gesture and speech might convey *distinct* information, suggesting that mathematical proof is a truly multimodal activity. Research in other domains has suggested that gestures can offer novel information about people's reasoning. Understanding whether gestures offer redundant or distinct information about proof construction may provide new insight into students' cognitive processes while proving.

4. Do relations between dynamic gestures and proof vary with task differences?

Prior research suggests the relations of dynamic gesture to proof activities may vary depending on the characteristics of a mathematical task. We hypothesize that the relationship between dynamic gestures and valid proofs will be stronger for a geometry task in which gestures often correspond to concrete spatial relations than for a parity task in which gestures often show abstract properties of the number system.

4. Methods

4.1. Participants

One hundred twenty undergraduates (51% female, *M* age = 19.2 years) from a large, public university in the Midwestern U.S. were prompted to read aloud and provide verbal justifications for two mathematical tasks. Eighty-four participants (70%) identified as Caucasian, 20 as Asian (17%), eight as Hispanic (7%), four as African-American (3%), and four as multiple races and/or ethnicities (3%). Participants' average self-reported SAT/ACT math percentile was 87.01 (*SD* = 13.60), with scores ranging from the 26th percentile to the 99th percentile. Thirty participants (25%) reported that their highest mathematics course was prior to Calculus I in the math sequence, 58 (48.3%) reported that Calculus I was their most advanced math course, 21 (17.5%) reported that Calculus II was their most advanced math course, and 11 (9.2%) reported that a math course above Calculus II was their most advanced math course.

4.2. Tasks and procedure

During the entire session, participants stood in front of a large, interactive whiteboard that displayed the experimental stimuli scaled to each participant's height and arm span. They were asked to read aloud two conjectures that were projected on the white board, and to think aloud as they attempted to provide a justification for each conjecture. We used two conjectures, one of which was drawn from prior studies on middle-school students' justification and proof activities (e.g., Williams et al., 2011; Knuth, Choppin, & Bieda, 2009) and one of which was drawn from studies analyzing speakers' gestures (e.g., Alibali, Spencer, Knox, & Kita, 2011; Boncoddo, Dixon, & Kelley, 2010; Schwartz & Black, 1996). The "triangle task" was a geometric conjecture (i.e., the Triangle Inequality Theorem). The prompt for the triangle task read:

> *Mary came up with the following conjecture: "For any triangle, the sum of the lengths of any two sides must be greater than the length of the remaining side." Provide a justification as to why Mary's conjecture is true or false.*

The "gear task" involved an inference about parity in a system of gears. The underlying mathematical idea involved understanding even/odd patterns in the number system, and then generalizing this pattern to predict even/odd for an unknown variable. The prompt read:

An unknown number of gears are connected together in a chain. If you know what direction the first gear turns, how could you figure out what direction the last gear turns? Provide a justification as to why your answer is true.

These two tasks were chosen because they involved very different kinds of mathematical reasoning: spatial, geometric relationships versus repetitive patterns in the number system. The two tasks were presented in counterbalanced order. The interviewer followed a script for all interactions with participants, including asking them to repeat their justification a second time for each conjecture. This request from the interviewer was added based on pilot work that showed that when initially justifying the conjecture, some participants forgot to give a justification and simply stated that the conjecture was true or false, regardless of the instructions they had been given. Participants rarely changed their justification from the first to the second explanation of the same conjecture; this happened in only 9 out of 240 (3.75%) cases. In these cases, the justification that the participant made last was the one considered for coding. Participants also rated their confidence in their answers on a 1–5 Likert scale. All sessions were videotaped with two cameras, one focusing on a close-up of the participant's upper body and one providing a full body shot.

Participants were randomly assigned to one of three prior action conditions not relevant to the present investigation. Additional information regarding these conditions has been reported elsewhere (Nathan et al., 2014); we briefly describe the three conditions here to provide a complete context for the study. Participants in Condition 1 ($n = 40$) performed *grounding actions*, directed physical actions that were directly related to one of the two tasks (i.e., triangle or gear), prior to engaging in the task itself. Participants in Condition 1 next performed *non-grounding* actions that were irrelevant to the second task, prior to engaging in the second task. Participants in Condition 2 ($n = 40$) performed the same sequence of actions, except that they performed *non-grounding actions* prior to the first task and *grounding* actions prior to the second task. Participants in Condition 3 ($n = 40$) performed *grounding actions* prior to each of the two tasks, and also received a prompt informing them that the actions they performed were directly related to the task they were about to complete. In the present investigation, we collapse across these prior action conditions, because the rate at which participants produced dynamic gestures while generating proofs did not vary across conditions ($\chi^2(2) = 2.09$, $p = .35$). We also checked whether condition was a significant predictor of each of the speech categories discussed in the analysis below, and there were no significant contrasts ($ps > 0.05$). We included experimental condition as a covariate in all analyses presented here, and it was never significant. Thus, in this paper, we do not consider these conditions any further.

4.3. Coding

Videotapes of each session were uploaded into Transana, a software program for transcribing and analyzing video data (Woods & Fassnacht, 2012). Analyses were conducted based on 120 participants generating two justifications each (one for each task), for 240 justifications total.

4.3.1. Gestures

We first identified gesture sequences; a gesture sequence began when a participant lifted his or her hands and ended when the participant dropped his or her hands. Thus, a gesture sequence could consist of a single gesture or several gestures. Each gesture sequence was then coded as either *dynamic* or *non-dynamic*. Fig. 1 provides several examples of dynamic and non-dynamic gesture sequences produced by participants. Appendix A provides more extended examples of gesture sequences that are dynamic. A dynamic gesture sequence involves movement-based operations of an imagined object as it is transformed through multiple states. A non-dynamic gesture sequence conveys an imagined object that is stationary and whose shape, size, orientation, and properties are unchanging—even if the hand itself is moving. Each gesture sequence produced by a participant was coded as dynamic or non-dynamic; note that we coded gesture *sequences* rather than individual gestures (though a sequence could potentially contain only one gesture)[1]. Throughout this paper, when we use the terms "dynamic gestures" or "non-dynamic gestures," we are referring to gesture sequences.

The full video clip for each conjecture was then coded into one of three categories: (1) *non-dynamic* if the participant produced only non-dynamic gesture sequences that represented individual, stationary objects; (2) *dynamic* if the participant produced *at least one* dynamic gesture sequence that depicted a movement-based transformation of an object; or (3) *none* if the participant made no representational gestures. Three independent coders achieved reasonable inter-rater reliability ($\kappa = 0.85$) for coding the justification-level gestures of a random 12.5% subset of all 240 video clips (i.e., 30 videos: 15 triangle, 15 gear).

4.3.2. Speech

Participants' verbalizations were transcribed from the video recordings by a trained undergraduate transcriber. All transcripts were examined by coders and if necessary were edited for correctness. The transcripts were then separated into 240 separate files (one for each conjecture) and cleaned in preparation for entry into text analysis software. Each file was separated into two paragraphs—one for the initial justification, and one for the second justification (since participants were always prompted to repeat their justification). We omitted any instances of the participants' immediately re-reading the conjecture; however, any repetitions of part or all of the conjecture in the middle of the justification were left intact.

The 240 transcripts were then entered into two computerized text analysis tools: Linguistic Inquiry and Word Count (LIWC; Pennebaker, Booth, Boyd, & Francis, 2015) and Coh-Metrix (McNamara, Louwerse, Cai, & Graesser, 2013). LIWC is a dictionary-based program that counts words assigned to more than 70 categories, such as "social process" words (e.g., words relating to family or friends) and "cognitive process" words (e.g., words describing causation or certainty). LIWC's output consists of the percentage of words in a transcript that is used from each dictionary, and thus, LIWC provides a measure of the content of a text. In contrast, Coh-Metrix analyzes the quality of a text. Coh-Metrix provides 108 different indicators of text readability (for a full list, see http://cohmetrix.memphis.edu), which are broadly organized into a range of categories. Some of these categories relate to surface features of the text, such as pronoun incidence or word concreteness, while others relate to deeper features of the text, such as measures of the propositional structure or the cohesiveness of the text. Coh-Metrix's output provides continuous quantitative measures of the degree to which these characteristics are present in a text. Therefore, using both software programs in tandem allows for analysis of the technical aspects of the language gathered from Coh-Metrix, as well as the content and topic of the language from LIWC, providing a more comprehensive description of participants' speech.

We made some adjustments to the LIWC dictionaries, as is a common practice (e.g., McCullough, Root, & Cohen, 2006; Tull, Medaglia, & Roemer, 2005; Vasalou, Gill, Mazanderani, Papoutsi, & Joinson, 2011). In line with previous work (e.g., Walkington, Clinton, Ritter, & Nathan, 2015; Williams-Pierce et al., 2017), we removed polysemous words (i.e., words with multiple meanings) from certain dictionaries in cases in which participants used the words in a manner incongruent with the dictionary the words were listed in. For example,

---

[1] In past publications (e.g., Walkington et al., 2014), we have used the terms *dynamic gestures* and *static gestures*. By using the terms *dynamic gesture sequences* and *non-dynamic gesture sequences* in this paper, we aim to be more precise with our terminology.
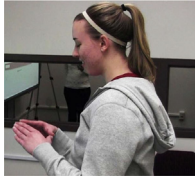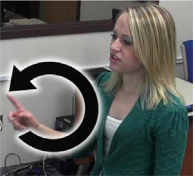
Task	Example of Non-Dynamic Gesture		Example of Dynamic Gesture	
Triangle		*Participant uses both hands to create a full triangle that does not move or change.*		*Participant uses both hands to make two sides of a triangle and fluently moves from making a flattened triangle to a normal triangle.*
Gear		*Participant uses right index finger to show a single gear turning in one direction.*		*Participant uses both hands to show two gears moving in opposite directions.*

Fig. 1. Examples of non-dynamic and dynamic gestures for both tasks.

we removed the word *value* from the "affective processes" dictionary and the words *foot* and *feet* from the "biology" dictionary, since participants used these terms in a mathematical sense and not in the sense that corresponded to the topic of the dictionary. In addition, we created an amended "quantitative words" category omitting words that were included in the prompts themselves, since participants often re-read the prompt in part during their justifications.

Because this is an exploratory study, we included all speech categories from Coh-Metrix in our initial analyses (Stevens, Ronan, & Davies, 2017), with two exceptions. First, we omitted the macro-categories that the creators of Coh-Metrix had derived from a principal components analysis (e.g., narrativity, connectivity, syntactic simplicity). We excluded these macro-categories because we wanted our analyses to focus on fine-grained individual measures, and the corpora from which the principal components were originally derived were substantially different from the texts analyzed here. Second, we omitted those categories specific to the number of paragraphs or the length of the sentences, as we were using natural speech instead of written text. The final list of all speech categories we examined is given in Appendix B. From Coh-Metrix, we used indices related to describing text (e.g., word length), referential cohesion and latent semantic analysis (i.e., overlap), lexical diversity, connectives, situation model measures (e.g., causal verbs), syntactic complexity and pattern density, word-level measures (e.g., number of pronouns), and readability measures. From LIWC, we used language metrics (e.g., word length), function word measures (e.g., number of articles), other grammatical measures (e.g., use of number words), and informal speech measures, as well as measures of words for topics related to affect, socialization, cognitive processes, perception, biological processes, core drives, time orientation, relativity, and personal concerns.

4.3.3. Proof validity

Each justification was analyzed to determine whether the participant constructed a valid, transformational proof for the conjecture. We followed the definition of a transformational proof provided by Harel and Sowder (2007). Table 1 provides examples of a valid proof and an invalid proof for each of the two conjectures. Note that although these examples include only the verbalizations, we attended to participants' speech *and* gestures simultaneously when coding for proof validity, since speakers' gestures often emphasized or clarified key elements of the speakers' reasoning; for example, saying the words "the triangle" while depicting a growing triangle with one's hands suggests the speaker was thinking about general triangles rather than a single instance. Three independent coders achieved acceptable inter-rater reliability ($\kappa = 0.84$) for coding the validity of a random 20% subset of all 240 video clips; half of the subset came from the triangle task, and half

came from the gear task.

4.3.4. Analyses

To answer our research questions about the relationships among dynamic gestures, speech, and proof validity, we calculated zero-order correlation coefficients (point biserial correlations) between (1) proof validity and production of dynamic gesture sequences, (2) proof validity and speech indicators from both LIWC and Coh-Metrix, and (3) production of dynamic gesture sequences and speech indicators. We also conducted the analyses using partial correlations that controlled for the number of words in each proof, and results were similar. All correlations were calculated across the $n = 240$ justifications (i.e., on a trial-by-trial basis). For the speech categories that were significantly correlated with valid proofs across both of the conjectures, we then examined the 20 transcripts that scored highest and the 20 that scored lowest on each category. This provided a holistic sense of what features of the transcripts are driving the significant correlations and enabled us to interpret those correlations in terms of general characteristics associated with valid proofs.

Our approach to the analyses was data-driven, because our aims were exploratory. This approach was chosen because it allows prediction and insight in a manner that is not possible with hypothesis-driven approaches (Schwartz & Ungar, 2015). Data-driven approaches allow for a range of predictors to be explored in an open, unconstrained manner. In contrast, hypothesis-driven approaches are useful for targeted examinations of language use, but findings are limited to the particular types of speech examined (Schwartz et al., 2013). We wanted to inclusively test a wide range of speech categories, so we chose the former.

When testing whether various speech categories were significantly correlated with proof and gesture, we did not perform *p*-value corrections. We made this decision for several reasons. First, requiring the correlations between the speech categories and gesture/proof each to show statistical significance in the same direction for each of two very different tasks (gear, triangle) is a higher benchmark for inclusion than collapsing the data from the different tasks together and examining significant correlations across the entire dataset (i.e., the training set approach discussed by McNamara et al., 2014). More importantly, our goal in this step was to identify *candidates* for the regression analyses—not to identify a final set of predictors from which to draw conclusions. Finally, this inquiry was exploratory, and as such, our aim was to mine patterns in the speech data. Our search for significant variables was performed heuristically, rather than with the strict analytic criteria we would apply in a confirmatory study. This approach limits Type II errors, and it has been used in text-mining studies with small sample sizes such as ours (e.g., Proyer & Brauer, 2018; Robinson, Nyea, & Ickes,

Table 1
Examples of valid and invalid participant proofs for each task.

Task	Validity	Example
Triangle	Valid	So the shortest distance between two points is a straight line, and if you go off of that path, such as by taking any two other straight lines, then it has to be longer than the shortest distance, which is a straight line
	Valid	Say we have triangle ABC, A and B added together must be more than side C or C plus B or C plus A … must be greater than the length of the remaining side… It's true. The two sides of the triangle added together equal the same as the third side of the triangle, then it would just be two lines on top of each other rather than a complete triangle. And if it was less than, the points wouldn't be able to connect, assuming they were straight lines
	Valid	Mary's conjecture is true, because if the one side is long–is longer than the sum of the other two sides then the other two sides won't be able to touch at the top and it won't be a triangle
	Invalid	That isn't true. Uh, it's false, because you could have a triangle where one side is very long and the other two sides are shorter, um very short, and so they add up to a length that is shorter than the longest side
Gear	Valid	When gears are connected, one gear would spin one way, and the gear that's connected to it would spin the opposite way. So if the gear is spinning clockwise, the gear on the next to it would spin counterclockwise… all you'd have to do is divide that number of gears by two. If it is divisible, then it is an even number … it would be clockwise if it starts out clockwise. And if it's not divisible by 2, it would be counterclockwise, if it starts out counterclockwise
	Valid	Um, obviously the gear after the first one turns in the opposite direction, and the next one turns in the opposite direction and so on and so on, so I guess if there's an odd number of gears it will turn in the same direction as the first gear, and if there's an even number of gears it'll turn in the opposite direction
	Valid	So the first gear is spinning one direction, I would think that the other gear, the second gear, would spin in the opposite direction of the first. So each odd numbered gear would spin… the same way, and the second, the fourth, and the sixth would spin opposite way as the first one
	Invalid	Um, I feel that all the gears should turn the same way, because it's a chain reaction, so it should turn in the same direction as the first gear

Note. There are several distinct ways to visualize the triangle conjecture, each of which we considered valid. The first is two short sides each connected to the endpoint of a longer side, and unable to connect at the top ( ). The second is two short sides connected to each other and to one endpoint of a long side and being unable to connect to the long side at its other endpoint ( ). A third is two short sides that are exactly as long as the long side, making a straight line ( ).

2013).

In order to answer our third research question (whether speech and gesture uniquely predict proof validity), we ran logistic regression models using the *lmer* function (Bates, Maechler, Bolker, & Walker, 2015) within the R software environment. In order to control Type I error rate inflation in the regression models, only speech categories significantly correlated with dynamic gesture or valid proofs for both of the two tasks were entered as candidates into the regression models (see similar approaches in Ickes & Cheng, 2011; Walkington et al., 2014). This substantially cuts down the number of predictors tested.

For both models, the dependent variable was whether the participant generated a valid proof for the task (coded as a 0/1); we included participant as a random effect and included control variables of task (i.e., triangle or gear), prior experimental condition (Condition 1, 2, or 3), and word count of the proof (mean centered) as fixed effects. To control for students' mathematical background, we included a predictor indicating their most advanced previous math course (Below Calculus I, Calculus I, Calculus II, and Above Calculus II). We also collected data on participants' self-reported ACT/SAT math score (converted to a percentile), but we chose not to use these data in the analysis because it was missing for 5 participants, and the distribution was skewed towards the upper end of the scale. However, we repeated all regression analyses with ACT/SAT percentile (rather than highest math course) as a control variable and found no differences in the results relevant to our research questions. The only difference was that ACT/SAT math scores had a stronger association with performance on the triangle task ($p = .008$) than on the gear task.

In the first regression analysis, we added predictors to the model in the following order: dynamic gestures, speech indicators significantly correlated with dynamic gestures (regardless of whether they were correlated with valid proofs), and speech indicators significantly correlated with valid proofs (regardless of whether they were correlated with dynamic gestures). Predictors were tested for inclusion in the model using the *anova* function, which tests for significant reductions in deviance using a χ^2 reference distribution. In the second analysis, we added predictors in the following order: speech indicators significantly correlated with valid proofs (regardless of whether they were correlated with dynamic gestures), speech indicators significantly correlated with dynamic gestures (regardless of whether they were correlated with valid proofs), and dynamic gestures. Thus, in one analysis we added terms for dynamic gestures into the model first, and in the other

analysis we added dynamic gestures into the model last. Both analyses resulted in the same final model. In both analyses, we tested for interactions between speech indicators, dynamic gestures, and control variables. We also tested interactions of task type (triangle vs. gear) with dynamic gesture production and speech patterns in order to determine if there were task-dependent differences in the predictive power of gesture or speech for constructing a valid proof. All speech predictors were mean-centered; dynamic gesture was a two-level categorical variable (i.e., any dynamic gestures versus no dynamic gestures). Predictors in the final models were checked for multicollinearity (all VIFs < 2.0). It was also of interest to quantify the proportion of variance explained by the various models, in order to estimate the size of the effects. For this purpose, we used Xu's (2003) metric, Ω^2, which gives the percentage of reduction in residual variance between a null model and a full model.

5. Results

For the triangle task, 50.00% of the participants constructed a valid proof, whereas for the gear task, 40.83% of the participants constructed a valid proof. In terms of dynamic gestures across the two tasks, 46.67% of the participants made at least one dynamic gesture sequence; this was the case for 35.00% of the participants during the triangle task, and 58.33% during the gear task. In addition, the average length of participants' justifications for both tasks was approximately 140 words ($SD = 71.87$); for the triangle task, the average length was approximately 142 words ($SD = 76.95$) and for the gear task, the average length was 138 words ($SD = 66.69$).

5.1. Research question #1: dynamic gesture and valid proofs

Producing at least one dynamic gesture sequence was significantly associated with verbalizing a valid proof for both the triangle task ($r = 0.454$, $p < .001$; Cohen's $d = 1.09$) and the gear task ($r = 0.255$, $p = .005$, Cohen's $d = 0.54$). For the triangle task, if a participant made a dynamic gesture sequence, their chance of getting the proof correct was 80.95%; if they did not, their chance was 33.33%. For the gear task, these probabilities were 51.43% and 26.00%, respectively. We also tested "producing only non-dynamic gestures" (i.e., producing gesture sequences that did not contain dynamic gestures) as a predictor. This category is distinct from the category of "no dynamic gestures," as

Table 2
Significant correlations between speech indicators and accuracy for each task.

Coh-Metrix/LIWC indicator	Indicator category	Significant correlations	
		Triangle task	Gear task
Type-token ratio – all words[C] (*LDTTRa*)	Repetition in justification	−0.250**	−0.48***
Type-token ratio – content word lemmas[C] (*LDTTRc*)	Repetition in justification	−0.279**	−0.500***
Number of modifiers per noun phrase[C] (*SYNNP*)	Syntactic variety/Complexity	0.248**	0.222*
Discrepancy words[L]	If…then statements	0.264**	0.202*
Temporal connectives[C] (*CNCTemp*)	If…then statements	0.265**	0.240**
First person singular pronouns[C] (*WRDPRP1s*)	Self-conscious statements	−0.260**	−0.403***
I[L]	Self-conscious statements	−0.259**	−0.414***
Pronouns[L]	Self-conscious statements	−0.246**	−0.294**
Pronouns[C] (*WRDPRO*)	Self-conscious statements	−0.204*	−0.321***
Present tense[L]	Self-conscious statements	−0.280**	−0.328***
Insight words[L]	Self-conscious statements	−0.305***	−0.407***

* *p* < .05.
** *p* < .01.
*** *p* < .001.
[L] Measure from LIWC.
[C] Measure from Coh-Metrix.

it does *not* include cases in which there was no gesture of any kind. We found that producing *only* non-dynamic gesture sequences was significantly negatively correlated with valid proofs for the triangle task (*r* = −0.183, *p* = .045), but not for the gear task.

5.2. Research question #2: speech and valid proofs

Table 2 presents the significant correlations between speech indicators and proof validity for each task. We grouped speech indicators into categories based on an analysis of proof transcripts that scored high versus low on each speech indicator; by examining the transcripts with the highest scores for an indicator and comparing them to those with the lowest scores, we were able to deduce which language patterns were captured by each speech indicator. This allowed us to inductively determine which indicators were capturing related language constructs. Language categories significantly associated with valid transformational proofs included *repetition in one's justifications* and *syntactic variety*. Valid proofs were also significantly associated with the use of logical statements, measured by *discrepancy* words (e.g., "should, would, could") and *temporal connective* words (e.g., "then"). In addition, we found that "self-conscious statements" (e.g., "I don't know" or "I'm not sure") were significantly negatively correlated with valid proofs, as indicated by four measures of pronoun use, the present tense, and *insight* words (e.g., "know," "understand").

5.3. Research question #3: dynamic gesture and speech

In order to measure whether dynamic gesture and the speech indicators from Table 2 each uniquely predict whether students produce valid proofs, we first examined whether dynamic gestures and speech were significantly correlated with one another. This allowed us to examine the degree of overlap between the two modalities and ensured that any effect of dynamic gesture in the regression model would be an effect over and above any relationship between gesture and speech. Table 3 presents the significant correlations between dynamic gestures and the speech indicators. There were two positive, significant correlations between dynamic gesture and speech indicators related to logical statements (*all connectives* and *temporal connectives*). Additionally, dynamic gestures were significantly negatively correlated with measures indicating the use of "self-conscious statements" (see Table 3).

Table 4 displays the results of the logistic regression analysis examining whether speech indicators and dynamic gesture production each explained unique variance in models predicting proof validity. The first column of Table 4 shows the base model with control variables

Table 3
Significant correlations between speech categories and dynamic gestures.

Coh-Metrix/LIWC indicator	Indicator category	Significant correlations	
		Triangle task	Gear task
All connectives[C] (*CNCAll*)	If…then statements	0.267**	0.225*
Temporal connectives[C] (*CNCTemp*)	If…then statements	0.248**	0.214*
Present tense[L] (Present)	Self-conscious statements	−0.188*	−0.207*
Insight words[L] (Insight)	Self-conscious statements	−0.214*	−0.337***
Cognitive processes words[L] (CogMech)	Self-conscious statements	−0.228*	−0.321***

* *p* < .05,
** *p* < .01,
*** *p* < .001.
[L] Measure from LIWC,
[C] Measure from Coh-Metrix.

only; the second column shows the model with control variables and those speech predictors significantly associated with dynamic gesture and valid proof (i.e., the speech categories in Tables 2 and 3); the third column shows the model with control variables and the dynamic gesture predictor only; the fourth column shows the model with control variables, speech predictors significantly associated with dynamic gesture and valid proof, and the dynamic gesture predictor; and the fifth column shows the model with two-way interaction terms.

Examining the null model with control variables only (first column), we see that participants whose most advanced math course was prior to Calculus I were less likely to generate valid proofs than those whose most advanced math course was above Calculus II (*p* = .002). Longer proofs were also more likely to be correct (*p* = .008). The model with control variables and speech predictors (second column) shows that participants were more likely to generate valid proofs on the triangle task than on the gear task (*p* = .013). We also see that three speech indicators significantly predicted whether participants articulated valid transformational proofs: the use of *discrepancy words* (e.g., logical statements like "if…then;" *p* = .005), a lower *type-token ratio* for content words (i.e., more word repetition; *p* = .024), and the use of fewer *cognitive processes* words (i.e., avoiding self-conscious statements; *p* = .003). The model with control variables and dynamic gestures (third column) indicated again that participants whose most advanced math course was prior to Calculus I were less likely to generate valid proofs than those whose most advanced math course was above

Table 4
Results of five logistic regression models predicting correct/incorrect proofs.

	1. Control Variables Only	2. Control + Speech	3. Control + Gesture	4. Control + Speech + Gesture	5. Control + Speech + Gesture + Interactions
Random Component – Participant ID (variance)	3.21	4.97	2.69	2.38	4.83
	B(SE)^{Sig}	B(SE)^{Sig}	B(SE)^{Sig}	B(SE)^{Sig}	B(SE)^{Sig}
(Intercept)	0.95(0.91)	−0.13(1.12)	−0.46(0.95)	−1.07(0.97)**	−0.05(1.25)
Highest Math Below Calc I	−3.46(1.12)**	−2.91(1.51)	−3.46(1.15)**	−2.48(1.12)*	−3.17(1.72)
Highest Math Calc I	−1.28(0.92)	−0.94(1.18)	−1.11(0.93)	−0.62(0.93)	−0.93(1.23)
Highest Math Calc II	−1.06(1.03)	−1.20(1.32)	−0.94(1.05)	−0.94(1.06)	−1.30(1.40)
Highest Math Above Calc II	(ref.)	(ref.)	(ref.)	(ref.)	(ref.)
Condition (Non-grounding actions)	(ref.)	(ref.)	(ref.)	(ref.)	(ref.)
Condition (One grounding action)	0.23(0.43)	0.54(0.55)	0.12(0.46)	0.26(0.50)	0.52(0.61)
Condition (Two grounding actions)	−0.06(0.57)	0.14(0.71)	−0.29(0.59)	. −0.18(0.59)	−0.20(0.74)
Task - Gear	(ref.)	(ref.)	(ref.)	(ref.)	(ref.)
Task - Triangle	0.63(0.36)	1.51(0.61)*	1.28(0.44)**	1.85(0.59)**	0.62 (0.77)
Word count (centered)	0.009(0.003)**	0.006(0.004)	0.007(0.003)	0.001(0.004)	0.004(0.005)
Discrepancy words (centered)		0.57(0.20)**		.0.41(0.13)**	0.56(0.26)*
Cognitive processes words (centered)		−0.27(0.09)**		−0.17(0.06)**	−0.23(0.10)*
Type-token ratio – Content words (centered)		−8.24(3.66)*		−8.64(3.23)**	−9.41(4.33)*
Dynamic Gestures			2.24(0.54)***	1.56(0.54)**	0.12(0.79)
Dynamic × Task-Triangle					3.36(1.65)*
Model Deviance	281.9	226.0	257.1	215.9	208.3
Deviance Reduction Compared to Model 1		19.8%	8.8%	23.4%	26.1%

Note. "(ref.)" denotes the reference category. Columns give the raw regression coefficients (B) and their respective standard errors, as well as their significance level (p-value). Raw B coefficients can be transformed into Odds Ratios by exponentiating the B coefficient. LIWC and Coh-Metrix predictors are centered (rather than normalized) to keep them on their original scale and allow for comparability to other LIWC/Coh-Metrix studies. Word count (centered) varied from −115 to 307, discrepancy words (centered) varied from −4.8 to 9.7, cognitive process words (centered) varied from −11 to 21, and type token ratio (centered) varied from −0.25 to 0.42. Thus, although type-token ratio appeared to have the largest effect, this is in part because it had the smallest scale.

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

Calculus II ($p = .003$), and that participants were more likely to generate valid proofs on the triangle task than on the gear task ($p = .004$). In addition, this model indicates that dynamic gestures were strongly positively associated with generating valid proofs (Odds Ratio = 9.35, $p < .001$).

Examining the model that included both speech and dynamic gesture predictors, but no interactions (fourth column), we see that the same three speech indicators significantly predicted whether participants verbalized valid proofs: the use of *discrepancy words* (i.e., logical statements; $p = .002$), a lower *type-token ratio* for content words ($p = .008$), and the use of fewer *cognitive processes* words ($p = .006$). We also found that producing dynamic gestures was a significant, positive predictor of verbalizing a valid proof ($p = .004$), with the presence of a dynamic gesture sequence associated with an increase in the relative odds of formulating a valid proof (i.e., an odds ratio) of 4.75. As shown in the last row of Table 4, once we included speech indicators alone as predictors, the model's overall deviance was reduced by 19.8% and when adding dynamic gesture alone as a predictor, the model's deviance was reduced by 8.8%. However, when adding both the dynamic gesture and speech predictors, the model's deviance was reduced by 23.4%—which is an additional 3.6% over speech alone and 14.6% over gesture alone. Thus, together, the speech and gesture predictors combined explained approximately 23% of the variance in whether participants generated valid transformational proofs. These are novel findings; notably, these findings quantify the degree to which dynamic gesture production predicts proof performance, above and beyond indicators in participants' speech.

5.4. Research question #4: task differences

There was a difference between the two tasks (triangle vs. gear) that

emerged when we included interaction terms that allowed effects to vary across the two tasks. The final column of Table 4 displays results for the model with interaction terms; we found a significant interaction of task and dynamic gestures in predicting proof validity. The simple effects contrasts stemming from the interaction term indicate that producing a dynamic gesture sequence was associated with a significantly higher likelihood of formulating a correct proof for the triangle task ($p = .002$), but not for the gear task ($p = .88$). Thus, when the interaction term was added to the model that controlled for other predictors such as speech indicators, it provided the additional information that the effect of dynamic gesture on valid proof production was being driven by the triangle task *only*. This result suggests that the identification and coding of dynamic gestures may be useful for understanding proof validity on some mathematical tasks, but not others. Adding the interaction of gesture and task to the model reduced deviance another 2.7%, for a total reduction of 26.1%

6. Discussion

We found that producing dynamic gestures was positively and significantly associated with formulating a valid mathematical proof (Research Question #1); however, this was the case for the triangle task and not for the gear task. We also found that certain speech patterns were significantly correlated with generating valid proofs (Research Question #2).

We also examined the relationship between gesture and speech and found that dynamic gestures were related to speech indicators for connective words used in logical statements for both the triangle and the gear task. Additionally, we found that making self-conscious statements such as "I don't know" or "I'm not sure" was significantly negatively associated with generating valid proofs, as well as with

producing dynamic gesture sequences.

The logistic regression analysis showed that including both dynamic gestures and speech indicators that were significantly correlated with valid proofs in the statistical model reduced the overall deviance of the model, relative to models that included only speech indicators or only dynamic gesture production. These findings suggest that each modality explained unique variance in participants' likelihood of generating a valid proof (Research Question #3).

Finally, there was a significant interaction of task and dynamic gesture production in the logistic regression analyses, indicating that the association of dynamic gestures and formulating valid proofs held true for the triangle task *only,* and not for the gear task. We delve more deeply into each research question in turn below.

6.1. Gesture and valid proof

Dynamic gesture production was significantly correlated with generating a valid proof. Furthermore, once we accounted for the speech indicators that were significantly correlated with valid proof production, dynamic gesture production still accounted for some of the variability in participants' likelihood of generating a valid proof for a Euclidean geometric conjecture, but not for a number theory conjecture relating to a gear system. Past work has shown that dynamic gestures are prevalent during proving practices, both for novices (Walkington et al., 2014) and for experts (Marghetis et al., 2014), across a range of geometric and non-geometric conjectures. The current findings converge with this past work to underscore the importance of dynamic gestures in understanding students' mathematical reasoning. Consequently, these results begin to paint a more nuanced picture of the conditions under which dynamic gesture production may be particularly important to interpreting student thinking.

Our findings suggest that producing dynamic gestures is associated with valid, transformational proof generation specifically for geometric conjectures, due to properties of dynamic gestures that do not overlap with speech. Other researchers have provided evidence that speakers often convey information in gestures that they do not convey in speech (e.g., Goldin-Meadow, 1999; Kita, 2000; McNeill, 1992; Pine, Bird, & Kirk, 2007). This work not only extends these findings to the particular discursive activity of geometric proof construction, but also demonstrates the relative contributions of particular kinds of gestures and particular categories of speech. These results build upon prior research showing that dynamic gestures indicating movement or transformation of depicted entities are important in spatial reasoning tasks (e.g., Göksun et al., 2013; Newcombe & Shipley, 2015; Uttal et al., 2013).

One implication of these findings is that researchers, teachers, and others performing assessments should attend to dynamic gestures—in addition to speech—when evaluating mathematical arguments. Indeed, if dynamic gestures capture unique information about students' reasoning in a way that goes above and beyond their speech patterns, then attending to, understanding, and analyzing these gestures may be a critical element of enacting valid and useful assessments of students' mathematical thinking. Students who poorly express verbal proofs may actually convey crucial information in their gestures that they cannot yet express in speech (Goldin-Meadow & Sandhofer, 1999; Goldin-Meadow & Singer, 2003). If a teacher can attend to what students are expressing with their hands, he or she may be able to support students in expressing that proof in words—perhaps as a scaffold for a subsequent formal, written proof. It also may be the case that inhibiting students from gesturing (e.g., by having them hold a pencil; see Walkington et al., 2014) may allow for less rich and informative inferences to be drawn about students' understanding.

6.2. Speech and valid proof

Using automated text analysis tools to extract speech characteristics during proof, we identified specific linguistic features that are key predictors of generating valid proofs. We found that using self-conscious "I" statements was significantly negatively correlated with generating a valid proof, but measures of repetition, syntactic variety, and logical statements were all significantly positively related to generating a valid proof. Our findings for logical statements echo those of González and Herbst (2013), who demonstrated that students use "if…then" statements while discussing a geometry proof to connect different ideas within an argument or to link different arguments together. We also found that repetitive language and connective words are used to connect ideas together, in the same way that they index cohesion in texts (McNamara et al., 2011). Although findings related to logical statements (i.e., "if… then") have been discussed in prior research, the importance of repetition of words in the argument discovered in the present study is a novel finding. The speech measures we found to be significantly associated with successfully verbalizing a valid proof appear to indicate cohesive, logical argumentation—what Harel and Sowder refer to as *logical inference*, which they emphasize as particularly crucial for constructing successful deductive proofs:

> Logical inferencing ability is a basic tool for the process of proving in mathematics and likely enters also into many justifications of a less sophisticated sort. But logic is central to the deductive proof schemes. For example, the transformational proof scheme, which constitutes the essence of the proving process in mathematics and is expected to develop with at least college-bound students and mathematics major students, should be present in students' mathematical behavior. (p. 23)

Although "if… then" statements appear to be particularly important, using text mining approaches allowed for a more flexible and inclusive categorization of logical statements during proofs. For example, many students in our sample made "if…then"-type statements without actually saying the word "then"; for example, one participant said, "If it was odd, it would be going in the opposite direction." Although logical statements of this kind might be missed when coding only for "if… then" statements, our categories of discrepancy words and temporal connectives captured a wide range of language patterns that showed logical deduction. Examining the 20 transcripts that scored highest on the discrepancy words indicator, we found that although 10 had formal "if… then" statements, only 1 of the 20 lacked logical statements. Of the top 20 transcripts for temporal connectives, 18 contained "if… then" statements, and none lacked logical statements.

The use of text-mining tools to examine speech patterns in tasks such as oral proof presents the potential to explore student reasoning in new ways. Such tools enable researchers to evaluate text from much larger corpora than could be handled by human coders, and they provide broader information about the nature of the transcribed speech than human coders could realistically extract without hundreds of hours of coding. Furthermore, these tools may provide a means for examining the reliability of human coders in an automated way, so they could potentially assist in establishing inter-rater reliability. For example, if we observed a proof that had been coded as valid, but it contained frequent self-conscious statements and lacked logical statements, we might investigate whether the human coder had miscoded the proof's validity. Intelligent tutoring systems and student assessments could benefit from leveraging such text analysis programs to evaluate students' written or oral responses to various tasks (McNamara et al., 2012; Williams & D'Mello, 2010).

6.3. Dynamic gesture and speech as unique predictors

The third main finding that emerged from this study is that, although speech and dynamic gesture production are significantly correlated with one another and are each predictive of generating a valid proof, both modalities contribute *uniquely* to models predicting valid proof generation. Although prior work has established that the speech category of "if…then" statements is related to valid arguments

(González & Herbst, 2013), this study offers new insights by combining speech and gesture analyses and examining both their overlapping and unique contributions.

The significant association between valid proofs and dynamic gestures in ways that go beyond speech patterns underscores the inherently grounded and embodied nature of mathematical reasoning, providing further support for and extending theories of embodied mathematical knowledge (e.g., Lakoff & Nuñez, 2000; Nathan, 2014; Nemirovsky & Ferrara, 2009). The study of embodiment is an important ongoing area of exploration in mathematics (Schoenfeld, 2016), and investigations of how different kinds of mathematical reasoning become grounded are important for understanding how mathematical thinking and learning take place (Shapiro, 2010). Indeed, mathematics is often seen as an abstract domain, disconnected from the senses and from the body (Lakoff & Nuñez, 2000), with mathematical arguments or proofs being particularly abstract. Our findings contribute to a growing consensus that this assumption is not necessarily valid.

6.4. Task differences

The regression analyses revealed a noteworthy statistical interaction between task and dynamic gestures, indicating that production of dynamic gestures was significantly related to formulating valid proofs for the triangle task, but not for the gear task. One difference between the two tasks is that the gear task was more difficult, perhaps because it involved more abstract mathematical thinking. The gear task had a 41% success rate for valid proof, compared to a 50% success rate for the triangle task. This difference was also manifested in the positive coefficients for the triangle task in the regression models. The gear task also elicited more dynamic gestures than the triangle task (on 58% of proof attempts versus 35%). Thus, when examining correlations, the triangle task had a stronger correlation between dynamic gesture and proof ($r = 0.454$) than the gear task ($r = 0.255$). However, it is worth noting that both of these correlations are positive and each differs significantly from zero. The weaker correlation for the gear task manifested itself in the final regression model, which found dynamic gesture to be significantly related to proof performance on the triangle task, but not significantly related to proof performance on the gear task.

It is worth noting that adults typically use their hands to gesture during problem-solving tasks involving gear systems (Alibali et al., 2011; Goldin-Meadow, 2003; Schwartz & Black, 1996), so we might expect those gestures that depict the relationships between gears in the system to be associated with valid proof. However, one study investigating whether participants could solve a gear task similar to ours found that most participants were able to arrive at a correct solution, regardless of whether they were allowed to gesture or were prohibited from gesturing—although gesture inhibition did influence the specific strategies participants chose to rely on (Alibali et al., 2011). These authors found that participants who were allowed to gesture tended to simulate the actions of the gears using their hands, whereas those who could not gesture were more likely to focus on the number of gears. Although participants in that study were asked simply to make a prediction about the movement of the gears, and not to formulate a justification about the gears' movement, the results here align with their finding that gesture did not seem to affect participants' *accuracy* on this type of gear task. One interesting hypothesis is that using gestures to embody an abstract idea like parity may actually be harmful in some cases, because it tends to focus learners on concrete, salient, spatial relations (e.g., relations among the gears themselves), rather than the hypothetical abstractions related to the underlying structure of the number system (see Alibali & Kita, 2010; Nathan et al., 2014). Nevertheless, additional research investigating the role of gesture in reasoning during this task is needed. The task differences that we observed suggest the need for a more nuanced theory of embodied cognition that takes into account the properties of the tasks and the relations between a task's surface features and its deep structure.

7. Conclusion

This study showed that speech and dynamic gestures each reveal students' mathematical proof practices related to geometry, a finding that aligns with recent research into the multimodal and embodied nature of expert mathematicians' proving behaviors (Marghetis et al., 2014) as well as with research on other mathematical reasoning activities (e.g., Alibali & Nathan, 2012; Broaders et al., 2007; Cook, Mitchell, & Goldin-Meadow, 2008). More specifically, dynamic gestures and speech conveying logical inference play an important role in mathematical reasoning during oral proof construction, although the two modalities may be differentially involved depending on the nature of the task.

The present work goes beyond extant descriptive accounts of gesture and successful mathematical reasoning. In particular, we show important and theoretically-grounded *patterns* that reveal the manner in which gesture, speech, and mathematical proof practices are related across different kinds of tasks. These patterns have begun to show consistency across different studies of students' mathematical reasoning (e.g., Abrahamson, 2015; Garcia & Infante, 2012; González & Herbst, 2013; Nathan & Walkington, 2017). Identifying such patterns is a critical step in understanding the role of gestures in mathematics reasoning, and is a necessary step towards the investigation of causal relationships. Thus, with this work, we seek to bridge prior descriptive accounts of gesture and mathematical justification to future work designed to investigate these causal relations.

7.1. Limitations

Our analyses of participants' language were limited by the particular categories included in the LIWC and Coh-Metrix computer programs. Although we considered more than 100 speech indicators in our initial analyses, using a different approach to coding and analyzing participants' speech might yield different results. In addition, given our exploratory, data-driven approach and small sample size, we did not perform conventional *p*-value corrections. Exploratory approaches that attempt to narrow down many text categories to a few should be interpreted with caution. Such approaches can provide a basis for future studies that use confirmatory approaches with stricter criteria in their analyses.

Additionally, we focused on one particular category of gestures—dynamic representational gestures—in a manner that excludes some other potentially relevant gestures, including metaphoric and deictic gestures, as well as other means for categorizing gestures, such as Kendon's (2004) distinction between gestures conveying enactment, depiction, and modeling. Our choice in focusing on dynamic gestures stemmed from our hypotheses about the nature of transformational proofs, as well as from existing work suggesting that dynamic gestures are particularly frequent and important during proof production (e.g., Marghetis et al., 2014).

Despite these limitations, we believe that this exploratory study provides initial insights into the types of language patterns that might support mathematical reasoning, as well as ways that certain types of gestures are involved in mathematical reasoning. This work thus serves as a starting point for research aiming to better elucidate how learners employ their language capabilities and their body-based resources in service of mathematical reasoning.

7.2. Future directions

This study showed that both dynamic gestures and logical statements are significantly associated with valid mathematical reasoning; however, based on these data, we cannot make causal claims about these relationships. Future research could explore causal claims by manipulating students' gestures or speech through prompts, instructions, or structuring of the environment. If dynamic gestures were

found to play an important causal role in students' formulations of mathematical arguments, then instructional scaffolding could guide students to produce such gestures, following other research showing benefits for directed gestures on mathematics reasoning (e.g., Broaders, Cook, Mitchell, & Goldin-Meadow, 2007; Smith, King, & Hoyte, 2014). Furthermore, having teachers themselves model dynamic gestures may facilitate students' own use of such gestures (Alibali & Nathan, 2007). With the increased popularity and introduction of touch-based classroom technologies such as SMART boards and iPads, the potential exists for such technologies to be used to detect students' gestures—for example, using an Xbox ® Kinect ® or comparable system—and to guide students in producing dynamic gestures. Students' mathematical reasoning might benefit, for example, if they are called upon to match the movements of video game characters (Nathan & Walkington, 2017). However, the effectiveness of such approaches may vary as a function of several factors, including the content domain, the nature of the embodied actions, and the nature of the task.

If future work were to find that speech patterns play a similar causal role in supporting students' mathematical argumentation, this could have implications for the teaching of mathematical proof practices. A stronger understanding of the verbal and discursive structure of valid mathematical arguments could lead to recommendations for how teachers might support students who struggle to express themselves mathematically. Making explicit that valid mathematical arguments tend to use logical statements and repetition of key ideas may provide guidance for scaffolding that teachers and curriculum developers alike could implement to support students' development of proof practices, such as having students use "scripts" that would facilitate their production of logical "if…then" statements, as documented by Rummel and Spada (2005).

Based on the findings from this study, we plan to further investigate the effectiveness of dynamic gestures for improving students' reasoning about tasks from different mathematical and non-mathematical domains. For example, interactive math games and learning environments that use players' actions are being used to promote conceptual understanding of related rates (Abrahamson, 2015), rational numbers (Williams-Pierce, 2016), elementary school geometry (Smith et al., 2014) and algebraic symbol manipulation (Ottmar & Landy, 2016). In this vein, we have developed a video game focusing on middle- and high-school geometry content that leverages recent advances in motion capture technology. The game directs students to make particular directed actions based on the dynamic gestures we have observed being spontaneously produced by successful students. It also uses real-time camera data to evaluate whether students perform the appropriate dynamic actions, and then prompts students to justify their reasoning about the truth of a geometric conjecture that is consistent with the mathematical relationship modeled by the dynamic actions. By scaffolding students' uses of dynamic gestures in this way, students generate key body-based mathematical insights related to geometric properties and relationships (Nathan & Walkington, 2017). We plan to develop this game into a classroom-based intervention that supports students' mathematical reasoning and proof development through dynamic gestures and speech. Given the current findings that dynamic gestures are particularly important for justifying geometric proofs in particular, we anticipate that this game will allow us to extend the findings presented here into a school-based context with students from a K-12 population engaged in academic tasks.

7.3. Concluding remarks

This study demonstrates that dynamic gestures and speech conveying logical inference are instrumental to understanding students' proof generation. We found that dynamic gestures were significantly correlated with generating valid proofs in geometry. Additionally, using computerized text analysis software, we were able to identify key speech patterns significantly associated with valid proof generation

across multiple mathematical conjectures. Finally, we showed that producing dynamic gesture sequences is associated with valid geometric proof generation above and beyond speech. Although the precise nature of the information conveyed by each modality still remains to be studied, our work demonstrates that speech and gesture can each serve as grounding mechanisms during mathematical reasoning, and that the two modalities quantifiably convey both overlapping and distinct information. At the same time, however, this work suggests that different mathematical tasks have different affordances for embodied simulation. Although dynamic gestures may support the grounding of symbols in some tasks, that grounding may be more or less successful, depending upon the nature of those symbols. In sum, this research not only demonstrates the importance of attending to students' gestures in addition to the speech they produce during mathematical proof generation, but also reveals the fundamentally embodied basis of mathematical proof.

Acknowledgements

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cedpsych.2019.01.012.

References

Abrahamson, D. (2015). The monster in the machine, or why educational technology needs embodied design. In V. R. Lee (Ed.). *Learning technologies and the body* (pp. 21–38). New York: Routledge.

Alibali, M. W., Boncoddo, R., & Hostetter, A. B. (2014). Gesture in reasoning: An embodied perspective. In L. Shapiro (Ed.). *The Routledge handbook of embodied cognition* (pp. 150–157). New York: Routledge.

Alibali, M. W., & Goldin-Meadow, S. (1993). Gesture-speech mismatch and mechanisms of learning: What the hands reveal about a child's state of mind. *Cognitive Psychology, 25*, 468–523.

Alibali, M. W., & Kita, S. (2010). Gesture highlights perceptually present information for speakers. *Gesture, 10*(1), 3–28.

Alibali, M. W., & Nathan, M. J. (2007). Teachers' gestures as a means of scaffolding students' understanding: Evidence from an early algebra lesson. In R. Goldman, R. Pea, B. J. Barron, & S. Derry (Eds.). *Video research in the learning sciences* (pp. 349–365). Mahwah, N.J.: Erlbaum.

Alibali, M., & Nathan, M. (2012). Embodiment in mathematics teaching and learning: Evidence from students' and teachers' gestures. *The Journal of the Learning Sciences, 21*(2), 247–286.

Alibali, M. W., Spencer, R. C., Knox, L., & Kita, S. (2011). Spontaneous gestures influence strategy choices in problem solving. *Psychological Science, 22*(9), 1138–1144.

Alibali, M. W., Yeo, A., Hostetter, A. B., & Kita, S. (2017). Representational gestures help speakers package information for speaking. In R. B. Church, M. W. Alibali, & S. D. Kelly (Eds.). *Why gesture? How the hands function in speaking, thinking, and communicating* (pp. 15–37). Amsterdam: John Benjamins.

Allen, L. K., Snow, E. L., & McNamara, D. S. (2014). The long and winding road: Investigating the differential writing patterns of high and low skilled writers. In J. Stamper, S. Pardos, M. Mavrikis, & B. M. McLaren (Eds.). *Proceedings of the 7th international conference on educational data mining, London, UK* (pp. 304–307). .

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617–645.

Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1521), 1281–1289.

Barsalou, L. W. (2010). Grounded cognition: Past, present, and future. *Topics in Cognitive Science, 2*, 716–724.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Boncoddo, R., Dixon, J. A., & Kelley, E. (2010). The emergence of a novel representation from action: Evidence from preschoolers. *Developmental Science, 13*, 370–377.

Broaders, S. C., Cook, S. W., Mitchell, Z. A., & Goldin-Meadow, S. (2007). Making children gesture brings out implicit knowledge and leads to learning. *Journal of Experimental*

Psychology: General, 136, 539–550.

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

Chung, C., & Pennebaker, J. (2007). The psychological functions of function words. In K. Fielder (Ed.). *Social communication* (pp. 343–359). New York: Psychology Press.

Church, R. B., & Goldin-Meadow, S. (1986). The mismatch between gesture and speech as an index of transitional knowledge. *Cognition, 23*, 43–71.

Clinton, V., Carlson, S. E., & Seipel, B. (2016). Linguistic markers of inference generation while reading. *Journal of Psycholinguistic Research, 45*(3), 553–574.

Cook, S., Mitchell, Z., & Goldin-Meadow, S. (2008). Gesturing makes learning last. *Cognition, 106*, 1047–1058.

De Vega, M., Glenberg, A., & Graesser, A. (2012). *Symbols and embodiment: Debates on meaning and cognition.* Oxford University Press.

Duggleby, S. J., Tang, W., & Kuo-Newhouse, A. (2016). Does the use of connective words in written assessments predict high school students' reading and writing achievement? *Reading Psychology, 37*(4), 511–532.

Duran, N. D., McCarthy, P. M., Hall, C., & McNamara, D. S. (2010). The linguistic correlates of conversational deception: Comparing natural language processing technologies. *Applied Psycholinguistics, 31*, 439–462.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*, 3–71.

Garcia, N., & Infante, N. E. (2012). Gestures as facilitators to proficient mental modelers. In L. R. Van Zoest, J.-J. Lo, & J. L. Kratky (Eds.). *Proceedings of the 34th annual meeting of the North American chapter of the international group for the psychology of mathematics education* (pp. 289–295). Kalamazoo, MI: Western Michigan University.

Glenberg, A. (1997). What memory is for. *Behavioral and Brain Sciences, 20*, 1–55.

Glenberg, A. M., Gutierrez, T., Levin, J. R., Japuntich, S., & Kaschak, M. P. (2004). Activity and imagined activity can enhance young children's reading comprehension. *Journal of Educational Psychology, 96*(3), 424–436.

Göksun, T., Goldin-Meadow, S., Newcombe, N., & Shipley, T. (2013). Individual differences in mental rotation. *Cognitive Processing (Special Issue on Spatial Learning and Reasoning), 14*, 153–162.

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Sciences, 3*(11), 419–429.

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think.* Cambridge, MA: Harvard University Press.

Goldin-Meadow, S., & Alibali, M. W. (2013). Gesture's role in speaking, learning, and creating language. *Annual Review of Psychology, 64*, 257–283.

Goldin-Meadow, S., & Sandhofer, C. M. (1999). Gestures convey substantive information about a child's thoughts to ordinary listeners. *Developmental Science, 2*(1), 67–74.

Goldin-Meadow, S., & Singer, M. A. (2003). From children's hands to adults' ears: Gesture's role in the learning process. *Developmental Psychology, 39*(3), 509–520.

González, G., & Herbst, P. (2013). An oral proof in a geometry class: How linguistic tools can help map the content of a proof. *Cognition and Instruction, 31*(3), 271–313.

Harel, G., & Sowder, L. (1998). Students' proof schemes: Results from exploratory studies. *Issues in Mathematics Education, 7*, 234–283.

Harel, G., & Sowder, L. (2007). Toward comprehensive perspectives on the learning and teaching of proof. In F. K. Lester (Ed.). *Second handbook of research on mathematics teaching and learning* (pp. 805–842). Charlotte, NC: Information Age Publishing.

Harnad, S. (1990). The symbol grounding problem. *Physica D, 42*, 335–346.

Havas, D. A., Glenberg, A. M., Gutowski, K. A., Lucarelli, M. J., & Davidson, R. J. (2010). Cosmetic use of botulinum toxin-A affects processing of emotional language. *Psychological Science, 21*(7), 895–900.

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

Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological Bulletin, 137*(2), 297–315.

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review, 15*(3), 495–514.

Hostetter, A. B., & Alibali, M. W. (2018). Gesture as simulated action: Revisiting the framework. *Psychonomic Bulletin & Review.* https://doi.org/10.3758/s13423-018-1548-0 (in press).

Ickes, W., & Cheng, W. (2011). How do thoughts differ from feelings? Putting the differences into words. *Language and Cognitive Processes, 26*(1), 1–23.

Jeannerod, M. (2001). Neural simulation of action: A unifying mechanism for motor cognition. *NeuroImage, 14*, S103–S109.

Kaschak, M. P., Jones, J., Carranza, J., & Fox, M. R. (2014). Embodiment and language comprehension. In L. Shapiro (Ed.). *The Routledge handbook of embodied cognition* (pp. 118–126). New York: Routledge.

Kelly, S. D., Iverson, J., Terranova, J., Niego, J., Hopkins, M., & Goldsmith, L. (2002). Putting language back in the body: Speech and gesture on three timeframes. *Developmental Neuropsychology, 22*, 323–349.

Kendon, A. (2004). *Gesture: Visible action as utterance.* Cambridge: Cambridge University Press.

Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.). *Language and gesture* (pp. 162–185). New York: Cambridge University Press.

Kita, S., & Ozyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence from an interface representation of spatial thinking and speaking. *Journal of Memory and Language, 48*, 16–32.

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

Knuth, E. J., Choppin, J., & Bieda, K. (2009). Middle school students' production of mathematical justifications. In D. Stylianou, E. Knuth, & M. Blanton (Eds.). *Teaching and learning proof across the grades* (pp. 153–170). Mahwah, NJ: Erlbaum.

Lakoff, G., & Nuñez, R. (2000). *Where mathematics comes from: How the embodied mind brings mathematics into being.* New York: Basic Books.

Louwerse, M. M., & Jeuniaux, P. (2008). Language comprehension is both embodied and symbolic. In M. de Vega, A. Glenberg, & A. C. Graesser (Eds.). *Symbols and embodiment: Debates on meaning and cognition* (pp. 309–326). Oxford, England: Oxford University Press.

Marghetis, T., Edwards, L. D., & Núñez, R. (2014). More than mere handwaving: Gesture and embodiment in expert mathematical proof. In L. D. Edwards, F. Ferrara, & D. Moore-Russo (Eds.). *Emerging perspectives on gesture and embodiment in mathematics* (pp. 227–246). Charlotte, NC: Information Age Publishing.

McCullough, M. E., Root, L. M., & Cohen, A. D. (2006). Writing about the benefits of an interpersonal transgression facilitates forgiveness. *Journal of Counseling and Clinical Psychology, 74*(5), 887–897.

McNamara, D. S., Graesser, A. C., Cai, Z., & Kulikowich, J. M. (2011). *Coh-Metrix easability components: Aligning text difficulty with theories of text comprehension. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.*

McNamara, D. S., Louwerse, M. M., Cai, Z., & Graesser, A. (2013). Coh-Metrix (Version 3.0) [Software]. Available from < http://cohmetrix.com/ > .

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge: Cambridge University Press.

McNamara, D. S., Raine, R., Roscoe, R., Crossley, S., Jackson, G. T., & Dai, J. (2012). The Writing-Pal: Natural language algorithms to support intelligent tutoring on writing strategies. In P. M. McCarthy, & C. Boonthum (Eds.). *Applied natural language processing and content analysis: Identification, investigation, and resolution.* Hershey, PA: IGI Global.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought.* Chicago: The University of Chicago Press.

McNeill, D. (2005). *Gesture and thought.* Chicago: The University of Chicago Press.

Nathan, M. J. (2014). Grounded mathematical reasoning. In L. Shapiro (Ed.). *The Routledge handbook of embodied cognition* (pp. 171–183). New York: Routledge.

Nathan, M. J., Srisurichan, R., Walkington, C., Wolfgram, M., Williams, C., & Alibali, M. W. (2013). Building cohesion across representations: A mechanism for STEM integration. *Journal of Engineering Education, 102*(1), 77–116.

Nathan, M. J., Walkington, C., Boncoddo, R., Pier, E., Williams, C. C., & Alibali, M. W. (2014). Actions speak louder with words: The roles of action and pedagogical language for grounding mathematical proof. *Learning and Instruction, 33*, 182–193. https://doi.org/10.1016/j.learninstruc.2014.07.001.

Nathan, M., & Walkington, C. (2017). Grounded and embodied mathematical cognition: Promoting mathematical insight and proof using action and language. *Cognitive Research: Principles and Implications, 2*(1), 9. https://doi.org/10.1186/s41235-016-0040-5.

Nemirovsky, R., & Ferrara, F. (2009). Mathematical imagination and embodied cognition. *Educational Studies in Mathematics, 70*(2), 159–174.

Newcombe, N. S., & Shipley, T. F. (2015). Thinking about spatial thinking: New typology, new assessments. In J. S. Gero (Ed.). *Studying visual and spatial reasoning for design creativity.* Dordrecht: Springer.

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice-Hall.

Ottmar, E. R., & Landy, D. (2016). Concreteness fading of algebraic instruction: Effects on mathematics learning. *The Journal of the Learning Sciences, 26*(1), 51–78.

Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). Linguistic Inquiry and Word Count (LIWC2015) [Software]. Available from < http://www.LIWC.net/ > .

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54*, 547–577.

Pine, K. J., Bird, H., & Kirk, E. (2007). The effects of prohibiting gesture on children's lexical retrieval ability. *Developmental Science, 10*(6), 747–754.

Proyer, R. T., & Brauer, K. (2018). Exploring adult Playfulness: Examining the accuracy of personality judgments at zero-acquaintance and an LIWC analysis of textual information. *Journal of Research in Personality, 73*, 12–20.

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience, 6*(7), 576.

Raskin, R., & Shaw, R. (1988). Narcissism and the use of personal pronouns. *Journal of Personality, 56*, 393–404.

Robinson, R. L., Navea, R., & Ickes, W. (2013). Predicting final course performance from students' written self-introductions: A LIWC analysis. *Journal of Language and Social Psychology, 32*(4), 469–479.

Rude, S., Gortner, E., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion, 18*(8), 1121–1133.

Rummel, N., & Spada, H. (2005). Learning to collaborate: An instructional approach to promoting collaborative problem solving in computer-mediated settings. *The Journal of the Learning Sciences, 14*(2), 201–241.

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

Schoenfeld, A. H. (2016). Research in Mathematics Education. *Review of Research in Education, 40*(1), 497–528. https://doi.org/10.3102/0091732X16658650.

Schwartz, D. L., & Black, J. B. (1996). Shuttling between depictive models and abstract rules: Induction and fallback. *Cognitive Science, 20*, 457–497.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE, 8*(9), e73791. https://doi.org/10.1371/journal.pone.0073791.

Schwartz, H. A., & Ungar, L. H. (2015). Data-driven content analysis of social media: a systematic overview of automated methods. *The Annals of the American Academy of Political and Social Science, 659*(1), 78–94.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*,

417–457.

Shapiro, L. (2010). *Embodied cognition.* Routledge.

Smith, C. P., King, B., & Hoyte, J. (2014). Learning angles through movement: Critical actions for developing understanding in an embodied activity. *The Journal of Mathematical Behavior, 36*, 95–108.

Steels, L. (2008). The symbol grounding problem is solved, so what's next? In M. de Vega, A. Glenberg, & A. Graesser (Eds.). *Symbols and embodiment: Debates on meaning and cognition.* Oxford, England: Oxford University Press.

Stevens, K. A., Ronan, K., & Davies, G. (2017). Treating conduct disorder: An effectiveness and natural language analysis study of a new family-centred intervention program. *Psychiatry Research, 251*, 287–293.

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

Tull, M. T., Medaglia, E., & Roemer, L. (2005). An investigation of the construct validity of the 20-item Toronto Alexithmia Scale through the use of a verbalization task. *Journal of Psychosomatic Research, 59*, 77–84.

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin, 139*(2), 352–402.

Vasalou, A., Gill, A. J., Mazanderani, F., Papoutsi, C., & Joinson, A. (2011). Privacy dictionary: A new resource for the automated content analysis of privacy. *Journal of the American Society for Information Science and Technology, 62*(11), 2095–2105.

Walkington, C., Boncoddo, R., Williams, C., Nathan, M. J., Alibali, M. W., Simon, E., & Pier, E. L. (2014). Being mathematical relations: Dynamic gestures support mathematical reasoning. In W. Penuel, S. A. Jurow, & K. O'Connor (Eds.). *Learning and becoming in practice: Proceedings of the eleventh international conference of the learning sciences* (pp. 479–486). Boulder, CO: University of Colorado.

Walkington, C., Clinton, V., Ritter, S. N., & Nathan, M. J. (2015). How readability and topic incidence relate to performance on mathematics story problems in computer-

based curricula. *Journal of Educational Psychology, 107*(4), 1051. https://doi.org/10.1037/edu0000036.

Williams, C., Akinsiku, O., Walkington, C., Cooper, J., Ellis, A., Kalish, C., & Knuth, E. (2011). Understanding students' similarity and typicality judgments in and out of mathematics. In *Proceedings of the 32nd annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 1180–1189). Reno, NV.

Williams, C., & D'Mello, S. (2010). Predicting student knowledge level from domain-independent function and content words. *Intelligent tutoring systems* (pp. 62–71). Heidelberg: Springer, Berlin.

Williams-Pierce, C. (2016). Provoking mathematical play through hidden deep structures. In C. K. Looi, J. L. Polman, U. Cress, & P. Reimann (Vol. Eds.), *Transforming learning, empowering learners: The international conference of the learning sciences: Vol. 2,* (pp. 1241–1242). Singapore: National Institute of Education, Nanyang Technological University.

Williams-Pierce, C., Pier, E. L., Walkington, C., Boncoddo, R., Clinton, V., Alibali, M. W., & Nathan, M. J. (2017). What we say and how we do: Action, gesture, and language in proving. *Journal for Research in Mathematics Education, 48*(3), 248–260.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin and Review, 9*, 625–636.

Williams, C. C., Walkington, C., Boncoddo, R., Srisurichan, R., Pier, E., Nathan, M., & Alibali, M. (2012). Invisible proof: The role of gestures and action in proof. In L. R. Van Zoest, J.-J. Lo, & J. L. Kratky (Eds.). *Proceedings of the 34th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp. 182–189). Kalamazoo, MI: Western Michigan University.

Woods, D., & Fassnacht, C. (2012). Transana (Version 2.52) [Software]. Available from < http://transana.org/ > .

Xu, R. (2003). Measuring explained variation in linear mixed effects models. *Statistics in Medicine, 22,* 3527–3541.