Learning and Instruction 64 (2019) 101225

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc

Does restricting hand gestures impair mathematical reasoning?

Candace Walkington[a,*], Dawn Woods[a], Mitchell J. Nathan[b], Geoffrey Chelule[a], Min Wang[a]

[a] Department of Teaching and Learning, Southern Methodist University, United States
[b] Department of Educational Psychology, University of Wisconsin, Madison, United States

ARTICLE INFO

Keywords:
Gesture
Proof
Embodied cognition
Gesture inhibition
Geometry

ABSTRACT

Gestures are associated with powerful forms of understanding; however, their causative role in mathematics reasoning is less clear. We inhibit college students' gestures by restraining their hands, and examine the impact on language, recall, intuition, and mathematical justifications of geometric conjectures. We test four mutually exclusive hypotheses: (1) gestures are facilitative, through cognitive off-loading, verbal support, or transduction, (2) gestures are not facilitative, but being inhibited from gesturing increases cognitive demands, (3) gestures are a byproduct of reasoning processes that would take place with or without the gestures' overt presence, and (4) gestures can cause learners to focus on concrete, salient representations, inhibiting abstraction. We find support for the third hypothesis, concluding that learners making or being inhibited from making gestures does not seem to impact their mathematical problem-solving, cognitive, or language processes. This suggests that being unable to overtly perform personally-generated gestures is not a hindrance to learners in this context; however this would not necessarily hold for directed or structured gestures.

1. Introduction

Embodied views on cognition posit that all mental processes are rooted in perceptual and motor systems (Wilson, 2002) and that mental representations of objects are experiential and multimodal in nature (Barsalou, 2008). Embodied approaches to teaching mathematics have become a particularly important area of study, challenging a tradition where mathematics is seen as disconnected from the body, action, and perception (Lakoff & Núñez, 2000). Geometry is an important area for embodiment investigations because of its spatial, dynamic relations and the complex interplay between language, symbols, and action (Nathan & Walkington, 2017).

One way in which mathematical reasoning is embodied is through *gesture*. We define gesture as personally-generated movements of the body that people use during reasoning about or communication of mathematical ideas. This follows McNeill (1992), who defines gestures as "all visible movements by the speaker" (p. 78) that do not involve object manipulation or actions like stroking one's hair. Under this definition, tracing a circle in the air while thinking about a geometric problem involving a circle would be a gesture, as would tilting the head to indicate the movement of an object translated on a Cartesian plane. Here we consider only gestures that are *mathematical* in nature – gestures that relate to mathematical reasoning, rather than gestures given for emphasis or to show other social cues (like nodding).

Learners' tendencies to produce mathematical gestures has been shown to predict learning and performance in mathematics (Cook & Goldin-Meadow, 2006), and students who gesture more and gesture in particular ways communicate more accurate geometry proofs (Nathan et al., 2014; Nathan & Walkington, 2017; Pier et al., 2019). *Dynamic gestures* (Garcia & Infante, 2012; Göksun, Goldin-Meadow, Newcombe, & Shipley, 2013) – gestures where the learner is depicting a motion-based transformation of a mathematical object through multiple states – are strongly associated with proof performance (Pier et al., 2019). An example of a dynamic gesture is formulating a triangle with thumbs and forefingers, and then having the triangle grow and shrink to show mathematical similarity. However, it is unclear whether such gestures are simply a byproduct of valid mathematical reasoning, or a causative factor. In other words, does formulating gestures provide conceptual support that allows students to be more successful, or do students who tend to have stronger mathematical knowledge also tend to gesture more?

One way to experimentally manipulate gesture is through gesture inhibition - physically inhibiting learners from being able to gesture and examining how this impacts reasoning. In the present study, we examine gesture inhibition for university students proving geometric conjectures. We examine how inhibition impacts mathematical

* Corresponding author. 6401 Airline Blvd., Suite 301, Dallas, TX, 75205, United States.
E-mail addresses: cwalkington@smu.edu (C. Walkington), dwoods@smu.edu (D. Woods), mnathan@wisc.edu (M.J. Nathan), nchelule@smu.edu (G. Chelule), minwang@smu.edu (M. Wang).

https://doi.org/10.1016/j.learninstruc.2019.101225
Received 1 June 2018; Received in revised form 17 June 2019; Accepted 1 July 2019
Available online 11 July 2019
0959-4752/

reasoning and speech patterns, and how this effect is moderated by student-level characteristics.

2. Theoretical framework

2.1. Gesture as Simulated Action

Hostetter and Alibali (2008) proposed the *Gesture as Simulated Action* (GSA) framework, where gestures come about as a result of perceptual and motor simulations which arise from mental imagery and language processing. Gestures arise when pre-motor activation is activated beyond a speaker's current *gesture threshold* - the level of motor activation needed for a simulation to be expressed in overt action. This threshold can vary depending on factors such as the current task demands (e.g., strength of motor activation when processing spatial imagery, task difficulty), individual differences (e.g., level of spatial skills), and situational considerations (e.g., social contexts). Hostetter and Alibali (2007) hypothesized that people with low verbal skills but high spatial visualization skills would gesture most, as their mental images may not be well-connected to verbal forms they can orally communicate. Gestures are theorized to assist with "packaging" ideas for speech production (Alibali, Kita, & Young, 2000; Alibali, Yeo, Hostetter, & Kita, 2017); therefore, gesture production may be highest when speakers who have difficulty with verbal skills are presented with an organizationally demanding task (Alibali et al., 2000). A related idea is that gestures facilitate lexical retrieval – they allow learners to produce more fluent speech by facilitating better retrieval of words from memory (Krauss, 1998).

GSA suggests that inhibiting gesture may increase cognitive load, which may be particularly detrimental when learners are confronting a challenging task. *Cognitive load* is cognitive processing demands experienced by learners due to the relationship between a task's difficulty and a learner's cognitive system (Van Merriënboer & Sweller, 2005). These demands draw upon *working memory* – the learner's cognitive capacity for in-the-moment processing, holding, and manipulation of information. Working memory demands can be reduced by utilizing external resources in the environment (e.g., writing down a phone number) through *cognitive offloading* (Risko & Gilbert, 2016). Gestures may relieve cognitive load by acting as an off-loading mechanism, allowing learners to bring new memory stores, such as spatial working memory, to bear. Thus, restricting the availability of gestures may prevent this beneficial off-loading from happening. Alternately, gesture inhibition itself may increase extraneous cognitive load, as stopping oneself from gesturing could be an effortful activity that utilizes working memory. This may be particularly detrimental to people with a low gesture threshold. GSA remains neutral on whether gesture inhibition itself utilizes cognitive resources because it is an effortful task, or whether gestures function to relieve cognitive demand (Hostetter & Alibali, 2008).

2.2. Cognition-action transduction

Nathan (2017) proposed to expand the GSA framework, citing recent research suggesting a *bidirectional* relation between action and cognition; this new theory is called *cognition-action transduction*. In addition to the hypothesis that mental simulations give rise to gestures, as proposed by the GSA framework of Hostetter and Alibali (2008), Nathan (2017) describes emerging evidence that the act of gesturing can itself activate mental simulations. In accordance with this view, superior problem-solving performance has been demonstrated when students follow directions to perform specific actions that correspond to cognitive operations that contribute to effective problem-solving strategies (e.g., Ginns, Hu, Byrne, & Bobis, 2016; Goldin-Meadow, Cook, & Mitchell, 2009; Hu, Ginns, & Bobis, 2015; Nathan et al., 2014; Novack, Congdon, Hemani-Lopez, & Goldin-Meadow, 2014). Cognition-action transduction allows that gesture inhibition could be detrimental, as people who are inhibited are not able to use gestures as a resource to understand new ideas. However, studies where learners are instructed to formulate particular highly-effective gestures are different than studies that allow learners to engage in their own personally-generated gesturing. Being instructed might be beneficial for low-knowledge learners who lack the resources to create their own effective gestures; these low-knowledge learners may actually generate misleading gestures illustrating incorrect relationships. However, Nathan et al. (2014) found that even purposefully-chosen directed gestures, when not properly understood, could lead learners down an incorrect solution path if they did not understand the relationships they were intended to be physically representing. Thus, it may only be when the person-environment system offers appropriate feedback that cognition-action transduction would predict that the outward processes would be correct and lead to a desired change in cognitive state.

An alternate view is that gestures or directed actions may focus learners' attention on the specific concrete, spatial qualities of the objects they are physically representing or pointing to, therefore learners may not engage in generalization or abstraction – as described by Walkington, Nathan, Wolfgram, Alibali, and Srisurichan (2014), they develop *modal-specific epistemological commitments*. These are situations where learners are so focused on immediately present, salient representations of concepts (like a concrete gesture) that they struggle to transfer this knowledge to other representations (Sloutsky, Kaminski, & Heckler, 2005), like a generalized mathematical proof. From a cognitive-action transduction standpoint, this could also lead them to give an incorrect or incomplete problem solution.

One way to study whether and how gestures affect cognition is to manipulate gestures through gesture inhibition.

2.3. Prior gesture inhibition studies

Having learners tap with one hand in particular patterns that periodically change to hinder automaticity (spatial tapping) or that involve tapping repeatedly in one place (simple tapping) while solving problems has been studied as a method of gesture inhibition. Results from this line of research (Hegarty, Mayer, Kriz, & Keehner, 2005; Nathan & Martinez, 2015) with participants solving problems about mechanical and biological systems show that spatial tapping does indeed inhibit problem-solving performance ($\eta^2 = 0.2$), while simple tapping does not. Hegarty et al. (2005) also found no effect for gesture inhibition through hand restriction. Together, these findings suggest that it is not the production of the gestures themselves that impact performance (i.e., not simple tapping or gesture inhibition). Instead, it is the demands of redirecting the processes involved in monitoring and executing particular motor sequences through the spatial tapping condition that selectively disrupts model-based reasoning, hampering performance on inference-making tasks. However, other researchers have examined the effects of hand restriction and found different results.

Several studies have looked at the impact of hand restriction on speech. Graham and Heywood (1975) found that gesture inhibition was associated with students using more spatial relation words, fewer demonstrative words, and more time spent pausing. Hostetter, Alibali, and Kita (2007) found that participants free to gesture used more semantically rich verbs ($d = 0.80$) and were less likely to begin sentences with "and" ($d = 1.24$). A similar study (Rauscher, Krauss, & Chen, 1996) found that gesture inhibition was associated with speaking more slowly and having more dysfluencies when discussing spatial content but speaking more quickly when discussing non-spatial content ($ds = 0.30$–0.58). However, in a study of undergraduates giving others instructions on how to perform simple tasks, Hoetjes, Krahmer, and Swerts (2014) found no differences between hand-restricted and unrestricted participants on any speech category they measured – including speech duration or rate, number of words, pauses, or acoustic measures.

Other studies have examined the effect of hand restriction on recall.

Frick-Horbury and Guttentag (1998) found gesture inhibition led to lower retrieval (*ds* = 0.64–1.04) and recall (*ds* = 1.32–1.98) of words and that this effect did not vary based on SAT scores. Beattie and Coughlan (1999) found that gesture-inhibited students were actually *more* likely to recall words, although this difference did not reach significance ($d = 0.29$). Those who were inhibited were significantly less likely to report a "tip of the tongue" state (i.e., where they believed they knew the word but could not retrieve it; $d = 0.60$) but were also less likely to be able to resolve this state when it happened ($d = 0.72$).

Two recall studies have attempted to clarify the mechanisms through which gesturing impacts recall. Goldin-Meadow, Nusbaum, Kelly, and Wagner (2001) asked participants to solve math problems while keeping words or letters in memory, and found recall was higher when gesturing ($d = 0.35$). They also compared instances where the participant was free to gesture but chose not to, versus being free to gesture and choosing to gesture, and found similar advantages of gesturing for recall. They concluded that gesture allows for cognitive offloading. In a similar study (Wagner, Nusbaum, & Goldin-Meadow, 2004) participants were asked to hold either a string of letters or a visuospatial configuration of dots in memory. No differences were found in speech patterns when participants did versus did not gesture, but not gesturing was associated with weaker recall. In addition, not gesturing when uninhibited was associated with negative outcomes that were similar to being inhibited from gesturing. The researchers conclude that gesturing reduces the load on both visuospatial and verbal working memory. However, they found that gesturing was only beneficial when it conveyed information that was also in speech, supporting the idea that gesture is beneficial because it helps learners organize information into the propositional form needed for speech. These studies are both of limited relevance because they do not contrast participants who were inhibited versus non-inhibited.

Alibali and Kita (2010) examined student explanations of Piagetian conservation tasks and found gesture inhibition caused children to express less perceptually present information (e.g., the glass is tall versus short) but more non-present information ($\eta^2 = 0.16$). When inhibited, participants were more likely ($\eta^2 = 0.28$) to call upon past events (e.g., "they were the same"), make hypotheses about things that might happen (e.g., "if the glass was fatter …"), and talk about transformations (e.g., "you moved it over"). When considering mathematical justifications, focusing on hypothetical states, past states, and transformations (rather than immediately-present, salient characteristics of mathematical objects) could be desirable (Harel & Sowder, 2005). This study points to the need to examine gesture inhibition for tasks like mathematical proof.

2.4. Research purpose

Prior research has not examined how gesture inhibition impacts mathematical reasoning generally, or geometric reasoning specifically, both of which have important visual, spatial, and motoric properties, in addition to powerful uses of language as a grounding mechanism (Lakoff & Núñez, 2000; Nathan, 2014). Mathematical reasoning is more complex than the simple recall or descriptive tasks examined in previous studies and has directly-actionable pedagogical implications. Geometric reasoning is an especially important area for the examination of gesture inhibition, given its spatial nature, use of transformational reasoning, and the prevalence of gestures (Nathan & Walkington, 2017; Walkington, Chelule, Woods, & Nathan, in press). Prior research has also not often examined whether effects of gesture inhibition vary based on learner characteristics, as this may influence students' ability to formulate their own effective gestures.

Prior gesture inhibition studies have also often utilized small sample sizes and between-subjects comparisons. This makes it difficult to draw firm conclusions or conduct analyses of moderation effects, a primary focus here. Many prior studies have defined gestures as being done with the hands (e.g., Frick-Horbury & Guttentag, 1998; Goldin-Meadow et al., 2001; Rauscher et al., 1996; Wagner et al., 2004) and have not considered the various ways learners can still gesture with their hands restrained. Indeed, one study suggests learners with their limbs and hands restricted become more likely to gesture with other available body zones (e.g., eyebrows; Rimé, Schiaratur, Hupet, & Ghysselinckx, 1984). Finally, advances in computerized text analysis of speech patterns (McNamara, Louwerse, Cai, & Graesser, 2013) and the recent identification of dynamic gestures (e.g., Garcia & Infante, 2012) open up new possibilities for how gesture inhibition can be studied. The current study addresses each of these gaps.

2.5. Research questions and hypotheses

Our research questions (RQs) are:

1) How do (a) speech patterns, and (b) gesture and dynamic gesture patterns, vary when participants are inhibited versus not inhibited from gesturing?
2) Does gesture inhibition impair recall, intuition, insight, or proof performance on geometric tasks, and does inhibition interact with participant characteristics?
3) How is the presence of gesture and dynamic gesture associated with recall, intuition, insight, and proof for only those trials when participants were free to gesture, and how do gesture effects interact with participant characteristics?

We pose four mutually exclusive hypotheses, with each hypothesis leading to different specific predictions for each research question. The first is the ***facilitation hypothesis***, which posits that personally-generated gestures can causatively help learners. There are several conceptually distinct accounts in the literature as to why this might occur. First, gestures might reduce cognitive load through cognitive offloading. Second, gestures may allow learners to communicate their ideas better verbally, facilitating packaging of ideas into speech and/or allowing for lexical retrieval. Third, gestures might serve a transductive purpose by activating mental simulations, giving learners new, actionable ideas.

Second is the ***interference hypothesis***, which posits that gestures do not have facilitative properties, but that preventing learners from gesturing in and of itself is an effortful activity that increases cognitive load, perhaps due to the novelty or discomfort of being inhibited from gesturing. Goldin-Meadow et al. (2001) describe this situation as "the observed effect is not due to the beneficial effects of gesture, but to the deleterious effects of the constraining instructions. Asking speakers not to gesture is, in effect, asking them to do yet another task …" (p. 519).

Third is the ***byproduct hypothesis***, which posits that gestures do not have facilitative properties, and that being inhibited from gesturing is not a cognitively effortful activity. This hypothesis would view gestures as merely an outgrowth of valid mathematical reasoning, rather than a causative factor. This hypothesis posits that gestures tend to co-occur with valid mathematical reasoning, without taking a role in causing that reasoning to happen. Fourth is the ***concreteness hypothesis***, which posts that gestures may cause learners to focus on currently-present, spatial, salient forms of mathematical concepts, and disrupt possibilities for engaging in mathematical abstraction or hypothetical or deductive reasoning.

For **RQ1a** relating to speech patterns, the facilitation hypothesis would suggest that gesture inhibition would change speech patterns (H1a-facilitation), as facilitation in the form of offloading or improved retrieval or communication may allow for learners to give longer proofs that use more logical statements, describe more operations on objects, or that show more generalized or abstract thinking (Harel & Sowder, 2005; see Pier et al., 2019). In addition, a transduction effect may show proofs with more action and body-related words, as well as increased verb and cohesion measures related to situation models. Finally, when inhibited we may see speech changes consistent with prior studies

reviewed above, such as fewer demonstrative words and semantically rich verbs.

The interference hypothesis would also posit that gesture inhibition changes speech patterns (H1a-interference). When cognitive load is increased, learners may use speech patterns that involve shorter proofs, less description of mathematical operations, less abstract or generalized language, and more dysfluent words or informal speech. The byproduct hypothesis would posit no relationship between gesture inhibition and speech patterns (H1a-byproduct). The concreteness hypothesis would posit that gesture inhibition causes participants to use language patterns that involve less concrete and spatial words, and more abstract terms and deductive language processes (H1a-concreteness).

For **RQ1b** regarding the impact of gesture inhibition on gesture, all four hypotheses would assume that gesture inhibition reduces gesture usage (H1b).

For **RQ2** relating to problem-solving performance, the facilitation hypothesis and the interference hypothesis would posit that gesture inhibition dampens problem-solving performance (H2-facilitation and H2-interference). The byproduct hypothesis would predict no relationship between gesture inhibition and performance (H2-byproduct), while the concreteness hypothesis would posit that gesture inhibition improves performance (H2-concretenes).

For **RQ3** relating to participant performance while free to gesture, the facilitation hypothesis and the byproduct hypothesis would predict that when free to gesture, gestures are associated with improved performance (H3-facilitation and H3-byproduct). The interference hypothesis would predict no association between gestures and performance when free to gesture (H3-interference), while the concreteness hypothesis would predict gesture use to be associated with *dampened* performance when uninhibited (H3-concreteness). The hypotheses and the research questions are summarized in Table 1.

Research Questions 2–3 also allow that these hypotheses may not be uniform across participants. For example, offloading might only be beneficial when the learner has a less strong background in mathematics and needs to relieve working memory demands for in-the-moment problem solving, while the hypothesized verbal support from gestures may only be beneficial when the learner has low fluency skills and needs gestures to help retrieve words and package ideas into speech. Transduction might only occur when the learner has strong enough spatial skills to produce useful gestures to give them new ideas. Interference might only occur when the learner has a weak mathematical background and the cognitive system is already overloaded. Gestures may only be a byproduct of valid mathematical reasoning when learners have high enough spatial skills to formulate mathematical gestures to accompany speech. Participants with low spatial skills may be especially likely to fall prey to difficulty generalizing from concrete representations. By examining a variety of moderators, we can consider whether differential effects might be occurring.

Finally, research question 3 also considers differences related to dynamic gestures as a special class of gesture that is particularly relevant to mathematical reasoning. We hypothesize that the presence of dynamic gestures will have a stronger association with insight and proof during uninhibited trials, compared to simply any gesturing being present. This is because dynamic gestures show transformations, which are central to the processes involved in understanding and generalizing mathematical relationships, which are key to having insights and formulating proofs of geometric conjectures.

Table 1
Summary of research questions and hypotheses.

	Hypotheses			
	Facilitation	Interference	Byproduct	Concreteness
Inhibited versus not inhibited				
Speech (RQ1a)	(H1a-facilitation) Inhibition may change language patterns, as retrieval and/or packaging and/or communication is facilitated, or transduction is instigated	(H1a-interference) Inhibition may change language patterns, as learners experience more cognitive load (e.g., more difficult word retrieval)	(H1a-byproduct) No differences	(H1a-concreteness) Inhibition may be associated with fewer concrete/spatial words, more abstract words, more hypothetical statements
Gesture (RQ1b)	(H1b) Less gesture occurs when learners are inhibited from gesturing			
Performance (Intuition, Insight, Proof; RQ2)	(H2-facilitation) Inhibition is harmful	(H2-interference) Inhibition is harmful	(H2-byproduct) No differences	(H2-concreteness) Inhibition is beneficial
Choose to gesture versus do not choose to gesture (Uninhibited trials only)				
Performance (Intuition, Insight, Proof; RQ3)	(H3-facilitation) Gesture use associated with higher performance	(H3-interference) No differences	(H3-byproduct) Gesture use associated with higher performance	(H3-concreteness) Gesture use associated with lower performance

3. Methods

3.1. Participants

Undergraduate and graduate students (n = 108; 48 male and 60 female) from a private university were recruited to participate in a laboratory study lasting 30–45 min. Math and statistics majors and graduate students were specifically targeted through department

emails, signs, and class visits. Of the 108 participants, 34 were math or statistics majors, 7 were math or statistics graduate students, 5 were engineering majors, and 5 were science majors. The remaining 57 participants were undergraduate non-STEM majors or undeclared majors. The mean age was 20.41 years (SD = 2.18). Fifty-five participants had taken a math class above calculus 1, 26 had taken up to calculus 1, and 27 had taken below calculus 1. Fifty-eight participants identified as Caucasian, 17 as Asian, 12 as African-American, 12 as Hispanic, and 9 as other races or biracial; 23 participants reported being non-native English speakers. One participant (a female math major) was omitted during the coding phase because she was not able to speak English well enough to respond to the prompts, for a final sample of 107.

3.2. Power analysis

An a priori power analysis was conducted with G*Power 3.1.9.2 (Faul, Erdfelder, Buchner, & Lang, 2009) using $\beta = 0.80$ and $\alpha = 0.05$. Based on previous data (Nathan & Walkington, 2017), correlations among a participant solving repeated geometry proofs were estimated at 0.6. We estimate an effect size of $d = 0.6$ for gesture inhibition on proof performance from Walkington, Boncoddo, et al. (2014) and Walkington, Nathan, et al. (2014), a small pilot study where 15 participants proved a similar set of geometry conjectures while inhibited or not inhibited for all trials. This study was used because it had the closest outcome variable to the present study (mathematical proof performance); however, this effect size was generally consistent with effect sizes in other studies reviewed earlier. We additionally took into account being powered to detect mediation effects of gesture and speech (assuming partial mediation and small to medium-sized paths for alpha and beta, d = 0.26–0.39; Fritz & MacKinnon, 2007), which led to a sample size of 86, taking into account the design effect. This was our minimum, but we accepted all participants who responded to the ads, with the restriction that we wanted to keep the number of non-STEM majors relatively balanced with the number of STEM majors. With 107 participants, a post-hoc sensitivity analysis showed that we should be able to detect effects for gesture inhibition that are as small in size as $d = 0.2$.

3.3. Procedures

Participants engaged in a one-on-one session with an interviewer. They were given four pre-measures (described later) and presented with 8 geometry conjectures (Table 2). The conjectures were ordered via a Latin Square and projected onto a screen one at a time. Participants were asked to read each conjecture out loud and state whether each conjecture was true or false and why it was true or false. For 4 of the 8 conjectures, participants were inhibited from gesturing by putting their hands in oven mitts that were attached to bottles attached to a music stand (Fig. 1). Participants were either inhibited for the first four conjectures or the final four conjectures; inhibition order was counterbalanced. When they completed all 8 conjectures, participants were asked to, while uninhibited, recall as many of the conjectures as possible. All interviewer interaction with the participant was scripted, to ensure uniform treatment of inhibited versus uninhibited trials.

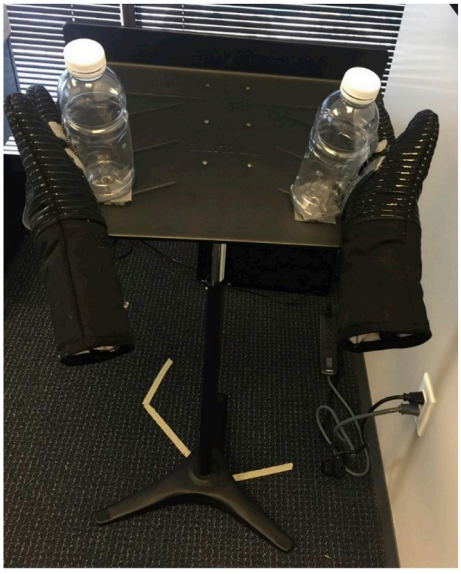

Fig. 1. Gesture inhibition rig.

3.4. Measures

Geometry Pretest. Participants were given a geometry knowledge pretest, developed in a prior study (Nathan & Walkington, 2017; r = 0.56 with performance on conjectures similar to those considered here), composed of twelve statements about triangles, parallelograms, and circles. Although the pretest had a moderate correlation with student performance on the conjectures in the present study (r = 0.41 with proof), it was ultimately not used in the models because of issues with internal consistency. Results for gesture inhibition were the same with or without pretest included, and pretest did not significantly interact with gesture inhibition in any model.

Spatial Skills Test. Participants were given the Paper Folding Test, from The Kit of Factor-Referenced Cognitive Tests (Ekstrom, French, Harman, & Derman, 1976), which measures participant's ability to visualize and manipulate images. Scores are calculated as the number of items correct minus one-quarter of the number incorrect. Reliability for this measure is 0.75 for males and 0.77 for females.

Phonemic Fluency Test. The phonemic fluency measure tests speakers' ability to manage the organizational demands of speaking by

Table 2
Conjectures used in study, with average success rates for proof and insight.

	Conjecture	Verity	Proof Correct	Insight Correct	Intuition Correct	Recall Correct
1	An angle bisector of any angle of a triangle also bisects the opposite side.	False	6.5%	23.4%	31.8%	61.68%
2	Any translation can be expressed as the composition of two reflections.	True	5.6%	23.4%	43.0%	53.27%
3	If one angle of a triangle is larger than a second angle, then the side opposite the first angle is longer than the side opposite the second angle.	True	29.9%	63.6%	88.8%	56.07%
4	The area of a parallelogram is the same as the area of a rectangle with the same length and height.	True	29.0%	75.7%	81.3%	70.09%
5	The segment that joins the midpoints of two sides of any triangle is parallel to the third side.	True	11.2%	46.7%	63.6%	50.47%
6	Any rotation can be expressed as the composition of two reflections.	True	4.7%	24.3%	51.4%	54.21%
7	Given that you know the measure of all three angles of a triangle, there is only one unique triangle that can be formed with these three angle measurements.	False	35.5%	39.3%	52.3%	43.93%
8	If there are three points P, R, and Q in space, and the distance between P and Q equals the sum of the distance between P and R and the distance between R and Q, then points P, R, and Q must all lie along the same line.	True	25.2%	43.0%	56.1%	50.47%

rapidly generating words in a way that they do not typically organize them in their lexicon; here, by naming as many words as they can in 60 s that begin with the letter 's' and then the letter 't.' Score is the number of words generated in 60 s (omitting proper nouns and simple variants). Retest reliability is 0.88 (desRosiers & Kavanagh, 1987). For the present data we examined the correlation between the count of words each participant generated beginning with "t" and the count of words each participant generated beginning with "s." The correlation for these values (including word variants) was 0.73. Approximately 19% of the words participants generated for this test were simple variants of words they had already generated or proper nouns (19.7% for "s" and 18.6% for "t"). The mean (14.8) and standard deviation (4.8) for the phonemic fluency measure here was similar to other studies that have used this measure (Hostetter & Alibali, 2007; Yeudall, Fromm, Reddon, & Stefanyk, 1986).

Demographic Questionnaire. A paper questionnaire asked participants to identify their class/year, race/ethnicity, Math ACT/SAT scores, college major(s), age, gender, native language, and prior and current math courses taken. ACT/SAT scores were not used as 42 of the 108 responses were missing/ambiguous.

3.5. Coding

Participants were video-recorded and their speech was transcribed in the Transana video analysis software (Woods & Fassnacht, 2012). Individual clips were made of each participant proving each conjecture for a total of 856 clips. One STEM major had his hands off-camera while uninhibited, thus his uninhibited data is omitted from gesture analyses.

Coding Proof. Participants' oral proofs for each conjecture were scored 0/1 in terms of correct or incorrect using a codebook (see Appendix A), which was developed from the criteria for valid mathematical proofs given by Harel and Sowder (2005). Cohen's kappa reliability of 0.81 was achieved for 100 randomly-selected double-coded clips.

Coding of Insight. Given the relatively low rate of valid proofs, a complementary measure (insight) was included to assess whether participants demonstrated understanding of some of the key ideas, without necessarily getting all the way to a full deductive argument. Insight was defined according to Zhang, Lei, and Li (2016) as conscious retrieval of activated mathematical properties and examples that are both validly applied and relevant to the conjecture at hand. See Appendix A for how this was operationalized. Cohen's kappa reliability of 0.80 was achieved for 100 randomly-selected double-coded clips.

Coding of Intuition. Participants' justifications were also coded for whether they correctly concluded the conjecture was true or false. If the participant changed their answer, only the final answer was considered for the coding. One hundred clips were randomly selected from the corpus for independent double-coding; a Cohen's kappa reliability of 0.97 was achieved.

Coding of Recall. Participants were asked at the end of the session to repeat as many of the conjectures as they could remember. If the participant recalled a conjecture, they received a code of "1", otherwise they received a code of "0." Cohen's kappa of 0.96 was achieved for 100 randomly-selected double-coded clips.

Coding of Gesture. Each clip was coded for whether the participant made (1) any gesture at all, which represented, pointed to, or indicated movement or transformation of, mathematical objects (e.g., sweeping the head left and upwards when indicating the upper left corner of a quadrilateral) or (2) any dynamic gesture, where participants showed a movement-based transformation of an object through multiple states (e.g., showing a translation of a mathematical figure by tilting the head or moving the hands). These categories were 0/1 variables, and gestures could fall into multiple categories. Reliability for inhibited versus uninhibited trials was calculated separately, due to the increased challenge of coding inhibited trials and differing gesture rates. Kappas of 0.90 (any gesture) and 0.87 (dynamic gesture) were achieved between two coders for the uninhibited trials on a random subset of 50 clips, while kappas of 0.81 (any gesture) and 0.81 (dynamic gesture) were achieved between two coders for the inhibited trials on a random subset of 50 clips.

Prior gesture inhibition studies have largely defined gestures in terms of hand movements, thus this coding represents a methodological expansion. Participants inhibited from gesturing would use their head to point to imagined figures in front of them or to show an object's movement. In addition, although not used in statistical analyses because of their rarity, we also looked at how often participants attempted hand gestures when gestures were inhibited. Occasionally the camera feed would capture slight but visible movements indicating participants were moving their fingers inside the oven mitts, that seemed to be related to their mathematical reasoning. We found that while hand gestures occurred in 62.74% of all trials while uninhibited, they only occurred in 4.45% of inhibited trials.

Coding of Speech. In order to examine the differences in speech patterns proposed in our hypotheses, the transcript of each participant's speech was entered into two text analysis software packages, Coh-metrix and LIWC. Coh-metrix (McNamara et al., 2013) codes texts based on 108 categories, which range from the number of words per sentence to the average concreteness and age of acquisition of words. Prior research has found that Coh-metrix can distinguish important elements of mathematical arguments like using connective words, actions on mathematical objects, and deductive statements, as well as progressing through a logical structure through sentence overlap (Nathan et al., 2018; Pier et al., 2019). LIWC (Pennebaker, Booth, Boyd, & Francis, 2015) is a dictionary-based computerized text analysis tool that counts the number of words occurring in the text in 70 categories. Although not all LIWC categories are relevant, many categories (like cognitive process words) have been found to be important in prior investigation of oral proofs (Pier et al., 2019). All language variables tested are listed in Appendix B.

As indicated by our hypotheses (Table 1), we only investigated speech differences for inhibited versus uninhibited trials. How speech categories relating to mathematical proof practices change when learners choose to gesture versus not gesture has been examined extensively in other work (see Pier et al., 2019). By comparing the coded speech categories used during inhibited trials to uninhibited trials, we can see if there were significant differences in the kinds of words, phrases, and patterns of speech participants used that reflected the differences we proposed in our hypotheses.

3.6. Analysis

Our hypotheses (Table 1) relate to three comparisons – (1) participants' performance on inhibited versus uninhibited trials, (2) participants' speech patterns in inhibited versus uninhibited trials, and (3) participants' performance when they chose to gesture versus not gesture (when uninhibited). For the first comparison, we modelled performance as an outcome with inhibited/uninhibited as a predictor. For the second comparison, we modelled our quantitatively coded speech categories from Coh-Metrix and LIWC as an outcome, with inhibited/uninhibited as a predictor. And for the third comparison, we modelled performance as an outcome (including only the uninhibited trials), with gesture/no gesture as a predictor.

For Research Question 1a, as an initial screening step, we first removed from consideration language categories in Coh-Metrix and LIWC that (1) were intended for text that and clear sentence delineations (rather than natural speech; e.g., number of sentences), (2) related to use of punctuation (e.g., incidence of semicolons), and (3) that were "0" (i.e., not present) for 75% or more of the data points. This left us with 53 different categories from LIWC and 95 different categories from Coh-Metrix (see Appendix B). All variables were continuous with the exception of word count. We then fit mixed effects linear models (with student and conjecture as random effects) predicting each language

category with inhibition condition as a predictor. We used a cluster bootstrap to estimate the standard error of the regression coefficients. The cluster bootstrap was used because it can deal with issues of non-normality and heteroscedasticity, and works for data that have a lot of "0" values, as is typical for readability variables. Bootstrap samples were drawn by sampling individuals rather than observations, thus taking into account the nesting of observations within participant. We implemented the cluster bootstrap procedure by (i) generating 1000 bootstrap samples using the ClusterBootstrap library in R, (ii) applying the linear mixed effects models to each bootstrap sample, and (iii) computing the regression coefficients' standard errors from the bootstrap distribution. We applied the False Discovery Rate p-value correction to the significance tests for the regression coefficients (Benjamini & Hochberg, 1995).

For Research Question 1b, mixed effects logistic regression models were fit using the *glmer()* command (Bates, Maechler, Bolker, & Walker, 2014) in the R software package. Models predicted any gesture (coded 0/1) and dynamic gesture (coded 0/1), with participant, conjecture, and which of the 8 Latin square orders they received as random effects. Fixed effects included whether the participant was inhibited for the conjecture, demographic variables (gender, language), and expertise variables (geometry pre-test score, spatial score, phonemic fluency score, STEM/non-STEM major, highest math course taken).

For Research Question 2, similar models were fit predicting correct proof (coded 0/1) and insight (coded 0/1). For the recall model only, an additional fixed effect was added to identify in what order during the session (1–8, a factor variable) participants had received the conjecture. We first fit main effects models examining the impact of gesture inhibition, and then examined two-way interactions between gesture inhibition and other fixed effects. Interactions were only retained if they significantly reduced deviance using the *anova()* command in R. The regression tables present raw coefficients that can be exponentiated to get odds ratios. Standardized mean difference-type effect sizes were calculated using the method in Chinn (2000).

For Research Question 3, the analysis process given above for Research Question 2 was repeated for the subset where participants were uninhibited (n = 428 trials). Additionally, dynamic gesture and any gesture were added as predictors.

4. Results

Table 3 gives descriptive statistics for our measures. As can be seen from the table, our four performance measures (intuition, recall, insight, and proof) had different average accuracy levels: 58.5%, 55.0%, 42.4%, and 18.5%, respectively. As we coded the validity of participants' mathematical reasoning at a variety of different levels of task difficulty, we offset concerns of either having ceiling or floor effects. We also conducted supplementary analyses where we only examined the easier or more difficult conjectures, and results were the same.

Table 3
Average rates of correctness and gesture incidence.

Measure	All Participants (N = 107)	
Geometry Pre-Test Mean (*SD*)	80.61% (12.26%)	
Spatial Skills Mean (*SD*)	12.20 (4.70)	
Phonemic Fluency Mean (*SD*)	14.87 (3.84)	
	Inhibited (428 trials)	**Not** (428 trials)
% Trials Correct Proof	18.22%	18.69%
% Trials Correct Insight	43.46%	41.36%
% Trials Correct Intuition	59.11%	57.94%
% Trials Correct Recall	55.61%	54.44%
% Trials Any Gesture	29.67%	69.81%
% Trials Dynamic Gesture	10.75%	32.78%

4.1. RQ1: association between gesture inhibition and speech and gesture

To examine how speech patterns varied when participants were inhibited versus not inhibited, we used a cluster bootstrap with mixed effects regression models to assess whether there were differences between uninhibited and inhibited trials according to the 148 different language measures. None were significant (consistent with H1a-by-product). See Appendix C for information on the regression coefficients and significance tests. Appendix B also includes Cohen's d values that show the effect size of gesture inhibition for each speech category – d values were less than or equal to 0.3, with only three categories (past tense, positive emotion, and affect) having a d value greater than 0.2.

Participants inhibited from gesturing were less likely to make any gesture (Odds = 0.09, $d = -1.32$, $p < 0.001$; consistent with H1b) and dynamic gestures (Odds = 0.14, $d = 1.07$, $p < 0.001$). When both including and removing the gesture inhibition variable, there were no instances where another fixed effect predicted gesture. Examining factors predicting gestures during only the uninhibited trials yielded only the significant effect that students who had taken calculus 2 or a higher math course were significantly more likely to make dynamic gestures than those who had taken only calculus 1 (Odds = 3.36, $d = 0.67$, $p = 0.034$) or those who had not taken calculus (Odds = 3.55, $d = 0.70$, $p = 0.036$).

4.2. RQ 2: association between gesture inhibition and outcomes

For the recall model (Model 1 in Table 4), gesture inhibition had no effect ($p = 0.62$, $d = 0.04$). The only fixed effect predictive of recall was course taking, with participants whose highest math class was below calculus 1 less likely to recall a conjecture than participants above calculus 1 (Odds = 0.46, d = -0.43, $p = 0.020$). The order in which participants received conjectures was highly significant but is not shown in the table for brevity. For the intuition model (Model 2), gesture inhibition had no effect ($p = 0.77$, $d = 0.02$). No other fixed effects predicted intuition. Models 1 and 2 were re-fit examining whether gesture inhibition interacted with each of the other variables. No interaction terms were statistically significant.

In the main effects model predicting insight (Model 3 in Table 5), spatial test score was significantly positively associated with correct insights ($d = 0.034$ per 1 point on test, $p = 0.014$). Being male was also associated with a greater likelihood of correct insight (Odds = 1.63, $d = 0.27$, $p = 0.024$), as was taking a math course beyond calculus 1, rather than no calculus (Odds = 0.49, $d = -0.40$, $p = 0.044$) or calculus 1 only (Odds = 0.51, $d = -0.37$, $p = 0.034$). There was no main effect for gesture inhibition ($p = 0.44$, $d = 0.07$).

Table 4
Models predicting recall and intuition.

Fixed Effects	Model 1: Main Effects Recall	Model 2: Main Effects Intuition
	B (SE)^{Sig}	B (SE)^{Sig}
(Intercept)	−0.39 (0.39)	0.54 (0.41)
Gestures Not Inhibited	(ref.)	(ref.)
Gestures Inhibited	0.08 (0.16)	0.05 (0.15)
Female	(ref.)	(ref.)
Male	0.03 (0.21)	0.18 (0.19)
Native English	(ref.)	(ref.)
Non-Native English	0.06 (0.28)	−0.06 (0.25)
Non-STEM major	(ref.)	(ref.)
STEM major	−0.31 (0.30)	−0.02 (0.26)
Phonemic Fluency	0.05 (0.03)	−0.01 (0.03)
Course Beyond Calc I	(ref.)	(ref.)
Course Below Calc 1	−0.78 (0.34)*	−0.31 (0.31)
Course Calc 1	−0.54 (0.31)	−0.39 (0.27)
Spatial Score	0.02 (0.02)	0.02 (0.02)

Note. (ref.) denotes the reference category. * = p < 0.05.

Table 5
Models predicting insight and proof.

	Model 3: Main Effects Insight	Model 4: Main Effects Proof
	B (SE)^{Sig}	B (SE)^{Sig}
(Intercept)	−0.24 (0.45)	−2.02 (0.56)***
Gestures Not Inhibited	(ref.)	(ref.)
Gestures Inhibited	0.12 (0.16)	0.03 (0.21)
Female	(ref.)	(ref.)
Male	0.49 (0.22)*	0.64 (0.28)*
Native English	(ref.)	(ref.)
Non-Native English	−0.57 (0.29)	−0.38 (0.37)
Non-STEM major	(ref.)	(ref.)
STEM major	0.06 (0.30)	0.23 (0.39)
Phonemic Fluency Score	0.01 (0.03)	−0.02 (0.04)
Course Beyond Calc I	(ref.)	(ref.)
Course Below Calc 1	−0.72 (0.36)*	−1.04 (0.48)*
Course Calc 1	−0.67 (0.32)*	−1.08 (0.42)*
Spatial Score	0.06 (0.02)*	0.12 (0.04)**

Note. (ref.) denotes the reference category. * = p < 0.05, ** = p < 0.01, *** = p < 0.001.

In the main effects model predicting proof (Model 4), spatial test score was significantly positively associated with correct proof ($d = 0.07$ per 1 point on test, $p = 0.001$). Being male was also associated with a greater likelihood of correct proof (Odds = 1.89, $d = 0.35$, $p = 0.023$). Having a highest math course above calculus 1 was associated with a higher likelihood of correct proof, compared to only calculus 1, (Odds = 2.96, $d = 0.60$, $p = 0.011$), and below calculus 1 (Odds = 2.81, d = 0.57, p = 0.030). There was no main effect for gesture inhibition ($p = 0.88$, $d = 0.02$). Models 3 and 4 were re-fit examining whether gesture inhibition interacted with each of the other variables. No interaction terms were statistically significant ($ps > .05$).

In sum, gesture inhibition did not predict any of the performance measures, consistent with H2-byproduct.

4.3. RQ3: association between gesture and outcomes when uninhibited

We next considered only trials where participants were free to gesture (Table 6). For the models predicting recall and intuition, neither gesture type (any gesture, dynamic gesture) was associated with outcomes. For the models predicting insight, dynamic gestures were associated with valid insights (Odds = 2.27, $d = 0.45$, $p = 0.014$), but any gesture was not ($p = 0.16$, $d = 0.21$). None of the other predictors had a significant interaction with dynamic gestures. For the models predicting proof, dynamic gestures were strongly associated with valid proofs (Odds = 3.81, $d = 0.74$, $p < 0.001$). Any gesture was not associated with valid proofs ($p = 0.45$, $d = 0.14$). None of the other predictors had a significant interaction with dynamic gestures. Being male was associated with a higher likelihood of insight and proof across models, while being a non-native English speaker was associated with a lower likelihood of insight. For all four models, spatial test score significantly positively predicted valid proofs and insights. For the dynamic gesture proof model, being a STEM major significantly positively predicted valid proofs.

As there were positive results for gestures predicting performance during uninhibited trials, these results are consistent with H3-by-product. We would also expect this effect under H3-facilitation, but the results from RQ2 rule out facilitation as being the appropriate hypothesis.

5. Discussion

Our results support our third hypothesis, the ***byproduct hypothesis*** (Table 1; H1a-byproduct, H1b, H2-byproduct, H3-byproduct). Gesture inhibition had no significant effect on a variety of outcome measures – including speech patterns, recall, and giving valid intuitions, insights, and proofs for geometry conjectures. Analyses of interaction effects suggested that inhibition had no significant effect regardless of gender, language status, spatial skills, phonemic fluency, college major, or math course-taking history. However, gesturing, particularly making dynamic gestures, was associated with improved insight and proof, and inhibiting gestures via our gesture inhibition rig dramatically reduced tendency to gesture. How can these results be reconciled?

An explanation for these findings is that gesture is merely a by-product of – rather than a causative factor in – valid geometric proof construction. In other words, college students in the study who tended to do better at these geometry tasks also tended to gesture more and to gesture in certain ways. But their gestures may not actually be *influencing* or *causing* their valid reasoning. If this was the case, when inhibited from gesturing, we would expect them to see no ill effect, since their gestures may have been simply an outgrowth of valid reasoning that was already established and that would have taken place with or without gesture.

This interpretation runs counter to other accounts in the literature – including the GSA framework (Hostetter & Alibali. 2008), which posits that gestures can relieve cognitive load, and/or that inhibiting gestures may increase cognitive load. There are a large number of other studies suggesting that gesture inhibition impacts a variety of outcomes, from language to recall to problem-solving performance (Alibali & Kita, 2010; Beattie & Coughlan, 1999; Frick-Horbury & Guttentag, 1998; Goldin-Meadow et al., 2001; Hostetter et al., 2007; Nathan & Martinez,

Table 6
Models predicting insight and proof, with gesture as predictor.

Fixed Effects	Model 5: Any Gesture Insight	Model 6: Dynamic Gesture Insight	Model 7: Any Gesture Proof	Model 8: Dynamic Gesture Proof
	B (SE)^{Sig}	B (SE)^{Sig}	B (SE)^{Sig}	B (SE)^{Sig}
(Intercept)	−0.74 (0.57)	−0.84 (0.58)	−2.79 (0.75)	−3.40 (0.79)
Female	(ref.)	(ref.)	(ref.)	(ref.)
Male	0.66 (0.26)*	0.70 (0.27)**	0.91 (0.34)**	1.02 (0.35)**
Native English	(ref.)	(ref.)	(ref.)	(ref.)
Non-Native English	−0.87 (0.36)*	−0.81 (0.36)*	−0.30 (0.44)	−0.18 (0.44)
Non-STEM major	(ref.)	(ref.)	(ref.)	(ref.)
STEM major	0.27 (0.36)	0.36 (0.36)	0.80 (0.49)	1.01 (0.51)*
Phonemic Fluency	0.03 (0.03)	0.03 (0.03)	0 (0.04)	−0.02 (0.04)
Course Beyond Calc I	(ref.)	(ref.)	(ref.)	(ref.)
Course Below Calc 1	−0.54 (0.41)	−0.48 (0.41)	−0.62 (0.58)	−0.13 (0.54)
Course Calc 1	−0.41 (0.38)	−0.33 (0.38)	−1.10 (0.53)*	−0.75 (0.52)
Spatial Score	0.06 (0.03)*	0.06 (0.03)*	0.12 (0.05)*	0.13 (0.05)**
Any Gesture	0.38 (0.27)		0.26 (0.34)	
Dynamic Gesture		0.82 (0.33)*		1.34 (0.40)***

Note. (ref.) denotes the reference category. * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$.

2015; Rauscher et al., 1996; Wagner et al., 2004). However, these studies had varied populations, inhibition methods, sample sizes, and content areas, and these methodological details may be important in explaining the differing results. The difference between child versus young adult cognitive processing may be particularly important. In addition, this study directly addresses a number of correlational studies that link more gesturing or gesturing in certain ways to better outcomes for mathematical activities (Gerofsky, 2010; Nathan & Walkington, 2017; Pier et al., 2019; Walkington et al., 2014). This effect, while reliably detected across studies and relatively large in size, may not be a causative relationship, and thus may not be a useful malleable factor to consider intervening upon.

Theories of cognition-action transduction posit that gestures can give learners new ideas (Nathan, 2017), and this hypothesis is supported by studies suggesting a positive effect for directing gestures on mathematical outcomes (Goldin-Meadow et al., 2009; Novack et al., 2014). While directing participants to gesture in specific ways that are known to be effective may have a transductive impact on cognitive states, this study suggests that for young adults in mathematics, it may not be the case that personally-generated gestures have this impact. Rather than our learners' gestures giving them new actionable ideas about geometric proofs, the present study suggests that these gestures may simply illustrate reasoning that would be happening one way or another. In order for transduction to reliably take place in a study like the present one, gestures may need to be explicitly directed by an outside agent or through structuring of the environment. Studies that come to conclusions that gesturing improves learning should be careful to qualify how those gestures were directed or structured to come about, to build a clear evidence base.

In addition, our results support the hypothesis that dynamic gestures tend to be significantly associated with insight and proof whereas a general category of "any gesture" is not, indicating dynamic gestures' important association with valid geometric reasoning. Dynamic gestures show transformations of mathematical objects through multiple states, and thus may signal learners' understanding of geometric relations. This is consistent with predictions about dynamic gestures given in Nathan and Walkington (2017). However, this again may not be a causative relationship. We also found that gestures and dynamic gestures did not show a significant association with recall or intuition. This is consistent with Nathan and Martinez (2015), who found spatial tapping selectively impaired making inferences from reading science text, but showed no difference on textbase recall or performance.

5.1. Limitations and future directions

Although it is difficult to extend implications for age groups beyond the one examined in this study (college students at a selective university), among a population similar to ours, this type of gesture inhibition may not be detrimental to students' reasoning on geometric proof tasks. Despite the fact that we used a college student population, note that the tasks we gave participants were challenging enough for them, as demonstrated by the low success rates on our most stringent measure of mathematical reasoning (proof).

In addition, although studies reviewed here did involve coding of gestures, most looked only at hand gestures. Here we took a broader view of what constitutes a gesture, as participants seemed able to point to and represent the movement of mathematical objects with their heads rather than their hands. Whether head gestures would serve the same purpose in other content domains is less clear – perhaps they are more useful in geometric reasoning. While the studies that involve spatial tapping do have some clear methodological limitations, one of the major reasons this method of gesture inhibition is used is because it may be more likely to inhibit *all* gestures – not just hand gestures. However, it is important to note that even with our liberal definition of gestures, we did see gesture rates fall dramatically with our inhibition method, with no accompanying significant change in any performance or language measure. It is also worth mentioning that participants could have been making "micro movements" inside the gloves or with other body parts that were too subtle to detect using visual means. The issue of micro movements is a limitation of all gesture inhibition studies that use a physical form of hand restriction. Despite this issue, most prior gesture inhibition studies using hand restriction have found significant differences.

A number of interventions for mathematics learning have been developed that direct learners to gesture in particular ways (e.g., Agostinho et al., 2015; Ginns et al., 2016; Goldin-Meadow et al., 2009; Hu et al., 2015; Nathan & Walkington, 2017; Ottmar & Landy, 2017; Petrick & Martin, 2012; Smith, King, & Hoyte, 2014). These interventions may be effective because they give learners well-thought-out gestural schemas to use that are more effective than the gestures that learners would use if left to their own devices. In other words, having learners simply gesture more may not be a particularly beneficial path – in order for gesture to give learners new ideas about geometry, learners may need to be taught to adopt specific types of gestures that are specially designed to demonstrate and embody powerful ideas and relationships. This may hold regardless of the level of mathematical expertise of the learner.

An important question is, how can these "effective" gestures be selected, and then how can they be passed on to learners in personally meaningful ways? In addition, to what degree do learners need to actually be directed to produce specific gestures? Research by Abrahamson and Trninic (2015) on the development of proportional reasoning using the body does not direct learners to do particular gestures, but rather creates an environment with a "field of promoted action" that loosely encourages particular kinds of body movement through feedback and interaction. These gestures are developed spontaneously and are personal, while at the same time are structured by the environment. An alternative approach is illustrated in a video game for promoting geometry reasoning developed by Nathan and Walkington (2017) that elicits directed actions by having players mimic the actions performed by in-game avatars. Other interventions that give the learner highly specific instructions about tracing relational parts of geometric figures have also been found to be effective (Ginns et al., 2015; Hu et al., 2015). If gesture is going to be leveraged to play a causative role in mathematical reasoning, these varied forms of intervention are important considerations to be addressed by future work.

6. Conclusions

Prior research has suggested somewhat uniform, detrimental effects for gesture inhibition; however, we discovered that personally-generated gestures may not play a causative role in geometric reasoning, supporting the *gestures as a byproduct hypothesis*. These results call into question a long line of studies suggesting detrimental effects for gesture inhibition. They also problematize theories that suggest that gesture can play a causative role in supporting learners' reasoning. However, this is one of the only studies looking at the effects of inhibition on *mathematical* reasoning, in the area of high school level geometry. This suggests that the causative role of gesture in promoting changes in language, recall, and reasoning might be different in other domains and developmental levels. While mathematical ideas are inherently embodied and perceptual (Lakoff & Núñez, 2000), it may be challenging for learners to spontaneously and meaningfully connect the embodied roots of mathematical ideas to the abstractions, definitions, and theorems they encounter in the typical classroom. What we tested here were highly academic tasks, firmly situated in the system of "school mathematics." In the mathematics classroom, learners are accustomed to only expressing their mathematical reasoning via written notation, rather than oral language accompanied by action. The ways in which learners use gestures and their bodies is certainly influenced by this overarching system of norms and beliefs in mathematics, and learners might need to be instructed upon particular gestural schemas to bridge this divide and

realize the power of gestures.

Acknowledgements

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A160020 to University of Wisconsin – Madison. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. Prior versions of this paper were presented at the 2018 Annual Meeting of the American Educational Research Association and the 2018 International Conference of the Learning Sciences. Thank you to Kasi Holcomb-Webb and Diana Vu for their assistance with the data, and to Peter Steiner for his assistance with the analysis. We also appreciate the contributions of our advisory board – Dor Abrahamson, Kristen Bieda, Martha Alibali, Elise Lockwood, Caro Williams-Pierce, Carmen Smith, and David Landy.

Appendix D. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.learninstruc.2019.101225.

Appendix A. Insight and proof codebook

Harel and Sowder (2005) describe how valid proofs involve (1) operational thought, where provers perform valid operations and transformations on mathematical objects and observe their results, (2) logical inference, where provers progress through a deductive structure of goals and subgoals to formulate an argument, and (3) generality, where provers show the conjecture holds for all cases under consideration.

Conjecture	Criteria for Correct Insight	Criteria for Correct Proof
An angle bisector of any angle of a triangle also bisects the opposite side.	Theparticipant states the conjecture is false because will not cut in half or not always cut in half. The participant states the conjecture is only true for some triangles but not for others, but does not get more specific than that.	The participant states that the conjecture is false because the angle bisector will not always cut the opposite side in half - the bisector will only cut the opposite side in half if it is an equilateral triangle, or if it is an isosceles triangle with the angle bisector bisecting the base. The participant states the conjecture is false and gives a specific counter-example. The participant states the conjecture is false and invokes the angle bisector theorem, stating that the two sides that form the angle must be equal for the angle bisector to bisect the opposite side. The participant states the conjecture is false and uses law of sines to observe that the angle bisector will bisect the opposite side when the two other sides are equal.
Any translation can be expressed as the composition of two reflections.	The participant states the conjecture is true based on specific case(s) (e.g., just proves it for a point not a figure; or just for an equilateral triangle; just proves it for reflections over x/y axes) where 1 translation equals 2 reflections. The participant states the conjecture is true because two flips/reflections return the figure to its original orientation but in a different location.	The participant states the conjecture is true because creating two lines of reflection (that are perpendicular to the translation vector) will flip the figure over twice, returning it to its original orientation and moving it in the direction of the translation. The participant states the conjecture is true because any translation can be expressed as the addition of two reflection vectors.
If one angle of a triangle is larger than a second angle, then the side opposite the first angle is longer than the side opposite the second angle.	The participant states the conjecture is true because the size of the angles and the lengths of the sides are related or dependent on each other. The participant states the conjecture is true because angles and sides are in a proportional relationship or in correspondence. The participant states the conjecture is true based on testing specific examples.	The participant states the conjecture is true because as one angle of a triangle becomes larger, it forces the opposite to open up, making the side longer. The participant states the conjecture is true because as one angle becomes smaller, it forces the opposite side to compress, making the side shorter. The participant states the conjecture is true and invokes the law of cosines or law of sines.
The area of a parallelogram is the same as the area of a rectangle with the same length and height.	The participant states the conjecture is true because a parallelogram is a rectangle that is tilted over or pushed over. The participant states the conjecture is true because a parallelogram and a rectangle have the same area formula, or both have the formula area equals length/base times width/height.	The participant states the conjecture is true because a right triangle can be cut off of one side of the parallelogram and added to the other side, making a rectangle. The participant states the conjecture is true because a right triangle can be cut off of one side of the rectangle and added to the other side, making a parallelogram. The participant states the conjecture is true because all rectangles are parallelograms, thus the formula for the area of a parallelogram would hold for a rectangle as well.
The segment that joins the midpoints of two sides of any triangle is parallel to the third side.	The participant states the conjecture is true based on specific cases (case may be expressed in gesture only or they may say they are visualizing).	The participant states the conjecture is true because the addition of the midsegment would form a smaller triangle that is mathematically similar to the full triangle (by SAS similarity). Thus, the lines are parallel (by the converse of the Corresponding Angles Theorem).
Any rotation can be expressed as the composition of two reflections.	The participant states the conjecture is true based on specific case(s) or states that it is true only for specific cases (e.g., 180° rotation, 360° rotation) where 1 rotation equals 2 reflections.	The participant states the conjecture is true because creating two (intersecting) lines of reflection (with the point of intersection of the two lines being the center of rotation), and then reflecting a figure over them will give the same result as a rotation (when the angle of rotation is twice the angle between the intersecting lines). The participant states the conjecture is true because any rotation can be expressed as the addition of two reflection vectors.
Given that you know the measure of all three angles of a triangle, there is only one unique triangle that can be formed with these three angle measurements.	The participant states the conjecture is false because the "shape" or the "dimensions" can be different when angles are held constant. The participant states conjecture is false	The participant states the conjecture is false because although angles are held constant, side lengths can change (creating triangles that are similar but not congruent).

	because side lengths in addition to orientation can change. The partcipant states the conjecture is true because although they acknowledge side lengths can change, they imply that these are not "unique" new triangles.	The participant states the conjecture is false and gives a specific counter example.
If there are three points P, R, and Q in space, and the distance between P and Q equals the sum of the distance between P and R and the distance between R and Q, then points P, R, and Q must all lie along the same line.	The participant states the conjecture is true because the distances PR and RQ must be equal to PQ for it to be a line or that it is forced to be equal, but doesn't mention triangles. The participant states that point R must be in between points P and Q on the line.	The participant states the conjecture is true because if the points were not along the same line, they would form a triangle, and they cannot form a triangle when PR + RQ = PQ (by the Triangle Inequality). The participant states the conjecture is true because the shortest distance between any two points is a straight line. The participant states the conjecture is true because the shortest distance between P and Q would be segment PQ. Therefore, if PR + RQ = PQ, point R has to be on segment PQ. Hence they lie along the same line.

Note. The participants did not necessarily have to verbalize the portions of the proofs given in parentheses, as they are specific geometry facts and properties they may not remember off the top of their heads.

Appendix B. Summary Statistics for Coh-metrix and LIWC Categories Tested

Category	Mean	Standard Deviation	Cohen's *d* (uninhibited compared to inhibited)	Category	Mean	Standard Deviation	Cohen's *d* (uninhibited compared to inhibited)
DESWC	93.69	64.97	0.006	WRDADJ	84.57	40.39	0.119
DESWLsy	1.33	0.12	−0.003	WRDADV	70.35	40.11	−0.081
DESWLsyd	0.68	0.17	0.029	WRDPRO	90.37	50.85	0.052
DESWLlt	3.97	0.38	−0.050	WRDPRP1s	28.3	34.14	0.084
DESWLltd	2.25	0.37	−0.010	WRDPRP2	18.2	23.18	0.107
PCNARz	1.23	1.16	0.013	WRDPRP3p	7.56	16.65	−0.059
PCNARp	79.66	23.76	0.006	WRDFRQc	2.42	0.23	−0.008
PCSYNz	−1.00	1.48	−0.025	WRDFRQa	3.12	0.15	−0.004
PCSYNp	29.16	29.13	−0.014	WRDFRQmc	1.09	0.92	0.053
PCCNCz	−1.80	1.34	0.055	WRDAOAc	333.24	109.28	0.066
PCCNCp	13.11	21.59	−0.050	WRDFAMc	572.36	12.27	−0.048
PCREFz	1.83	2.67	−0.051	WRDCNCc	335.13	25.95	−0.029
PCREFp	72.43	32.77	0.016	WRDIMGc	373.35	26.92	0.015
PCDCz	2.26	2.38	0.072	WRDMEAc	391.64	22.28	0.024
PCDCp	81.62	29.88	0.080	WRDPOLc	4.24	0.76	−0.024
PCVERBz	0.72	1.78	−0.012	WRDHYPn	6.40	1.05	−0.003
PCVERBp	61.8	33.05	0.015	WRDHYPv	1.25	0.33	0.045
PCCONNz	−1.52	2.42	0.094	WRDHYPnv	1.30	0.34	0.036
PCCONNp	29.38	35.10	0.041	RDFRE	69.3	16.47	−0.040
PCTEMPz	0.48	5.58	0.056	RDFKGL	9.59	5.65	0.010
PCTEMPp	81.15	31.22	0.001	RDL2	26.70	14.91	−0.013
CRFNO1	0.39	0.38	0.031	WC	89.96	62.46	0.007
CRFAO1	0.57	0.36	0.041	Analytic	44.68	29.86	0.020
CRFSO1	0.44	0.38	0.032	Clout	37.79	23.93	0.063
CRFNOa	0.39	0.36	0.014	Authentic	49.22	33.24	0.047
CRFAOa	0.56	0.35	0.029	Tone	63.67	26.5	−0.002
CRFSOa	0.43	0.36	0.021	WPS	21.31	12.12	−0.075
CRFCWO1	0.25	0.21	−0.043	Sixltr	15.10	5.59	−0.027
CRFCWO1d	0.13	0.13	0.111	Dic	89.21	4.95	0.083
CRFCWOa	0.25	0.20	−0.063	function	57.09	5.83	0.068
CRFCWOad	0.14	0.11	0.097	pronoun	13.83	6.25	0.106
CRFANP1	0.47	0.38	−0.018	ppron	5.83	4.29	0.101
CRFANPa	0.44	0.36	0.008	i	3.07	3.89	0.105
LSASS1	0.34	0.25	−0.033	you	1.90	2.44	0.107
LSASS1d	0.17	0.16	0.073	they	0.60	1.44	0.030
LSAGN	0.25	0.11	0.058	ipron	7.99	3.95	0.055
LSAGNd	0.25	0.11	0.015	article	10.52	4.86	0.004
LDTTRc	0.67	0.14	0.001	prep	9.52	3.87	−0.065
LDTTRa	0.54	0.13	0.019	auxverb	13.20	4.24	0.065
LDMTLD	37.84	13.18	−0.043	adverb	4.73	3.55	−0.124
LDVOCD	13.72	19.89	−0.001	conj	10.00	4.23	0.035
CNCAll	109.91	40.57	0.073	negate	2.10	3.05	0.141
CNCCaus	44.32	26.63	0.018	verb	18.00	6.22	0.076
CNCLogic	76.35	31.81	0.098	adj	6.49	4.14	0.093
CNCADC	15.52	18.65	−0.025	compare	4.78	3.83	0.024
CNCTemp	14.46	16.42	−0.111	interrog	1.02	1.64	−0.253
CNCTempx	4.39	10.68	−0.160	number	3.41	3.67	−0.028
CNCAdd	39.61	28.83	0.136	quant	2.19	2.55	−0.137
CNCPos	98.82	42.39	0.043	affect	2.92	2.56	0.257
CNCNeg	11.25	15.67	0.048	posemo	2.73	2.49	0.286
SMCAUSv	17.39	21.78	0.021	social	3.94	3.32	0.027
SMCAUSvp	56.11	31.14	0.055	cogproc	18.36	7.02	0.045
SMINTEp	10.13	14.56	0.005	insight	3.78	4.04	0.112
SMCAUSr	1.86	1.60	0.080	cause	3.16	2.50	−0.018
SMINTEr	1.37	1.54	−0.074	discrep	3.31	2.78	0.028
SMCAUSlsa	0.16	0.12	−0.020	tentat	3.34	2.8	0.020

SMCAUSwn	0.41	0.15	−0.017	certain	3.09	2.87	0.150
SMTEMP	0.82	0.57	0.053	differ	5.41	3.50	−0.007
SYNLE	4.37	4.92	−0.059	percept	3.01	2.64	−0.015
SYNNP	0.75	0.29	−0.005	see	2.32	2.42	−0.066
SYNMEDpos	0.67	0.21	0.012	hear	0.53	1.03	0.005
SYNMEDwrd	0.79	0.22	−0.001	drives	2.43	3.00	−0.006
SYNMEDlem	0.77	0.22	−0.001	power	1.28	2.40	−0.048
SYNSTRUTa	0.08	0.10	−0.027	focuspast	0.72	1.29	0.301
SYNSTRUTt	0.08	0.09	−0.041	focuspresent	14.09	6.13	0.107
DRNP	334.09	53.40	0.021	focusfuture	1.61	1.77	−0.023
DRVP	226.57	57.65	0.047	relativ	12.17	6.21	0.115
DRAP	39.35	26.93	0.026	motion	1.78	2.42	−0.037
DRPP	73.21	33.98	−0.098	space	8.49	5.28	0.099
DRNEG	19.02	26.57	0.116	time	1.94	1.94	−0.111
DRGERUND	12.40	16.34	−0.072	informal	4.56	3.58	0.066
DRINF	10.35	14.83	−0.027	netspeak	0.72	1.56	0.189
WRDNOUN	172.93	56.03	0.076	assent	0.89	1.48	0.169
WRDVERB	93.42	37.99	0.046	nonflu	2.97	2.82	−0.013

Appendix C. Regression Coefficients for the effect of gesture inhibition on language categories

Category	Regression coefficient - Inhibited	Bootstrapped standard error of coefficient	*p*	*p*(i) (FDR)
pronoun	−0.674	0.303	0.028	0.000
SYNNP	0.032	0.015	0.032	0.001
WRDPRO	−4.722	2.217	0.036	0.001
focuspast	−0.178	0.086	0.040	0.001
LDMTLD	1.645	0.844	0.054	0.002
ppron	−0.401	0.207	0.056	0.002
Analytic	3.020	1.647	0.069	0.002
CRFSO1	0.036	0.021	0.084	0.003
WRDCNCc	2.187	1.412	0.124	0.003
SMCAUSr	−0.166	0.108	0.127	0.003
function.	−0.545	0.358	0.130	0.004
CRFSOa	0.029	0.019	0.132	0.004
quant	0.215	0.143	0.135	0.004
DRVP	−4.334	2.921	0.141	0.005
DRPP	2.788	1.900	0.145	0.005
CNCAdd	−2.489	1.700	0.146	0.005
WRDHYPv	0.031	0.021	0.149	0.006
WRDIMGc	2.297	1.605	0.155	0.006
CRFNOa	0.027	0.020	0.166	0.006
nonflu	0.220	0.162	0.176	0.007
DRNP	−3.983	2.948	0.179	0.007
PCNARz	−0.070	0.052	0.182	0.007
posemo	−0.222	0.169	0.191	0.008
number	0.268	0.210	0.204	0.008
affect	−0.224	0.177	0.208	0.008
CNCPos	−3.104	2.488	0.215	0.009
DRINF	0.968	0.781	0.218	0.009
WRDADJ	2.875	2.343	0.223	0.009
WRDVERB	−2.693	2.255	0.235	0.010
LSAGN	0.007	0.006	0.236	0.010
CRFANP1	−0.024	0.020	0.242	0.010
ipron	−0.276	0.240	0.254	0.011
CRFNO1	0.023	0.020	0.258	0.011
Dic	−0.341	0.303	0.263	0.011
CNCAll	−2.838	2.570	0.272	0.012
WRDMEAc	1.611	1.486	0.281	0.012
i	−0.231	0.217	0.289	0.013
SMINTEp	−0.917	0.878	0.299	0.013
netspeak	−0.075	0.073	0.307	0.013
WRDPRP1s	−1.997	1.951	0.308	0.014
assent	−0.102	0.100	0.311	0.014
LSASS1	0.013	0.013	0.315	0.014
SMCAUSv	1.233	1.237	0.321	0.015
SMINTEr	0.098	0.099	0.324	0.015
drives	−0.183	0.187	0.329	0.015
WRDPRP3p	1.053	1.094	0.338	0.016
CNCTempx	0.607	0.633	0.340	0.016
percept	−0.149	0.156	0.342	0.016
focusfuture	−0.110	0.119	0.356	0.017
Clout	−1.303	1.436	0.366	0.017
WRDFRQc	−0.012	0.013	0.367	0.017
cause	0.129	0.145	0.375	0.018
LSAGNd	0.006	0.006	0.379	0.018
article	0.216	0.245	0.380	0.018
PCCONNz	0.127	0.147	0.388	0.019

WRDFAMc	−0.540	0.624	0.389	0.019
WRDPRP2	−1.168	1.363	0.393	0.019
prep	0.197	0.230	0.394	0.020
you	−0.122	0.144	0.398	0.020
PCCNCp	1.373	1.639	0.404	0.020
WPS	0.540	0.672	0.423	0.021
compare	−0.148	0.193	0.445	0.021
DRGERUND	0.787	1.028	0.446	0.021
SMCAUSwn	−0.007	0.009	0.447	0.022
CRFANPa	−0.015	0.020	0.450	0.022
CRFCWO1d	−0.006	0.009	0.457	0.022
WC	2.057	2.798	0.464	0.023
verb	−0.242	0.333	0.469	0.023
RDFRE	−0.676	0.933	0.471	0.023
see	−0.102	0.145	0.485	0.024
DESWLlt	0.013	0.020	0.497	0.024
conj	−0.176	0.263	0.505	0.024
DESWC	1.896	2.913	0.517	0.025
focuspresent	−0.215	0.333	0.520	0.025
cogproc	0.237	0.382	0.536	0.025
WRDHYPnv	0.010	0.017	0.543	0.026
CRFCWOa	0.006	0.011	0.555	0.026
SYNLE	−0.184	0.311	0.556	0.026
adj	−0.130	0.236	0.584	0.027
SYNSTRUTa	−0.003	0.005	0.592	0.027
PCCNCz	0.051	0.096	0.594	0.027
RDFKGL	0.175	0.329	0.595	0.028
social	−0.092	0.178	0.605	0.028
CNCTemp	0.612	1.191	0.608	0.028
CRFCWOad	−0.003	0.007	0.615	0.029
SYNSTRUTt	−0.003	0.005	0.617	0.029
discrep	0.087	0.178	0.626	0.029
Sixltr	0.170	0.353	0.631	0.030
WRDADV	1.051	2.215	0.636	0.030
Tone	−0.763	1.614	0.637	0.030
PCDCz	−0.068	0.147	0.644	0.031
RDL2	−0.339	0.737	0.647	0.031
PCVERBz	−0.044	0.098	0.651	0.031
Authentic	−0.941	2.104	0.655	0.032
motion	−0.060	0.137	0.661	0.032
CRFAOa	−0.008	0.019	0.669	0.032
LDTTRa	−0.003	0.007	0.674	0.033
PCDCp	0.803	2.100	0.703	0.033
SMTEMP	0.016	0.041	0.704	0.033
they	0.035	0.091	0.706	0.034
PCVERBp	−0.650	1.855	0.727	0.034
CNCCaus	−0.547	1.568	0.728	0.034
PCTEMPp	0.775	2.233	0.729	0.035
DESWLsy	0.002	0.007	0.731	0.035
PCREFz	0.046	0.138	0.741	0.035
SYNMEDlem	−0.005	0.015	0.743	0.036
WRDHYPn	0.020	0.062	0.747	0.036
LDVOCD	−0.276	0.853	0.747	0.036
auxverb	−0.090	0.286	0.753	0.037
WRDNOUN	0.885	2.913	0.762	0.037
PCTEMPz	0.118	0.398	0.768	0.038
space	0.075	0.263	0.777	0.038
interrog	0.030	0.110	0.786	0.038
insight	−0.053	0.209	0.798	0.039
LSASS1d	0.003	0.010	0.805	0.039
adverb	−0.054	0.219	0.806	0.039
power	0.028	0.118	0.811	0.040
PCCONNp	0.440	2.023	0.828	0.040
time	0.028	0.134	0.833	0.040
WRDAOAc	1.310	6.383	0.838	0.041
SYNMEDpos	−0.003	0.014	0.839	0.041
tentat	0.040	0.197	0.840	0.041
differ	0.044	0.225	0.844	0.042
PCREFp	−0.366	1.880	0.846	0.042
SYNMEDwrd	−0.003	0.015	0.848	0.042
CRFCWO1	0.002	0.011	0.848	0.043
PCNARp	−0.233	1.243	0.851	0.043
LDTTRc	0.001	0.007	0.855	0.043
CNCNeg	−0.121	0.927	0.897	0.044
CRFAO1	−0.002	0.019	0.901	0.044
hear	−0.008	0.067	0.902	0.044
certain	−0.023	0.197	0.905	0.045
DESWLsyd	0.001	0.010	0.912	0.045
PCSYNz	0.008	0.079	0.914	0.045
SMCAUSlsa	−0.001	0.007	0.915	0.046

WRDPOLc	0.005	0.051	0.925	0.046
DESWLltd	−0.002	0.021	0.925	0.046
WRDFRQa	0.001	0.008	0.936	0.047
informal	0.014	0.208	0.945	0.047
CNCLogic	−0.125	2.138	0.954	0.047
relativ	−0.018	0.340	0.957	0.048
WRDFRQmc	−0.003	0.061	0.958	0.048
negate	−0.008	0.183	0.966	0.048
DRAP	0.067	1.792	0.970	0.049
CNCADC	0.040	1.077	0.970	0.049
PCSYNp	−0.059	1.595	0.970	0.049
SMCAUSvp	−0.051	1.601	0.975	0.050
DRNEG	0.002	1.655	0.999	0.050

Note. Under the FDR procedure, the value of $p(i)$ would have to be greater than the p-value for the coefficient to be considered significant.

References

Abrahamson, D., & Trninic, D. (2015). Bringing forth mathematical concepts: Signifying sensorimotor enactment in fields of promoted action. *ZDM Mathematics Education, 47*(2), 1–12. https://doi.org/10.1007/s11858-014-0620-0.

Agostinho, S., Tindall-Ford, S., Ginns, P., Howard, S. J., Leahy, W., & Paas, F. (2015). Giving learning a helping hand: Finger tracing of temperature graphs on an iPad. *Educational Psychology Review, 27*(3), 427–443. https://doi.org/10.1007/s10648-015-9315-5.

Alibali, M. W., & Kita, S. (2010). Gesture highlights perceptually present information for speakers. *Gesture, 10*(1), 3–28. https://doi.org/10.1075/gest.10.1.02ali.

Alibali, M. W., Kita, S., & Young, A. J. (2000). Gesture and the process of speech production: We think, therefore we gesture. *Language & Cognitive Processes, 15*(6), 593–613. https://doi.org/10.1080/016909600750040571.

Alibali, M. W., Yeo, A., Hostetter, A. B., & Kita, S. (2017). Representational gestures help speakers package information for speaking. In R. Breackinridge Chuch, M. W. Alibali, & S. D. Kelly (Eds.). *Why gesture: How the hands function in speaking, thinking, and communicating* (pp. 15–38). John Benjamins Publishing Company. https://doi.org/10.1075/gs.7.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617–645. https://doi.org/10.1146/annurev.psyc.59.103006.093639.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7.* Retrieved from http://CRAN.R-project.org/package=lme4.

Beattie, G., & Coughlan, J. (1999). An experimental investigation of the role of iconic gestures in lexical access using the tip-of-the-tongue phenomenon. *British Journal of Psychology, 90*(1), 35–56. https://doi.org/10.1348/000712699161251.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B,* 289–300. https://doi.org/10.2307/2346101.

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine, 19*(22), 3127–3131. https://doi.org/10.1002/1097-0258(20001130)19:22<3127::AID-SIM784>3.0.CO;2-M.

Cook, S. W., & Goldin-Meadow, S. (2006). The role of gesture in learning: Do children use their hands to change their minds? *Journal of Cognition and Development, 7*(2), 211–232. https://doi.org/10.1207/s15327647jcd0702_4.

desRosiers, G., & Kavanagh, D. (1987). Cognitive assessment in closed head injury: Stability, validity and parallel forms for two neuropsychological measures of recovery. *International Journal of Clinical Neuropsychology, 9*, 162–173.

Ekstrom, R. B., French, J. W., Harman, H., & Derman, D. (1976). *Kit of factor-referenced cognitive tests (rev. ed.).* Princeton, NJ: Educational Testing Service.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149–1160. https://doi.org/10.3758/BRM.41.4.1149.

Frick-Horbury, Donna, & Guttentag, Robert E. (1998). The effects of restricting hand gesture production on lexical retrieval and free recall. *American Journal of Psychology, 111*, 43–62. https://doi.org/10.2307/1423536.

Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science, 18*(3), 233–239.

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.

Gerofsky, S. (2010). Mathematical learning and gesture: Character viewpoint and observer viewpoint in students' gestured graphs of functions. *Gesture, 10*(2), 321–343. https://doi.org/10.1075/gest.10.2-3.10ger.

Ginns, P., Hu, F. T., Byrne, E., & Bobis, J. (2016). Learning by tracing worked examples. *Applied Cognitive Psychology, 30*(2), 160–169. https://doi.org/10.1002/acp.3171.

Göksun, T., Goldin-Meadow, S., Newcombe, N., & Shipley, T. (2013). Individual differences in mental rotation. *Cognitive Processing, 14*, 153–162. https://doi.org/10.1007/s10339-013-0549-1.

Goldin-Meadow, S., Cook, S. W., & Mitchell, Z. A. (2009). Gesturing gives children new ideas about math. *Psychological Science, 20*(3), 267–272. https://doi.org/10.1111/j.1467-9280.2009.02297.x.

Goldin-Meadow, S., Nusbaum, H., Kelly, S., & Wagner, S. (2001). Explaining math: Gesturing lightens the load. *Psychological Science, 12*, 516–522. https://doi.org/10.1111/1467-9280.00395.

Graham, J. A., & Heywood, S. (1975). The effects of elimination of hand gestures and of verbal codability on speech performance. *European Journal of Social Psychology, 5*, 189–195. https://doi.org/10.1002/ejsp.2420050204.

Harel, G., & Sowder, L. (2005). Toward comprehensive perspectives on the learning and teaching of proof. In F. Lester (Ed.). *Second handbook of research on mathematics teaching and learning.* Reston, VA: National Council of Teachers of Mathematics.

Hegarty, M., Mayer, S., Kriz, S., & Keehner, M. (2005). The role of gestures in mental animation. *Spatial Cognition and Computation, 5*(4), 333–356. https://doi.org/10.1207/s15427633scc0504_3.

Hoetjes, M., Krahmer, E., & Swerts, M. (2014). Does our speech change when we cannot gesture? *Speech Communication, 57*, 257–267. https://doi.org/10.1016/j.specom.2013.06.007.

Hostetter, A. B., & Alibali, M. W. (2007). Raise your hand if you're spatial: Relations between verbal and spatial skills and gesture production. *Gesture, 7*(1), 73–95. https://doi.org/10.1075/gest.7.1.05hos.

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review, 15*(3), 495–514. https://doi.org/10.3758/PBR.15.3.495.

Hostetter, A. B., Alibali, M. W., & Kita, S. (2007). Does sitting on your hands make you bite your tongue? The effects of gesture prohibition on speech during motor descriptions. *Proceedings of the 29th annual meeting of the cognitive science society* (pp. 1097–1102). Mawhah, NJ: Erlbaum.

Hu, F. T., Ginns, P., & Bobis, J. (2015). Getting the point: Tracing worked examples enhances learning. *Learning and Instruction, 35*, 85–93. https://doi.org/10.1016/j.learninstruc.2014.10.002.

Krauss, R. M. (1998). Why do we gesture when we speak? *Current Directions in Psychological Science, 7*(2), 54–59. https://doi.org/10.1111/1467-8721.ep13175642.

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

McNamara, D. S., Louwerse, M. M., Cai, Z., & Graesser, A. (2013). *Coh-metrix (version 3.0).* : http://cohmetrix.com/ Available from .

McNeill, D. (1992). *Hand and mind: What gestures reveal about 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. (2017). One function of gesture is to make new ideas. In R. Breckinridge-Church, M. W. Alibali, & S. Kelly (Eds.). *Why Gesture?: How the hands function in speaking, thinking and communicating* (pp. 175–196). Amsterdam, the Netherlands: John Benjamins Publishing Company.

Nathan, M. J., & Martinez, C. V. (2015). Gesture as model enactment: The role of gesture in mental model construction and inference making when learning from text. *Learning: Research and Practice, 1*(1), 4–37. https://doi.org/10.1080/23735082.2015.1006758.

Nathan, M. J., & 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.

Nathan, M. J., Walkington, C., Boncoddo, R., Pier, E., Williams, C., & Alibali, M. (2014). Actions speak louder with words: The roles of action and pedagogical language for grounding mathematical reasoning. *Learning and Instruction, 33*, 182–193.

Nathan, M. J., Walkington, C., Vinsonhaler, R., Michaelis, J., McGinty, J., Binzak, J. V., et al. (April, 2018). Embodied account of geometry proof, insight, and intuition among novices, experts, and English language learners. *Presentation at the 2018 annual Meeting of the American educational research association* New York, NY.

Novack, M. A., Congdon, E. L., Hemani-Lopez, N., & Goldin-Meadow, S. (2014). From action to abstraction using the hands to learn math. *Psychological Science, 25*(4), 903–910. https://doi.org/10.1177/0956797613518351.

Ottmar, E., & Landy, D. (2017). Concreteness fading of algebraic instruction: Effects on learning. *The Journal of the Learning Sciences, 26*(1), 51–78. https://doi.org/10.1080/10508406.2016.1250212.

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/.

Petrick, C., & Martin, T. (2012). Mind your body: Learning mathematics through physical action. *Paper presented at the Annual meeting of the American educational research association* (Vancouver, Canada).

Pier, E. L., Walkington, C., Clinton, V. E., Boncoddo, R., Williams-Pierce, C., Alibali, M. A.,

et al. (2019). Embodied truths: How dynamic gesture and transformational speech contribute to mathematical proof practices. *Contemporary Educational Psychology, 58*, 44–57. https://doi.org/10.1016/j.cedpsych.2019.01.012.

Rauscher, Frances H., Krauss, R., & Chen, Y. (1996). Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological Science, 7*, 226–231. https://doi.org/10.1111/j.1467-9280.1996.tb00364.x.

Rimé, B., Schiaratura, L., Hupet, M., & Ghysselinckx, A. (1984). Effects of relative immobilization on the speaker's nonverbal behavior and on the dialogue imagery level. *Motivation and Emotion, 8*(4), 311–325. https://doi.org/10.1007/BF00991870.

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688 doi: 0.1016/j.tics.2016.07.002.

Sloutsky, V. M., Kaminski, J. A., & Heckler, A. F. (2005). The advantage of simple symbols for learning and transfer. *Psychonomic Bulletin & Review, 12*(3), 508–513. https://doi.org/10.3758/BF03193796.

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. https://doi.org/10.1016/j.jmathb.2014.09.001.

Van Merriënboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review, 17*(2), 147–177. https://doi.org/10.1007/s10648-005-3951-0.

Wagner, S., Nusbaum, H., & Goldin-Meadow, S. (2004). Probing the mental representation of gesture: Is handwaving spatial? *Journal of Memory and Language, 50*, 395–407. https://doi.org/10.1016/j.jml.2004.01.002.

Walkington, C., Boncoddo, R., Williams, C., Nathan, M., Alibali, M., Simon, E., et al. (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., Chelule, G., Woods, D., & Nathan, M.J. (in press). Collaborative gesture as a case of extended mathematical cognition. The Journal of Mathematical Behavior. https://doi.org/10.1016/j.jmathb.2018.12.002.

Walkington, C., Nathan, M., Wolfgram, M., Alibali, M., & Srisurichan, R. (2014). Bridges and barriers to constructing conceptual cohesion across modalities and temporalities: Challenges of STEM integration in the precollege engineering classroom. In S. Purzer, J. Strobel, & M. Cardella (Eds.). *Engineering in pre-college settings: Research into practice* (pp. 183–210). West Lafayette, Indiana: Purdue University Press.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review, 9*, 625–636. https://doi.org/10.3758/BF03196322.

Woods, D., & Fassnacht, C. (2012). *Transana v2.52*. Madison, WI: The Board of Regents of The University of Wisconsin System. http://transana.org.

Yeudall, L. T., Fromm, D., Reddon, J. R., & Stefanyk, W. O. (1986). Normative data stratified by age and sex for 12 neuropsychological tests. *Journal of Clinical Psychology, 42*(6), 918–946.

Zhang, Z., Lei, Y., & Li, H. (2016). Approaching the distinction between intuition and insight. *Frontiers in Psychology, 7*, 1195. https://doi.org/10.3389/fpsyg.2016.01195.