2022

Does Co-Speech Gesture Support Children's Analogical Reasoning? An Investigation Into the Differential Effects of Gesture on Learning

Katharine F. Guarino

LOYOLA UNIVERSITY CHICAGO

DOES CO-SPEECH GESTURE SUPPORT CHILDREN'S ANALOGICAL REASONING? AN INVESTIGATION INTO THE DIFFERENTIAL EFFECTS OF GESTURE ON LEARNING

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE GRADUATE SCHOOL

IN CANDIDACY FOR THE DEGREE OF

DOCTORATE OF PHILOSOPHY

PROGRAM IN DEVELOPMENTAL PSYCHOLOGY

BY

KATHARINE F. GUARINO

CHICAGO, IL

MAY 2022

ACKNOWLEDGMENTS

This research was made possible by Dr. Elizabeth M. Wakefield, my wonderful mentor and advisor, and my committee who provided invaluable guidance throughout this process. I would also like to extend my gratitude to all undergraduate research assistants for their help with data collection and data processing, and to all the children, parents, and teachers who participated in this study. Lastly, I would like to thank my wonderful husband, parents, and friends for their continuous support and belief in me that I could finish what I started.

CO-AUTHORSHIP

A version of Chapter 2 is in press in *Cognitive Development*, and is therefore greatly influenced by co-authors Elizabeth Wakefield, Robert Morrison, and Lindsey Richland.

A version of Chapter 3 is published in *Frontiers in Psychology*, and is therefore greatly influenced by co-author Elizabeth Wakefield.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Although the general consensus is that gesture supports learning across a wide range of learning contexts, nuances to gesture's effects are found across the gesture-for-learning literature. The purpose of this body of research was to advance our understanding of gesture's effect on learning. Specifically, we explored the utility of gesture in a domain that had not been considered in the gesture literature previously: analogical reasoning (Study 1). We aimed to understand *whether* gesture supports children's analogical reasoning ability and *why* gesture might support this type of reasoning. Specifically, we investigated whether gesture could support learning through directing visual attention, thereby minimizing the limitations of children's immature inhibitory control. Next, we investigated *when* gesture is most beneficial for learning across a wide range of contextual and situational variations of the learning environment: We explored whether there are different effects of gesture on analogical reasoning depending on children's cognitive capacities (Study 2) and identified *under which* conditions gesture is most beneficial using a meta-analytic approach (Study 3). Taken together, we provide further evidence that gesture is generally beneficial for learning, but that there are nuances to these effects both within the domain of analogical reasoning and across learning contexts. The effects of gesture depend on a variety of factors that comprise the instructional environment, including factors related to the learner themselves, the content being learned, and the gesture itself.

CHAPTER ONE

INTRODUCTION

Analogical reasoning is the ability to identify underlying schematic structure shared between representations. In its mature form, it is a powerful cognitive mechanism that contributes to a range of skills unique to humans, including innovation (Markman & Wood, 2009), inductive reasoning (Gentner, 2010), adaptive general intelligence (Gentner, 2010), and creativity (Sternberg, 1988; for review see Gentner & Smith, 2013). At their core, each of these skills require identifying relational similarity shared between contexts in order to make a conclusion about the contexts as a whole.

Whereas children and adults are faced with different types of analogy problems, analogical reasoning is nevertheless an important way children make sense of the world around them (e.g., Ferry, Hespos, & Gentner, 2015; Goswami, 2001; Rattermann & Gentner, 1998). Promoting analogical reasoning supports learning across instructional domains, including mathematics, science, and history education (for review see Richland & Simms, 2015). And while this ability supports a host of skills necessary for future academic achievement and career success (Gentner, 2010; Gentner & Smith, 2013), developing analogical reasoning is not easy. Young children struggle to extract underlying relations from a comparison and are more likely to make judgments based on surface features, or perceptual similarities, rather than based on relational information (Gentner, 1988). In educational settings, the use of analogies require explicit support to

ensure that children see the importance of relational thinking (Richland & Simms, 2015). Whereas prior work suggests that facilitating alignment during comparison opportunities helps children notice commonalities between contexts or problems (Richland, 2015), research addressing classroom tools to support this ability is limited. One tool that might be well-suited to facilitate this type of alignment is *gesture* – movements of the hands that convey meaning through their shape and trajectory.

The purpose of my dissertation is to not only understand *whether* gesture supports children's analogical reasoning ability, but also to address *why* and *how* gesture can support this type of reasoning. While gesture supports learning in other domains such as mathematics (e.g., Cook, Duffy, & Fenn, 2013) and word learning (e.g., Wakefield, Hall, James, & Goldin-Meadow, 2018a), the impact of gesture is nuanced. That is, the effect of gesture instruction depends both on children's current cognitive state and the context in which the gesture is used. Therefore, Study 1 will address whether spoken instruction incorporating gesture can support children's analogical reasoning above and beyond only spoken instruction, and if one way gesture is effective for learning is because of its ability to facilitate effective looking patterns. Studies 2 and 3 will explore whether there are differential effects of gesture on analogical reasoning depending on children's cognitive capacities (Study 2) and identify under which conditions gesture is most beneficial for learning (Study 3). Before discussing this line of research, I will review literature that informs this work, including the factors that both hinder and support the development of children's analogical reasoning ability, the mechanisms by which gesture supports learning, and why those mechanisms may be nicely suited to support children's analogical reasoning.

Analogical Reasoning Ability's Protracted Development

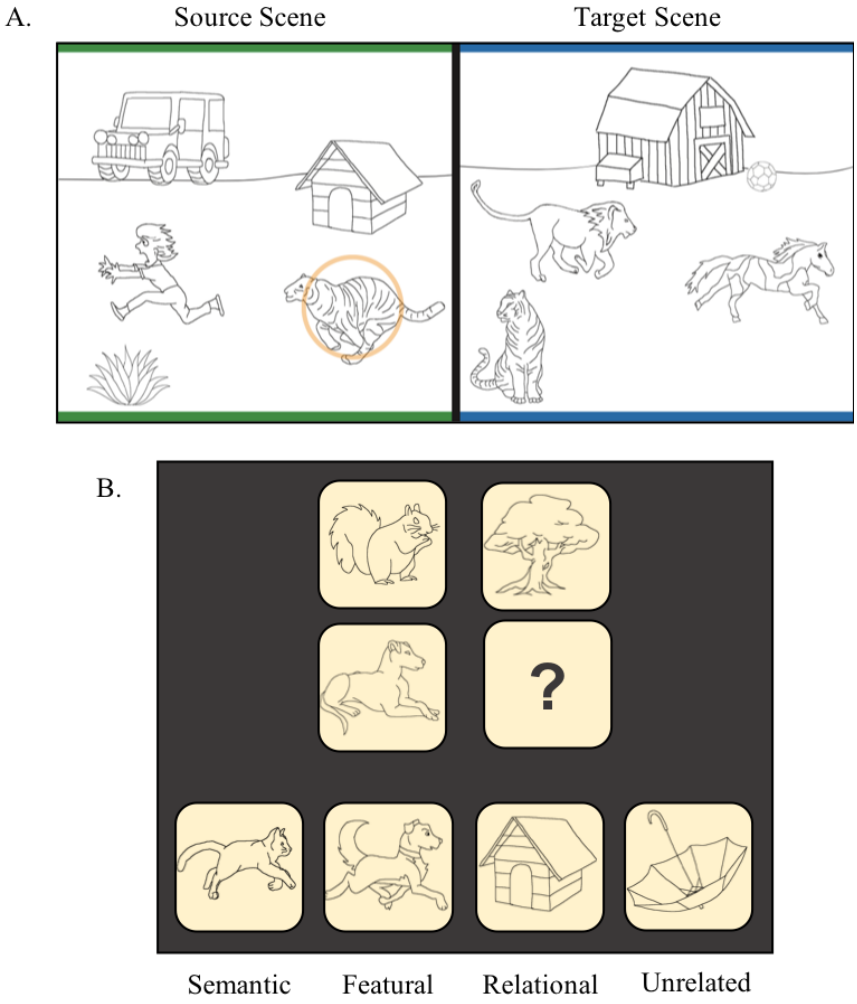
Analogical reasoning ability matures gradually over time, with rudimentary relational reasoning skills present as early as infancy and toddlerhood (e.g., Ferry et al., 2015; Goswami & Brown, 1989; Rattermann & Gentner, 1998). Most researchers consider the initial stages of an analogical reasoning ability that mirrors an underdeveloped adult-like ability to be present in children 3-5 years old (Alexander, Willson, White, & Fuqua, 1987; Goswami & Brown, 1989; Rattermann & Gentner, 1998), with nearly adult-like reasoning evident in early adolescence (e.g., Richland, Morrison, & Holyoak, 2006; Simms, Frausel, & Richland, 2018).

One type of analogy task that has been used to assess the development of children's analogical reasoning ability are *scene analogies*, in which children are asked to examine two scenes (e.g., a source and target scene) which contain both relational and featural similarities. Children are asked to identify an item in the target scene that corresponds *relationally* to a prompted item in the source scene. However, children often choose an item that corresponds *featurally* to the prompted item instead. This type of 'featural match' is one item in a target scene that is not incorporated in the relation of focus, but has great surface feature similarity to the prompted item in a source scene (Richland et al., 2006). For example, a source scene might show a tiger chasing a woman (relation of *chasing*), with the tiger prompted (Figure 1A). The corresponding target scene would contain a lion chasing a horse (relation of *chasing*) and a second tiger (the *featural match*). Here, the lion in the target scene would be the correct relational choice, and the tiger would be the incorrect featural match. When assessing ability to solve scene analogies across a wide age range, Richland and colleagues (2006) find that the lure of featural matches is greatest for the youngest children tested (3- and 4-year-olds), and decreases gradually over development,

with 6- and -7-year-olds, 9- and 10-year-olds, and 13- and 14-year-olds showing incrementally improved performance.

Researchers have also demonstrated protracted development of analogical reasoning using *propositional analogies* (e.g. Richland et al., 2006; Starr, Vendetti, & Bunge, 2018; Thibaut & French, 2016; Thibaut, French, & Vezneva, 2010). In these analogies, children are asked to select a D item from four choices that is relationally similar to the C item in the same way that the A and B items are similar. For example, in a pictorial propositional analogy, item A might be a picture of a squirrel, item B a picture of an tree, and item C a picture of a dog (Figure 1B). If the relation between items A and B is that squirrels *live in* trees, then the goal is to identify which option from the array is what dogs *live in*. In this example, four options would be presented in the choice array: three that are similar to the dog in different ways and an unrelated item (e.g., the umbrella). The three items include a relationally similar option (a dog house), a featural similar option (another dog), and a semantically or categorically similar option (a cat), where the correct choice would be the dog house, because dogs *live in* dog houses. In a study using pictorial propositional analogies, Thibaut and colleagues (2010) find that 6-year-old children consistently choose the featural match and rarely considered the other response options. While 8-year-olds are less likely to choose the featural matches, measures of reaction time suggest that it is more difficult for them to avoid the featural matches than 14-year-olds. This is further corroborated through the use of eye tracking, where children's visual attention is monitored while they solve analogies: Thibaut & French (2016) find that 5- and 8-year-olds, compared to adolescents or adults, spend a greater proportion of their time attending to featural matches rather than relational information.

Figure 1. A. An example of a scene analogy. B. An example of a propositional (A:B::C:D) analogy.



Note. A. If the prompted item in the source scene is the chasing tiger, the correct relational choice is the chasing lion, and the featural match to avoid is the sitting tiger in the target scene. B. The correct relational choice is the dog house, because dogs live in dog houses like squirrels live in trees, and the featural match to avoid is the other dog.

Relation Between Children's Cognitive Profile and Analogical Reasoning Ability

Across the field, the general consensus is that children struggle with analogical reasoning because they have a bias to focus on featural similarities shared between contexts, rather than structural or relational similarities. Under this premise, children are unable to reason analogically

until they have undergone a *relational shift*, at which point the bias to attend to featural matches decreases (Ratterman & Gentner, 1998). Two hypotheses have been put forth to account for the relational shift: 1) increases in domain relevant knowledge (i.e., relevant relational knowledge), and 2) increases in executive functions of working memory (WM) capacity and inhibitory control (IC). Both hypotheses suggest that development of analogical reasoning ability happens gradually across early childhood, and reaching adult-like ability in early adolescence. Development of analogical reasoning ability is best accounted for when taking into consideration both of these hypotheses, where children's *cognitive profile* determines their current analogical reasoning ability. In the case of analogical reasoning, children's cognitive profile is comprised of varying degrees of relational knowledge, WM, and IC – that affect analogical reasoning independently and in combination.

The first hypothesis suggests that increases in domain relevant knowledge may account for children's relational shift. In general, domain relevant knowledge refers to a familiarity with the relational orientations shared between objects within an analogy (e.g., size, locations). For example, knowledge about the relational category of 'size' would indicate an understanding that if triangle A is half the size of triangle B, then triangle A is smaller than triangle B. Researchers who believe that limited relational knowledge underlies children's difficulty solving problems of analogy also propose that young children *have* the ability to reason analogically – they are just *unable* to use the ability without the proper domain knowledge (Rattermann & Gentner, 1998). Supporting evidence demonstrates that children as young as 3 years old can reason relationally when they have sufficient familiarity with the tested domain (e.g., Goswami, 1989; Goswami & Brown, 1989). For example, 4- and 5-year-olds are able to solve A:B::C:D analogies using at-

tribute blocks that varied on dimensions of color, shape, and size (Alexander et al., 1987). Alexander and colleagues (1987) suggest that young children are able to successfully solve these analogies because they have the necessary relational knowledge about colors, shapes, and sizes, since these blocks are a common toy found in most homes and daycares.

However, while most researchers in the field agree that accretion of relational knowledge plays a necessary role in analogical reasoning development, the second hypothesis suggests this factor does not fully account for children's difficulty solving analogies: Other factors such as WM and IC also contribute to improvements in this ability over development (e.g., Doumas, Morrison, & Richland, 2018; Simms et al., 2018). The role of WM in analogical reasoning is well-supported. To successfully solve an analogy, one needs to simultaneously process multiple contexts and pieces of information. However, young children's ability to process multiple relations simultaneously is limited by their WM capacity. Over development, as the ability to accommodate larger demands in WM increases, so does children's analogical reasoning ability (e.g., Gick & Holyoak, 1980; Halford, 1993; Simms et al., 2018).

Similarly, researchers agree that successful analogical reasoning ability relies on effective IC. Whereas the role of IC in adult analogical reasoning has long been established (e.g., Morrison et al., 2004; Viskontas, Morrson, Holyoak, Hummel, & Knowlton, 2004), there is a growing body of work that posits that IC is also important for the development of analogical reasoning in children (e.g., Begolli et al., 2018; Doumas et al., 2018; Morrison, Doumas, & Richland, 2011; Richland, Chan, Morrison, & Au, 2010; Richland et al., 2006; Thibaut et al., 2010). When solving analogies, a reasoner might have to inhibit a more salient, featural match response, and select a less salient, relational match (e.g., Richland et al., 2006; Viskontas et al.,

2004). Across analogical reasoning paradigms (e.g., propositional and scene analogies), researchers suggest that children's limited IC accounts, in large part, for their difficulty solving analogies (Thibaut & French, 2016; Thibaut et al., 2010). That is, early in development young children struggle with tasks requiring inhibition of salient featural responses, yet over age as their IC develops, their analogical reasoning ability improves (e.g., Thibaut et al., 2010).

Increases in children's domain relevant knowledge, WM, and IC over development individually and jointly contribute to their relational shift, and ultimately, to successful analogical reasoning. Simulations in computational models of analogical reasoning suggest an interplay between these two executive functions, where IC serves as a gating mechanism for WM (Morrison et al., 2011). When faced with a scene analogy (e.g., Figure 1A), individuals activate information stored in long term memory to make sense of the scenes and complete the task. This activated information allows them to choose an item in the target scene that is in the same place in the pattern as the prompted item (e.g., the tiger) in the source scene. IC determines what information will remain active in WM, and therefore, predicts which item an individual will choose. By this theory, adults, who have higher IC, are able to more systematically gate relational knowledge in WM which helps them complete the task – in this case, understanding the idea of chasing, and holding that in WM while determining that the most relevant relational correspondence is that the tiger in the source scene and the lion in the target scene are both chasing. In contrast, if an individual has low IC, the gating mechanism does not function efficiently, the featural information that is most saliently related to the prompted item will remain activated and dominate WM. This is the case with young children, who are unable to inhibit the activation of the concept 'tiger' and therefore choose the salient, featural match. Age has been shown regularly to predict

the likelihood of selecting a distractor in scene analogy tasks, corroborating this account (Murphy, Zheng, Shivaram, Vollman, & Richland, 2021; Richland et al., 2006; Richland et al., 2010). This finding that IC, and executive functions more broadly, are important predictors of analogical reasoning ability has been corroborated across a number of methodological approaches including computational modeling (Doumas et al., 2018; Morrison et al., 2011; Morrison et al., 2004; Viskontas et al., 2004), longitudinal work (Richland & Burchinal, 2013), and eye tracking (e.g., Glady, Thibaut, & French, 2010; Starr et al., 2018).

In sum, children's analogical reasoning ability is predicted by the joint development of domain relevant knowledge, WM, and IC. However, because most researchers assess the development of analogical reasoning using propositional or scene analogies, and the domain relevant knowledge necessary to solve these problems is acquired early in life, most would agree that development of this ability hinges on maturation of WM and IC. In my work, I focus on the role of IC in the development of analogical reasoning ability, given that previous work seems to be at a consensus that the constraints of children's WM are due to limitations of IC. Furthermore, as discussed in detail in the following section, I anticipate that instruction incorporating gesture will be more closely associated with its impact on children's ability to avoid prepotent responses (i.e., IC), rather than its ability to reduce WM constraints.

Children's Analogical Reasoning May Benefit From Instructional Gesture

An effective instructional tool for supporting young children's analogical reasoning ability should help them focus on relational similarities during comparisons and avoid the lure of featural responses. There are two key ways this could be achieved: 1) driving attention away from irrelevant components of a comparison and towards relationally similar components, or 2)

facilitating comprehension of potentially ambiguous speech that may accompany analogical reasoning instruction. In other words, an effective instructional tool should focus attention *and* support speech comprehension. Fortunately, there is a readily-accessible tool that is naturally used in the classroom by teachers (Alibali & Nathan, 2007; Flevares & Perry, 2001) that may be nicely suited to serve these functions: co-speech gesture.

Although many hand and body movements can be considered *gestures* (Kendon, 2004; McNeill, 1992), here, we refer specifically to *representational gestures*, which are hand movements that represent information through their form and trajectory. Representational gestures can be further separated into three categories: *deictic gestures* point to a referent in the physical environment, *iconic gestures* represent a concrete action or object, and *metaphoric gestures* represent an abstract idea through a visuospatial representation. While both iconic and metaphoric gestures represent a referent in a concrete, visuospatial manner through the motion or shape of the hands, they do this in different ways depending on the accompanying spoken information. That is, the same concrete gesture can be either iconic or metaphoric depending on the verbal context. Take for example when a person uses a two-handed gesture to represent that one entity is above another in space and verbally stating that 'one entity is taller than the other'. In this case, the accompanying gesture is iconic because it represents a physical attribute of the referents. However, if the same two-handed gesture is accompanied by a sentiment stating that 'one entity is more important than the other', then the gesture is metaphoric because it represents an abstract quality of the referents. When explaining the relational structure of an analogy problem representational gestures can help highlight relations between items within scenes and comparisons across scenes. For example, iconic gestures should be able to illustrate the relations within the individual scenes

(e.g., a sweeping movement between two items in a scene could emphasize that one item is *chas-ing* the other), and two-handed deictic gestures should draw attention to commonalities across scenes by indexing the items in the problem as they are referenced in speech.

Gesture use during instruction supports children's learning across a variety of topics that require relational comparisons, including mathematics (Cook et al., 2013; Singer & Goldin-Meadow, 2005), symmetry (Valenzeno, Alibali, & Klatzky, 2003), and Piagetian conservation (Church, Ayman-Nolley, & Mahootian, 2004; Ping & Goldin-Meadow, 2008). The features of gesture that make it an effective teaching tool in other domains suggest that is has the potential to facilitate analogical reasoning ability. Next, I will review how these features of gesture could be utilized to support children's analogical reasoning.

Gesture should be able to drive attention away from irrelevant components of a comparison and towards relationally similar components during problems of analogy. As discussed previously, young children struggle to identify relational commonalities between problems or contexts. This struggle is often attributed to immature IC that causes them to rely on a default behavior of focusing on featural similarities during comparisons. Presumably, one way to ameliorate the effects of low IC during instruction is to visually direct attention toward relations of interest in a problem and away from irrelevant components. Instructional use of gesture has been shown to do just that: Using gesture increases children's ability to learn from instruction by directing attention to spoken referents and facilitating the link between words and what they map onto (Richland, Zur, & Holyoak, 2007; Wakefield, Novack, Congdon, Franconeri, & Goldin-Meadow, 2018b). Gesture, by its nature, grounds speech in the physical environment, and is, therefore, well-suited to index spoken words to relevant parts of the visual context (Glenberg &

Robertson, 1999). Because gesture can link ideas *and* embody the relationship between those ideas, when explaining how to solve an analogy, gesture should draw attention to the individual contexts *and* the relationship between two contexts at the same time.

Gesture should be able to facilitate comprehension of spoken analogical reasoning instruction. Children's difficulty identifying relational similarities is not only a result of their limited IC, but they may also struggle to understand the speech that accompanies analogical reasoning instruction. Therefore, an effective instructional tool should facilitate comprehension of spoken instruction. In the case of analogical reasoning, instruction may be complex and ambiguous at times. For example, spoken instruction about a scene analogy may be unclear as to which scene is being discussed at each time point, or which of the two featurally similar items are incorporated in the relation of interest (i.e., which of the two tigers). There is ample evidence that gestures are beneficial for children's language comprehension (Hostetter, 2011; Kendon, 1994), particularly when the spoken message is ambiguous (Thompson & Massaro, 1986) or complex (McNeil, Alibali, & Evans, 2000). The *Integrated-Systems Hypothesis* suggests that because gesture and speech form an integrated system of meaning during language production (Kendon, 1986; McNeill, 1992), gesture can serve to enhance language comprehension (Kelly, Ozyürek, & Maris, 2010). In a classroom setting, several studies have demonstrated that the inclusion of gesture in instruction promotes deeper learning in students than lessons without gesture (Alibali et al., 2013; Richland et al., 2007). For example, Wakefield and colleagues (2018b) found that by incorporating gesture during mathematical equivalence (e.g., 5+3+2 = _+2) instruction children were better able to synchronize their visual attention with the spoken instruction, and that this visual attention pattern was predictive of learning. They suggest that increased synchronization

of visual attention is evidence of gesture's ability to disambiguate which side of the equals sign is being spoken about throughout instruction, and this ability of gesture to disambiguate was the key to gesture's impact on children's learning. In the case of analogical reasoning, gesture may serve a similar role in disambiguating spoken instruction.

Together, gesture's ability to direct visual attention and disambiguate spoken instruction should mitigate the limitations of children's immature IC during tasks of analogy. Because children's immature IC causes them to focus on irrelevant components of an analogy (i.e., featurally similar components) rather than relationally similar components, gesture should alleviate these IC limitations by orienting attention to the relational structure of an analogy and clarifying which components of an analogy are relevant to solving the problem.

Although there is evidence to suggest that gesture will have an impact on children's IC during analogical reasoning tasks, it may also have an effect on their WM capacity during solving. By grounding speaker's words in the external environment (Goldin-Meadow, 2015), gesture has been shown to mitigate limitations of WM by off-loading aspects of cognitive processing to the physical environment (Alibali & Nathan, 2007; Goldin-Meadow, Nusbaum, Kelly, & Wagner, 2001). While this has been demonstrated in numerous domains, it may also be the case during analogical reasoning. Given that prior work suggests children's IC serves as a gating mechanism for their WM, and if gesture does facilitate IC, then gesture instruction should also increase children's WM capacity during tasks of analogical reasoning. By conveying meaning differently than speech and making clear connections between spoken instruction and a physical analogy problem, gesture could maximize limited processing capacity, which is necessary for children to reason analogically about multiple contexts at one time.

While children's analogical reasoning *should* benefit from instructional gesture, individual children may benefit differently depending on their cognitive profile. That is, if gesture is argued here to support children's IC, and perhaps their WM, their degree of existing IC and WM may play a factor in whether gesture instruction benefits their analogical understanding. Previous work has found that in other domains individual differences across learners impacts the benefits of gesture. For example, Wakefield and James (2015) taught children the concept of a palindrome (i.e., a word that reads the same forward and backward) through speech-alone or speech+gesture instruction. Children with high phonological ability benefitted more from speech+gesture instruction than speech-alone instruction, but children with low phonological ability did not show this advantage, suggesting that children need some degree of pre-existing knowledge within the domain to utilize gesture. Similarly, Congdon and colleagues (2018) found that having rudimentary skills on the mathematical concept of measurement units was necessary in order to benefit from gesture instruction. Therefore, while gesture may serve as an effective instructional tool for analogical reasoning, its impact may depend children's current cognitive development.

Broadening Our Understanding of When Gesture is Beneficial

Beyond asking how gesture may affect children's analogical reasoning depending on their current cognitive abilities, there are other questions to consider when deciding the best way to employ gesture during instruction. For example, is the effectiveness of teacher's gestures dependent on the learning context, such as the topic of instruction or the characteristics of the learner? Does the relation between the semantic content of gesture and speech matter? Overall, the consensus in the literature is that gesture facilitates learning. However, *how* this instruction is

administered and *to whom* does seem to matter: The impact of gesture instruction can depend on a number of factors, including task type (Driskell & Radtke, 2003), level of difficulty of the task (McNeil et al., 2000), or even population characteristics of the learners, such as their age or primary language (McNeil et al., 2000; Sueyoshi & Hardison, 2005). Two previous meta-analyses have explored *when* gesture is beneficial for comprehension and communication – that is, whether gesture's effects depend on characteristics of the context in which the gesture is provided. There are not only discrepancies across these two meta-analyses in terms of what factors moderator the impact of gesture, but additional factors should be explored to comprehensively understand *when* gesture is most advantageous as an instruction tool. By understanding the impact of gesture on learning more broadly through a meta-analysis that aims to resolve previous discrepancies and address open-questions in the field, this work can further inform on best practices for teaching analogical reasoning using gesture.

Overview of Present Studies

I have presented evidence that gesture may be one means to facilitate an important foundational ability: analogical reasoning. While gesture should support the types of skills that are necessary for analogical reasoning, the impact of gesture may depend on a number of factors, including children's age and current cognitive profile (i.e., their current knowledge state and cognitive abilities). Together, Studies 1 and 2 examine whether we can support children's analogical reasoning ability using gesture instruction and address why it may or may not have the anticipated advantages for all children. To understand *if* gesture can support children's analogical reasoning, we first need to better understand *why* children struggle with this form of reasoning. In Study 1, we delve deeper into the role of IC in children's analogical reasoning ability. Specifi-

cally, we ask how IC relates to children's orientation of visual attention, measured via eye tracking, while solving scene analogies. This will help researchers understand how limitations of IC contribute to children's difficulty with these tasks. By first establishing how IC, visual attention, and analogical reasoning interact during problem solving, we can then ask *how* gesture instruction impacts children's analogical reasoning as compared to instruction without gesture components. It is possible that gesture instruction produces behavioral changes in this ability through impacting children's visual attention.

To further understand how gesture impacts children's analogical reasoning, we need to ask *when* gesture is beneficial. In Study 2, we ask *when* in terms of children's developmental trajectory is gesture more or less beneficial compared to speech alone instruction. We have presented evidence that analogical reasoning is characterized by a protracted developmental trajectory and gesture's effects are nuanced. In the case of analogical reasoning, the impact of gesture may similarly depend on children's cognitive profile, where all children may not benefit from gesture equally. That is, gesture's ability to promote analogical reasoning may depend on children's current point in development. Here, we treat children's age as a proxy for their current cognitive profile, given that previous work has established relations between the relevant cognitive abilities (IC and WM), age, and analogical reasoning performance (e.g., Richland et al., 2006; Simms et al., 2018; Thibaut & French, 2016). Beyond behavioral measures, we use eye tracking to assess the impact of instruction on children's visual attention. Therefore, Study 2 builds from Study 1 by addressing the role of children's current cognitive profile on the impact of gesture instruction by assessing the relation between children's age, ability to learn from gesture, and the impact of instruction on visual attention.

Finally, to more comprehensively understand the factors that impact gesture's effect on learning, we conduct a meta-analysis as Study 3. Previous work has begun to look across the gesture-for-learning literature and ask when gesture is best utilized for communication and comprehension (Kendon, 1994; Hostetter, 2011; Dargue, Sweller, & Jones, 2019). Study 3 explores this question further by (1) aiming to resolve discrepancies found across previous meta-analyses, (2) consider additional theoretically important questions regarding situational or contextual factors that may impact when gesture is most beneficial for learning, and (3) attempt to replicate key findings of previous meta-analyses. By replicating findings using a sample that includes studies in both previous meta-analyses and those not included in previous work, we can provide further support for the role of particular methodological variations on the effect of gesture. Therefore, the aim of Study 3 is to add to the existing line of research, including both Studies 1 and 2 proposed here and decades of pre-existing literature by exploring under what conditions gesture supports learning across domains, including analogical reasoning.

In sum, because analogical reasoning ability is important for a host of skills and more generally academic success, identifying educational tools to strengthen early analogical reasoning can set children on the path to success in the future. However, while gesture should support the cognitive profile necessary for the development of children's analogical reasoning, the benefits of gesture may not be uniform – all children may not benefit equally. Across these studies we expect that gesture will have differential effects depending on a number of factors, including children's age, cognitive profile, and visual attention patterns. These differential effects will add to the gesture-for-learning literature. And by exploring discrepancies across the field using a me-

ta-analytic approach we can address more broadly under what conditions gesture is most suitable for instruction in the classroom.

CHAPTER TWO

EXPLORING HOW VISUAL ATTENTION, INHIBITORY CONTROL, AND CO-SPEECH GESTURE INSTRUCTION CONTRIBUTE TO CHILDREN'S ANALOGICAL REASONING ABILITY

Introduction

We have presented evidence that the mechanisms by which gesture supports learning in other domains suggests that it may be a useful tool for supporting children's analogical reasoning ability. Here, we examine whether incorporating gesture during spoken instruction can support analogical reasoning ability above and beyond spoken instruction alone. To understand *if* gesture can support children's analogical reasoning, we need to further understand *why* children struggle with reasoning analogically, and then ask *if* and *how* gesture is beneficial.

As discussed in the Introduction, analogical reasoning requires an individual to look beyond featural similarities between elements in two contexts and notice the shared relational structures between the contexts. But, even when told explicitly to look for relational similarities to solve problems, young children tend to focus on featural similarities. We see evidence for this in both scene analogy and propositional analogy tasks (Richland et al., 2006; Starr et al., 2018; Thibaut & French, 2016; Thibaut et al., 2010) and researchers suggest this is driven, in large part, by insufficient IC (Richland et al., 2006; Thibaut & French, 2016).

Whereas previous work has implicated IC as an important contributor to mature analogical reasoning ability, there are still many unknowns about how IC relates to other behavioral

measures which are relevant for problem-solving, and whether children can benefit from directed instruction despite immature IC. One such measure is visual attention (i.e., the orientation of one's foveal visual attention at a given timepoint). Whereas researchers have measured children's visual attention during analogical reasoning (e.g., Gordon & Moser, 2007; Thibaut & French, 2016) and found that allocation of visual attention impacts problem solving (e.g., Starr et al., 2018), no one has considered the relation between visual attention and IC directly in the context of analogical reasoning. In the present study, we consider: (1) whether IC might help children correctly solve analogical reasoning problems by promoting visual attention away from featural distractors, and (2) whether analogical reasoning ability could be facilitated through instruction using gesture, a tool that has been shown to promote learning in a variety of domains (e.g., Congdon, Kwon, & Levine, 2018; Cook, Mitchell, & Goldin-Meadow, 2008; Novack, Congdon, Hemani-Lopez, & Goldin-Meadow, 2014; Ping & Goldin-Meadow, 2008; Valenzeno et al., 2003; Wakefield et al., 2018a; Wakefield & James, 2015).

The Role of Visual Attention in Analogical Reasoning

In previous work, researchers have established that children and adults who are successful when solving analogical reasoning problems tend to show different patterns of visual attention than children who unsuccessfully choose the featural distractor. For example, propositional analogies in the format of A:B::C:D require individuals to identify the relation between A and B, and then choose 'D' from an array of options, such that C and D relate to each other in the same way that A and B relate to each other. When successfully solving these analogies, children and adults tend to initially direct their visual attention to the A:B pair before attending to the C item and response options (Starr et al., 2018; Thibaut & French, 2016). When unsuccessfully solving

A:B::C:D analogies, children tend to show a lack of attention to relations within analogies and a constrained focus on the C item – scanning between the C item and response options.

When considering visual attention to scene analogy problems, Guarino and colleagues (2019) also found differences in how adults and children allocate their attention: 4-5 year-old children attend more to the featural match than adults. Taken together, we see different visual attention patterns in adults who successfully solve problems of analogy versus children who fail to correctly solve problems of analogy, although the specific differences in attention are affected by problem format. While orienting attention away from featural matches is important for successfully solving either form of analogy, attention to the source relation depends on the format of the problem: Attention to the source relation in a propositional analogy (A:B pair) predicts children's performance (Starr et al., 2018), yet attention to the source relation in a scene analogy does not predict performance (Guarino, Wakefield, Morrison, & Richland, 2019).

What might the differences in visual attention to propositional and scene analogy problems mean? We suggest that in each case, the patterns of visual attention align with what we might expect from children with higher versus lower IC, and thus, looking patterns may be driven by this important predictor of analogical reasoning ability. In the case of propositional analogies, children with higher IC theoretically would be able to inhibit attention away from the immediate task goal (i.e., 'Find what goes with C') allowing them to first assess the relation between the A:B pair before reviewing response options. In contrast, children with lower IC who do not inhibit that task goal might instead jump to the C item and make choices based on featural similarities between response options and the C item (i.e., choosing a featural distractor because it is featurally similar to the C item), rather than consider the meaningful relation between the A:B pair. Indeed, children who have difficulty inhibiting the immediate task goal focus more of

their visual attention on the featural distractor compared to other response options (Starr et al., 2018). Similarly, when solving a scene analogy, children with lower IC will likely be unable to inhibit attention away from the item in the target scene that is featurally similar to the prompted item, in favor of focusing on relationally similar items in the problem.

In the present study, we will *directly* compare the relation between IC and visual attention during analogical reasoning to test this relationship using scene analogy problems. To measure IC, we will use a task appropriate for children, that has been used in previous research on analogical reasoning ability (Simms et al., 2018), the Eriksen flanker task, based on the Attention Network Task (Rueda et al., 2004). In analogical reasoning tasks, children must ignore featural similarities in favor of relational similarities. In other words, they must focus their attention on task-relevant features, and inhibit other features. The flanker task also requires children to focus their attention on task-relevant features, and inhibit other features in order to make the appropriate behavioral response. Specifically, in this version of the task they must determine the direction a center fish is 'swimming' (i.e., facing), embedded within a line of fishes. Whereas congruent trials show all fish swimming in the same direction, on incongruent trials, the fish surrounding the center fish are swimming in the opposite direction of the center fish. Thus, children must ignore the more salient information of the direction most fish are swimming, in favor of the center fish. While recent work suggests that children's performance on the flanker task does not relate to their ability to solve analogical reasoning problems, Simms and colleagues (2018) suggested that this might have been because a wide age-range had been considered. The researchers suggest that at a younger age – the age we use in the current study – this task may be a good predictor of analogical reasoning ability.

Gesture Instruction May be One Way to Support Analogical Reasoning Ability

In addition to considering the relation between a child's IC and visual attention during scene analogy problem solving, we ask whether co-speech gestures, movements of the hands that are produced simultaneously with speech, represent information (McNeill, 1992), and direct visual attention (e.g., Rohlfing, Longo, & Bertenthal, 2012; Wakefield et al., 2018b), can serve as a tool to support children's analogical reasoning through facilitating effective looking patterns (i.e., directing visual attention away from featural distractors and towards important, relational comparisons). Decades of research suggest that gesture helps children learn new concepts in a variety of domains, from mathematical principles (e.g., Cook et al., 2013) to word learning (e.g., Wakefield et al., 2018a). Although the mechanisms by which gesture facilitates learning are still being understood, there is evidence that gesture can organize children's visual attention (Wakefield et al., 2018b): Gesture use by an instructor increases the likelihood that children will focus on the important aspects of a problem or material being taught, and better align their visual attention with referents in a teacher's spoken instruction, which is predictive of learning gains (Wakefield et al., 2018b). In the case of an analogical reasoning problem, gesture could be used to call attention to and link related items as they are mentioned in spoken instruction. The benefits of this could be two-fold: First, promoting comparison between exemplars by visually aligning them through gesture could encourage children to attend to the relational or structural similarities shared among them (Namy & Gentner, 2002). Second, by drawing attention to the abstract relations in the source and target scene, this necessarily draws attention *away* from the featural distractor.

Present Study

In the present study, we expand on previous literature by considering how children's visual attention when solving scene analogy problems may be related to their IC, and whether gesture may be a useful tool for supporting successful problem solving, potentially through its ability to direct visual attention. To address these questions, we use a pretest-training-posttest design and monitor children's visual attention using eye tracking while they solve problems and receive instruction through speech alone or speech and gesture, and measure their IC using the Eriksen flanker task.

We hypothesize that individual differences in IC, as measured by the Eriksen flanker task, will be related to both behavioral performance and visual attention patterns as measured using eye tracking before children have received instruction. By measuring visual attention via eye tracking we can assess how children orient their foveal attention while solving scene analogies. Specifically, at pretest, the higher a child's IC, the more likely they should be to choose the correct, relational choice and the less likely they should be to choose the featural distractor. Further, the higher a child's IC, the less likely they should be to visually attend to the featural distractor. Based on previous literature from other domains, we anticipate that children will benefit more from instruction that includes gesture, than from instruction through speech alone, and that this may be driven by visual attention patterns. Gesture instruction may scaffold understanding of the relations through helping children avoid the featural matches and directing attention towards relational information in the problem.

Method

Participants

Fifty-seven 4- and 5-year-old children participated in the present study (29 females, M_{age} = 4;11 mo, SD_{age} = 5.6 mo)[1]. Children were randomly assigned to one of two instruction conditions: speech-alone *(n*=28) or speech+gesture (*n*=29). Two children in the speech+gesture condition were excluded from all analyses due to a lack of behavioral response. Participants represented a diverse sample from a large metropolitan city (48% White, 14% Black, 5% More than one race, 5% Asian, 28% Unreported). Informed consent was obtained from a parent or guardian of each participant, and verbal assent was obtained from children. Children were compensated with stickers and a certificate noting their participation in a research study. Additionally, study sites were compensated with a gift card to purchase classroom materials. Children participated individually in one experimental session at their school during a regular school day.

Materials

Pretest/posttest stimuli. Twenty-four scene analogy problems were created, based on the structure used by Richland and colleagues (2006). Each problem included a pair of scenes, a source scene on the left, and a target scene on the right. Scenes depicted one of two *relation categories* (i.e., chasing or reading) occurring between items (i.e., animals or people). Source scenes contained five items: the two items within the relation of chasing or reading, and three additional items (i.e., neutral inanimate objects that were not involved in the relation of interest). One of the items within the source scene relation was circled. Target scenes also contained five items: the

[1] The sample size was limited due to the onset of the COVID-19 pandemic. While we are not able to collect more data for this study, we acknowledge that a larger sample would benefit the study. Working with this constraint, we have attempted to gain power in our analyses by conducting analyses at the trial level, while accounting for random effects of participants. However, we do acknowledge that replication of this study with a larger sample is needed.

two items within the relation, two additional items, and a featural distractor. The featural distractor was similar to the circled source-scene item and centrally located, increasing the likelihood that the item would draw participants' attention.

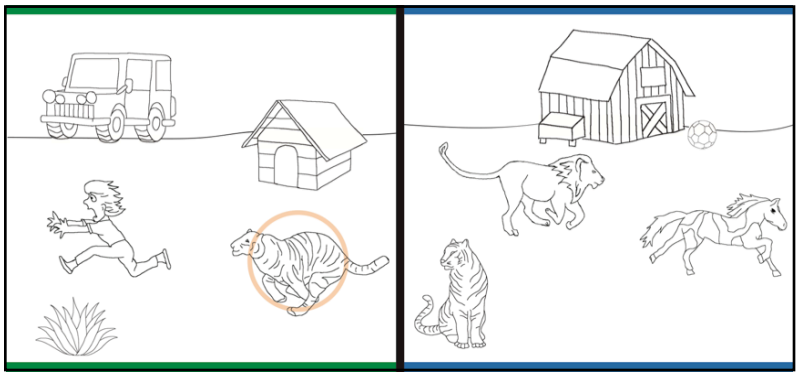
Figure 2A shows an example of a chasing *source* and *target* scene. The source scene on the left shows a tiger chasing a woman (items within the chasing relation), and a dog-house, jeep and plant (neutral items). The corresponding target scene on the right shows a lion chasing a horse (items within the chasing relation), a barn and soccer ball (neutral items), and a tiger (a distractor, featurally similar to the prompted tiger in the source scene). Figure 2B shows and example of a reading *source* and *target* scene. The source scene on the left shows an elephant reading to a rabbit (items within the reading relation), and a tree, bench, and see-saw (neutral items). The corresponding target scene on the right shows a bird reading to a frog (items within the reading relation), a floatie and tent (neutral items), and a rabbit (a distractor, featurally similar to the prompted rabbit in the source scene). Stimuli were displayed on a 15-inch Dell laptop.

Figure 2.

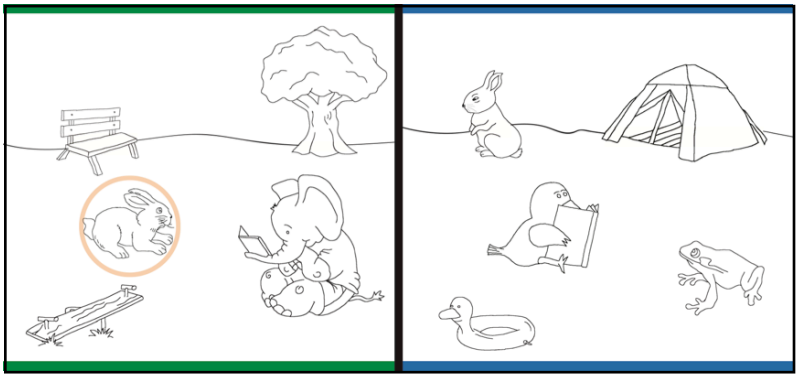
The directionality of relations within a pair of scenes was reversed to avoid children making choices based on spatial location alone. For example, in Figure 2A, the direction of chasing is right to left in the source scene (the tiger on the right is chasing the boy on the left), whereas the direction of chasing is left to right in the target scene (the lion on the left is chasing the horse on the right).

Figure 2. A. Example trial of chasing relation category. B. Example trial of reading relation category.

A.



B.



Instruction stimuli. Four speech-alone and four speech+gesture instructional videos about how to solve scene analogies were created. All videos depicted chasing relations, but unlike the pre- and posttest problems, no item was circled. For both conditions, scenes were accompanied by spoken instruction highlighting the relations between and within scenes. For example, "*See, the boy is chasing the girl, and the dog is chasing the cat. This means the boy is in the same part of the pattern as the dog because they are both chasing, and the girl is in the same part of the pattern as the cat because they are both being chased*". In addition to spoken instruction, gestures were used to emphasize items and relations for children in the speech+gesture condition (see Figure 3). In the previous example, when the instructor said '*The boy is chasing the*

girl,', a sweeping movement of the index finger traced a path from the boy to the girl, highlighting the chasing relation. The same sweeping gesture was used when the instructor said '*... and the dog was chasing the cat*'. Then, deictic gestures – pointing gestures used to indicate objects or locations – were used to simultaneously reference the items that were in the same parts of the relations. Items were indicated by a pointed index finger on each hand. When the instructor said, '*This means the boy is in the same part of the pattern as the dog because they are both chasing*', simultaneous deictic gestures, made with the pointer fingers on each hand, pointed to the boy and the dog. Similarly, when the instructor said, '*…and the girl is in the same part of the pattern as the cat because they are both being chased*', simultaneous deictic gestures pointed to the girl and the cat. Videos showed up to the entire hand depending on the item's location.

To control for potential differences in inflection, the same audio track was used for the speech-alone and speech+gesture versions of each video. Videos were approximately 29 seconds.

Flanker inhibitory control and attention test. An NIH Toolbox version of the Flanker Inhibitory Control and Attention Test for ages 3-7 was administered to all participants. This task was a child-friendly version of a typical Erikson Flanker task, such that the children were asked to indicate which direction a cartoon fish was swimming. On some trials, a row of fish would be oriented in the same direction, including the target fish in the middle, signifying swimming in the same direction. On other trials, all fish except the target fish would be oriented in the same direction. The latter incongruent trials were the measure of IC. Inhibitory control score is calculated from the output provided by the NIH toolbox software. Scores range from 0-10, where higher scores indicate mature IC.

Eye tracker. Eye tracking data were collected via corneal reflection using a Tobii X3-120 remote eye tracker mounted at the base of a 15-inch Dell laptop screen. Tobii software was

used to perform a 5-point calibration procedure using standard animation blue dots. This step was followed by the collection and integration of gaze data with the presented instructional videos (described below) using Tobii Studio (Tobii Technology, Sweden). Data were extracted on the level of individual fixations as defined by Tobii Studio software—an algorithm determines if two points of gaze data are within a preset minimum distance from one another for a minimum of 100 msec, allowing for the exclusion of eye position information during saccades. After extraction, fixation location was queried at 8.33 msec intervals, to align with the native sampling frequency of the eye tracker (120 Hz).

Procedure

Children participated individually in a quiet setting at their school. Children were told they were going to play a picture game and completed a warm-up trial, which oriented them to the layout of test trials (i.e., two pictures with different colored borders; The terms left and right were not used because of the age of the children) and their task: For each set of scenes, their job was to find the pattern in the pictures. During the warm-up trial, the experimenter described the chasing relation, using language similar to the instructional videos, and asked the child to solve the relation. The explanation was repeated until the child chose the correct item. This introduction ensured that when children incorrectly answered a trial, it was not because they misunderstood the task. After completing the warm-up, children were seated approximately 40 cm in front of the laptop. Their position was calibrated and adjusted if necessary, and they were asked to remain as still as possible during the rest of the game, during which eye tracking data were collected.

Next, children completed a pretest. All children saw 12 pairs of scenes in a randomized order: six depicting the chasing relation, and six depicting the reading relation. For each pair of

scenes, an item in the source scene was circled and children were asked to identify the related item in the target scene: "*Which thing in the picture with the blue edges is in the same part of the pattern as the circled thing in the picture with the green edges?*". Responses were recorded for each trial and children were re-prompted if they did not respond.

After completing the pretest, children were asked to pay attention to instructional videos. Children were randomly assigned to watch four videos with speech-alone or speech+gesture instruction. In each video, children saw a source and target scene that each contained items in a chasing relation and neutral items. The target scene also contained a featural distractor. In both conditions, they heard spoken instruction explaining the relations within and between the scenes. The spoken instruction was accompanied by gesture in the speech+gesture condition videos (see Materials for additional details). Following each video, children were presented with the scene-pair from the video with a source-scene item prompted, and asked to find the item in the picture with the blue edges that is in the same part of the pattern as the circled item. If an incorrect response was given, children were re-prompted until they selected the correct relational match to help emphasize the correct way to solve the scene analogies. Prompts to solve the instructional trials were only provided verbally for both conditions.

A posttest was administered after children watched the instructional videos. Children completed another set of 12 scene analogy problems (six chasing; six reading), identical in format to the pretest. Finally, children completed the NIH Toolbox Flanker Inhibitory Control and Attention Test for ages 3-7 to assess their IC.

Each part of the task was administered immediately following the previous part of the task. The entire procedure lasted 25-35 minutes (~10 minutes to familiarize the child with the

eye tracker, for calibration, and to introduce the task, ~10 min for pretest, ~5 minute for instruction, and ~10 minutes for posttest).

Data Selection for Eye Tracking Analyses

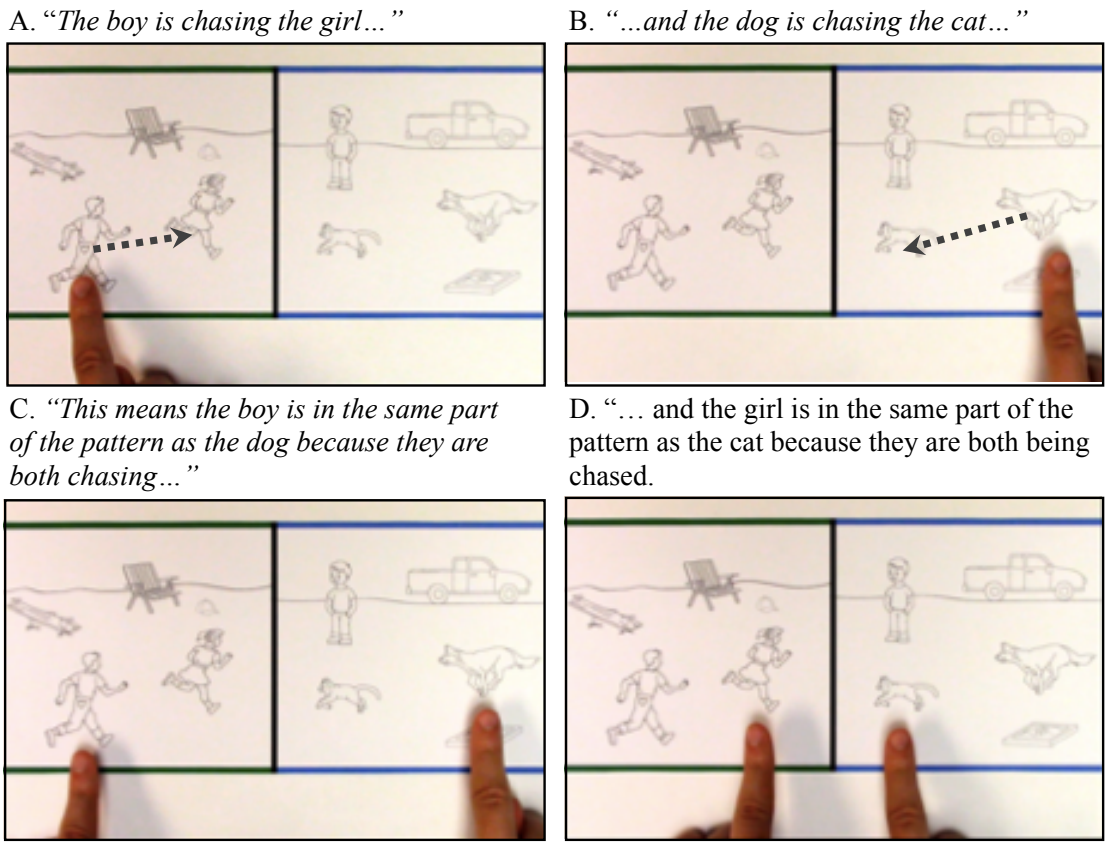
To address how visual attention related to children's IC and performance at pretest, and how visual attention was impacted by instruction at training and posttest, we considered the proportion of time spent looking at items within the scenes, and the degree to which children follow along with spoken instruction. Trials on which insufficient eye tracking data were collected (<65% tracking during a trial) were excluded from analyses. If a child had insufficient tracking on greater than 75% of the trials at a given timepoint (pretest, training, posttest), that child was excluded from analyses involving that timepoint of interest. This resulted in one child being excluded from analyses in which pretest was a timepoint of interest ($n_{Speech\text{-}Alone} = 1$), and three children being excluded from analyses in which posttest was a timepoint of interest ($n_{Speech\text{-}Alone} = 1$; $n_{Speech+Gesture} = 2$).

Fixation to the featural distractor. For pretest, training, and posttest trials, areas of interest (AOIs) were generated for each of the 10 items within the scene pairs using Tobii Studio, with a particular interest in how children visually attend to the featural distractor because previous research finds that children struggle to direct their attention away from featural matches within analogies (Guarino et al., 2019) and often choose the featural match rather than the correct, relational choice (e.g., Richland et al., 2006). The remaining spaces outside of these AOIs were collapsed into an "other" AOI. To facilitate comparisons to previous work using other analogy formats (French, Glady, & Thibaut, 2017; French & Thibaut, 2014; Gordon & Moser, 2007), proportion of time spent looking to each AOI was calculated by dividing the time looking to an AOI during the *latter third of solving time* by the total time looking during the latter third

of solving time. We examined this portion of solving time, rather than initial solving time, because all children will likely attend to the source relation in the source scene to some degree initially given the saliency of the circled item, but the degree to which they continue to look to the featural match will impact their performance (Guarino et al., 2019). In contrast, during instruction, proportion of time spent looking to AOIs was calculated by dividing the time looking to an AOI throughout an *entire trial* by the total time looking during the entire trial. Proportion of time spent looking to AOIs were used as dependent measures in our analyses.

Following along during instruction. Because previous work suggests that gesture can help children follow along with spoken instruction and that this is predictive of learning (Wakefield et al., 2018b), we also calculated a 'following score' for each training trial. Following scores were calculated by creating four time segments in which different relational comparisons were made by the instructor (see Figure 3) and assessing whether children looked to AOIs highlighted in speech during each segment (i.e., during a given segment, children received a score of '1' if they looked to the relevant AOIs as they were labeled in speech and a '0' if they did not). Children could receive a score of 0 to 4 on each training trial. Scores calculated for each instructional trial were used in analyses.

Figure 3. Example instructional trial from a speech+gesture trial, with corresponding verbal and gesture components.



Note. Figure shows video stills from speech+gesture condition, and accompanying speech. Dotted arrows show the path the pointer finger took to show the chasing relation. Children heard identical speech, but did not see gesture, in the speech-alone condition.

Results

All analyses were conducted using R Studio (version 1.1.456), supported by R version 3.6.0. Analyses relied on the *lme4* package, which allows for mixed effects modeling (Bates, Mächler, Bolker, & Walker, 2015). When running mixed effects models through *lme4*, dummy coding was used, the default option for coding in this package. Appropriate reference levels for factors were assigned before each model was run: When condition was included in the models, the speech-alone condition was set as the baseline and compared against the speech+gesture

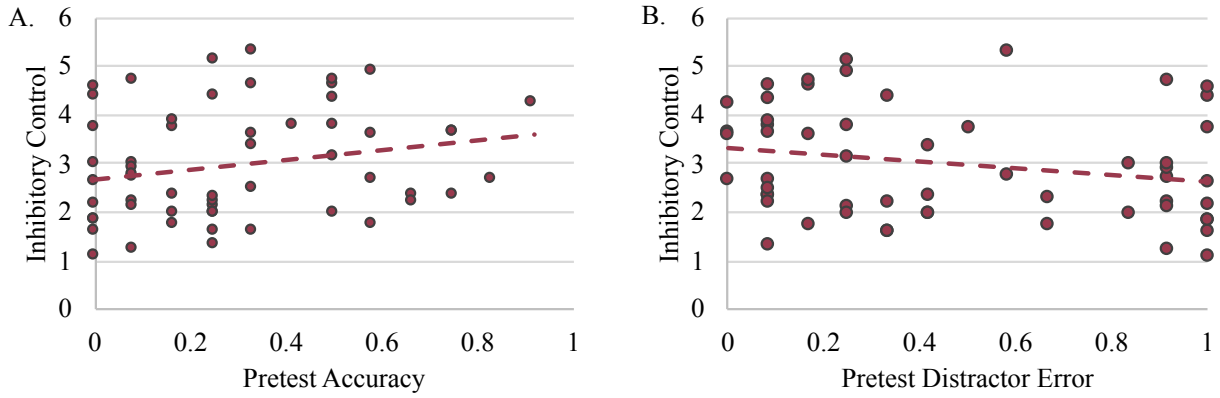
condition. Because trials at pre- and posttest included two types of relations (reading and chasing), trial type was included as a predictor in all pre- and posttest models to account for any difference in performance between the two types of relations. Participant was included as a random factor in all models. An alpha level of .05 was used when evaluating statistical significance.

Do Inhibitory Control and Visual Attention Predict Behavioral Performance?

Before considering our main questions of interest, we asked whether we could replicate previous findings showing that 4- and 5-year-old children perform poorly on scene analogy problems: They often choose the featural distractor rather than the correct relational choice (Richland et al., 2006), and this is associated with an inability to visually ignore the featural distractor (Guarino et al., 2019) and underdeveloped IC (Doumas et al., 2018; Morrison et al., 2004; Morrison et al., 2011; Viskontas et al., 2004). As expected, children performed poorly at pretest, correctly answering 3.54 trials (out of 12; SD = 3.04) and choosing the featural distractor on nearly half of the trials (M = 5.72 out of 12 trials, SD = 4.49), allocated 26% of their attention in the final third of the trial to the visual distractor, and showed variable but immature IC, scoring between 1.13 and 5.31 on the Eriksen Flanker task (scale range: 0 to 10).

To test whether there was a significant relation between children's likelihood to choose the correct choice or the incorrect featural distractor at pretest, and their IC and attention to the featural distractor, we constructed two binomial logistic regression models. In the first model, trial-level accuracy (0,1) served as the dependent measure with proportion looking to the distractor on a given trial and IC score as fixed factors. We found a main effect of proportion looking to the distractor, such that looking to the distractor negatively predicted accuracy (β = -6.48, SE = 0.80, t = -8.06, $p < .001$), but found no main effect of IC on accuracy (β = 0.21, SE = 0.15, t = 1.37, p = .172, Figure 4A).

Figure 4. A. Relation between inhibitory control and pretest accuracy. B. Relation between inhibitory control and pretest distractor error.



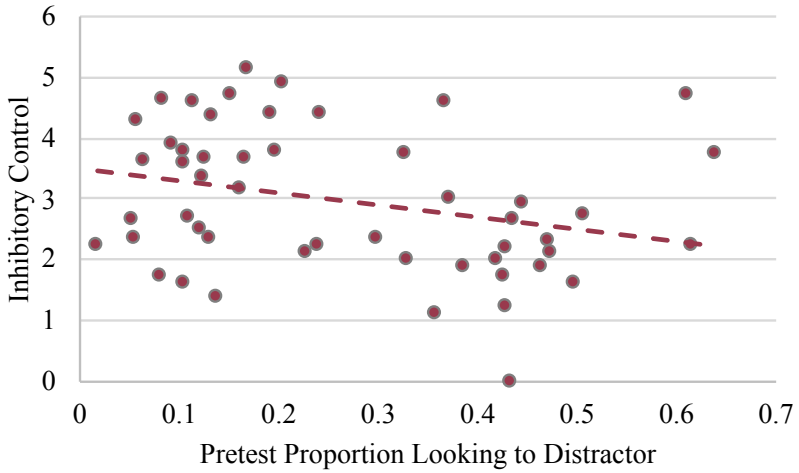
In the second model, the same fixed factors were used but trial-level distractor error (0,1) served as the dependent measure. We found a significant main effect of proportion looking to the distractor, but this time in the opposite direction: Looking to the distractor was positively related to choosing the distractor (β = 14.19, *SE* = 1.57, t = 9.05, $p < .001$). In contrast to the first model, the main effect of IC was also significant, such that the lower a child's IC, the more likely they were to choose the distractor (β = -0.76, *SE* = 0.38, t = -1.99, p = .047, Figure 4B). These findings suggest that looking to the distractor is associated with incorrectly choosing the distractor and failing to make the correct choice, but immature IC contributes more to children's bias toward featural matches than their accuracy.

Does Children's Inhibitory Control Relate to their Visual Attention Patterns?

Given our conceptual replication of prior work demonstrating relations between IC and visual attention with children's analogical reasoning ability, our first main goal was to determine if one reason IC promotes success on analogical reasoning tasks is because it organizes visual attention away from featural distractors during problem solving. We constructed a generalized linear model with proportion of time spent looking to the distractor on each problem as the de-

pendent measure, and children's Flanker IC score as the fixed factor. We found a marginal relation between IC and looking to the distractor, suggesting that the lower children's IC, the greater the proportion of time spent looking towards the distractor (β = -0.04, SE = 0.02, t = -1.87, p = .068; Figure 5). Importantly, a similar model showed that IC did not predict whether children look at the distractor at least once during the entire trial (β = -0.02, SE = 0.01, t = -1.42, p = .162), suggesting that all children 'checked-in' with the distractor, but only those with lower IC continued to fixate on the distractor in the final third of solving time. Although the relation between IC and proportion looking to the featural distractor was trending and did not reach statistical significance, in combination with the finding that simply checking in with the featural distractor was not at all related to IC, we suggest the issue of statistical significance may relate more to statistical power than a lack of relation. Still, we wish to emphasize that this result is *marginal*, and not robust evidence for a relation between IC and amount of looking to the distractor. While this is a relation that makes theoretical sense, it should be replicated in future studies with a larger number of participants.

Figure 5. Relation between inhibitory control and proportion of time spent looking to the distractor during the latter third of solving time.



Can Including Gesture in Instruction Facilitate Children's Analogical Reasoning Ability?

Our second goal was to ask whether gesture – a tool that has been shown to facilitate learning in many domains – can improve children's analogical reasoning ability, particularly by promoting beneficial looking patterns. To do this, we considered (1) whether behavioral performance at posttest differed between conditions, (2) whether visual attention patterns during instruction differed between conditions, (3) whether visual attention patterns during instruction predict posttest behavioral performance, and finally (4) whether the relation between visual attention at posttest changed due to instruction. This set of analyses addresses whether any behavioral changes at posttest, compared to pretest, are due to the effect of instruction on visual attention, either during instruction or following instruction. Before addressing these questions, we confirmed that there were no pretest differences between children who had been randomly assigned to each condition on behavioral performance measures (accuracy: $\beta = 0.22$, $SE = 0.42$, $t = 0.51$, $p = .607$; distractor error: $\beta = -0.38$, $SE = 0.70$, $t = -0.54$, $p = .590$) or proportion looking to the distractor ($\beta = -0.01$, $SE = 0.05$, $t = -0.16$, $p = .874$).

Does behavioral performance at posttest differ by condition? First, we asked if accuracy was impacted by instruction. On average, children in the speech-alone condition correctly answered 5.07 problems (out of 12, $SD = 3.42$) and children in the speech+gesture condition correctly answered 4.48 problems ($SD = 3.16$). We constructed a binomial logistic regression model with accuracy (0,1) as the dependent measure, and condition (speech-alone, speech+gesture), timepoint (pretest, posttest), and two 2-way interactions (condition x timepoint; trial type x timepoint) as fixed effects. IC was not included in models predicting accuracy at posttest because there was not a pre-existing relation between IC and accuracy prior to instruction.

We found no evidence of an interaction between condition and timepoint predicting accuracy (β = -0.42, *SE* = 0.27, *t* = -1.57, *p* = .116), which would have suggested that one condition supported *greater* gains than the other condition, and no main effect of condition (β = 0.22, *SE* = 0.42, *t* = 0.52, *p* = .604). But, we did find evidence that children made performance gains from pretest to posttest: We found a significant interaction between trial type and timepoint (β = 0.93, *SE* = 0.26, *t* = 3.53, *p* < .001). Posthocs revealed that children significantly improved on both reading (β = 0.58, *SE* = 0.27, *t* = 2.17, *p* = 0.029) and chasing (β = 1.03, *SE* = 0.27, *t* = 3.79, *p* < .001) trials across timepoint, suggesting that improvement is seen on both trial types, although this effect was stronger for trials similar to those on which children received instruction. Together, these results suggest that all children learn from instruction, regardless of what type of instruction they received, and that children's performance improves both for trials similar to those used during training and those that require more generalization.

Next, we asked if choice of the featural distractor was impacted by instruction. On average, children in both conditions chose the distractor less at posttest than they did at pretest. Children in the speech-alone condition chose the distractor on 4.18 problems *(SD* = 4.76) compared to 6.07 problems (*SD* = 4.99) at pretest, and children in the speech+gesture condition chose the distractor on 4.59 problems (*SD* = 3.74), compared to 5.38 problems (*SD* = 4.00) at pretest. To test whether children showed significant decreases in distractor error from pretest to posttest, we constructed a similar binomial logistic regression model with trial level distractor error (0,1) as the dependent measure, and condition (speech-alone, speech+gesture), timepoint (pretest, instruction), IC score, and three 2-way interactions (condition x timepoint; IC x timepoint; trial type x timepoint) as fixed factors. Due to the significant relation between IC and choice of distractor prior to instruction, IC was included in models predicting distractor error.

We found a significant main effect of timepoint on distractor choice (β = -2.55, *SE* = 0.54, t = -4.76, $p < .001$), where children were less likely to select the featural distractor on posttest trials than on pretest trials. However, in contrast to the previous model, we *did* find a significant interaction between condition and timepoint (β = 0.93, *SE* = 0.32, t = 2.91, p = .004), with posthocs revealing a marginal effect of timepoint for children in the speech+gesture condition (β = -0.98, *SE* = 0.57, t = -1.72, p = .085) and a significant effect of timepoint for children in the speech-alone condition (β = 0.00, *SE* = 0.00, t = -4.45, $p < .001$). Additionally, in contrast to the model predicting accuracy, there was no significant interaction between trial type and timepoint (β = -0.36, *SE* = 0.30, t = -1.18, p = .236). Within the context of the significant main effect of timepoint, these additional results suggest that children were less likely to choose the featural distractor on both reading and chasing trials at posttest as compared to pretest, but that overall this decrease was marginal for the speech+gesture condition.

When considering the impact of IC on distractor choice, we found a significant interaction between timepoint and IC (β = 0.46, *SE* = 0.14, t = 3.27, p = .001). This interaction suggests that the impact of IC changes when children solve problems before and after instruction. Posthocs revealed that the relation between IC and choice of distractor is only evident at pretest (β = -0.66, *SE* = 0.32, t = -2.06, p = .040) and not at posttest (β = -0.14, *SE* = 0.37, t = -0.37, p = .711), suggesting that instruction weakens the relation between IC and performance that was evident at pretest.

Together, these results suggest that whereas all children show improved accuracy after receiving instruction, the degree to which they improve does not differ based on receiving instruction with only speech or instruction with speech and gesture. However, while both conditions show a decrease in the likelihood of choosing the distractor after receiving instruction, the

degree of change from pre- to posttest is different between conditions: The significant interaction between condition and timepoint predicting distractor choice, in concert with posthoc analyses showing a significant effect of timepoint for the speech-alone condition, but a marginal effect of timepoint for speech+gesture condition, suggests that the decrease in distractor choice from pre- to posttest is *greater* for children in the speech-alone condition than children in the speech+gesture condition. Further, separate analyses show that the impact of IC is weakened by both forms of instruction, such that at posttest there is no longer a significant relation between children's IC and their analogical reasoning ability, suggesting that both types of instruction impact the role of children's existing IC in their ability to solve analogy problems.

Do visual attention patterns during instruction differ between conditions? Although we found no significant differences in behavioral performance between conditions at posttest, we sought to understand whether gesture instruction might drive visual attention differently than speech-alone instruction during instruction, and whether such instruction could change how children subsequently solve and attend to problems at posttest. Specifically, we asked if instruction impacts two visual attention patterns: Proportion of time spent looking to the featural distractor (mirroring the visual attention patterns assessed at pre- and posttest) and ability to synchronize visual attention with the spoken instruction (following score).

To test whether children oriented their visual attention differently between conditions, we constructed two generalized linear models with visual attention measures on each problem (proportion of time spent looking to the distractor, following score) as the dependent measures, and children's IC score, condition, and average pretest accuracy as fixed factors. We found main effects of condition in both models: Children in the speech+gesture condition allocated significantly less of their attention to the distractor ($M = 0.06$, $SD = 0.01$) than children in the speech-alone

condition (M = 0.11, SD = 0.05; β = -0.05, SE = 0.01, t = -5.34, $p < .001$). When considering children's ability to follow along with spoken instruction, children in the speech+gesture condition showed better alignment between their visual attention and spoken instruction than children in the speech-alone condition (speech+gesture: M = 2.81 following score, SD = 0.90); speech-alone: M= 2.02 following score, SD = 0.96; β = 0.82, SE = 0.26, t = 3.15, p = .002). These findings suggest that the inclusion of gesture helped orient children's attention to items as they were referenced in speech and drew attention away from the featural distractor, a visual attention pattern associated with higher IC. Interestingly, although our results at pretest suggest that IC marginally predicted looking to the distractor, there was no evidence of this relation during instruction (β = -0.00, SE = 0.01, t = -0.62, p = .536), suggesting that *if* this pattern did indeed exist during pretest, the presence of instruction was able to direct children's visual attention despite their IC. Finally, when considering relation between children's pretest performance and visual attention measures at instruction, we found that pretest performance predicted proportion of time looking to the distractor, such that children who scored higher on pretest looked less to the distractor (β = -0.04, SE = 0.02, t = -2.75, p = .009); However, pretest performance did not predict following instruction (β = 0.50, SE = 0.45, t = 1.13, p = .266) suggesting that instruction was enough to guide all children's attention to relationally important items in scene analogies.

Does visual attention during instruction predict subsequent performance? Next, we asked if how children orient their visual attention during instruction impacts their posttest behavioral performance. Because we did not find differences between conditions in posttest performance, even though speech+gesture supports more effective visual attention during instruction, we asked if, across both conditions, visual attention during instruction impacts later analogical reasoning ability. We constructed a pair of binomial logistic regression models: one assessing the

relation between looking to the distractor and choice of the distractor, and another assessing the relation between following score and accuracy. First, we constructed a model with trial-level distractor error at posttest (0,1) as the dependent measure and proportion looking to the distractor at instruction, IC score, and average pretest accuracy as fixed factors. We did not find a significant effect of IC (β = 0.15, *SE* = 0.19, *t* = 0.76, *p* = .447), supporting the previously discussed findings that IC no longer impacts posttest performance. And while we found that pretest accuracy related to posttest distractor error (β = -8.62, *SE* = 1.05, *t* = -8.21, *p* < .001), such that children who made fewer errors at pretest also made fewer errors at posttest, importantly, we did not find a relation between looking to the distractor during instruction and posttest distractor error (β = 3.98, *SE* = 4.59, *t* = 0.87, *p* = .386). Using a similar model, we asked whether children's ability to following along with spoken instruction related to their posttest accuracy. While we found a significant effect of posttest accuracy (β = 5.86, *SE* = 0.48, *t* = 12.22, *p* < .001), we did not find an effect of IC (β = -0.01, *SE* = 0.09, *t* = -0.11, *p* = .915) or following score (β = 0.03, *SE* = 0.10, *t* = 0.33, *p* = .742). Together, these findings suggest that while both types of instruction are able to weaken the relation between IC and performance, how children oriented their visual attention during instruction did not directly relate to their analogical reasoning ability during subsequent problem solving. In other words, while speech+gesture instruction can support effective visual attention patterns (i.e., less attention to the featural distractor and effective synchronization of visual attention with spoken instruction), this effect does not extend beyond the instructional session.

Does the relation between visual attention and behavioral performance at posttest change due to instruction? Finally, to further understand whether the relation between visual attention and performance changes due to instruction, we asked whether the relation between

visual attention and behavioral performance evident at pretest changes at posttest. Recall that at pretest visual attention to the featural distractor negatively predicted accuracy and positively predicted choice of the distractor. These relations held at posttest: Using models similar to those used at pretest, we constructed two binomial logistic regression models. In the first model, trial-level accuracy (0,1) served as the dependent measure with proportion looking to the distractor, IC score, and average pretest accuracy were treated as fixed factors. In the second model, we asked if trial-level distractor error (0,1) was predicted by attention to the featural distractor. We found that the relation between attention to the distractor and accuracy seen at pretest remained at posttest (β = -4.38, SE = 0.76, t = -5.74, $p < .001$), such that children who attended more to the featural distractor were less likely to choose the correct, relational item. Similarly, the relation between attention to the distractor and distractor error remained at posttest (β = 12.36, SE = 1.56, t = 7.94, $p < .001$), suggesting that children who attended to the featural distractor more made more errors at posttest. Together, these results suggest that while instruction supports effective visual attention in the moment, how children orient their visual attention still plays a large role in their analogical reasoning ability after instruction.

Discussion

The goals of the current study were to (1) examine how IC might support children's ability to solve scene analogies by directing visual attention away from featural distractors, and (2) determine whether providing instruction with gesture supports successful analogical reasoning, potentially by changing the way children allocate their attention during problem solving. Although marginally significant, our results are suggestive that IC promotes attention away from the featural distractor before instruction. At instruction, we see evidence that instruction with gesture can encourage similar patterns of visual attention: Children who received instruction through

speech+gesture instruction were better able to follow-along with instruction that highlighted relational items, and allocated less attention to the featural distractor than children who received speech-alone instruction. However, these differences in visual attention patterns did not predict posttest performance: Children showed similar performance gains across the two conditions as well as gains in accuracy for both chasing and reading trials, suggesting that children were learning about relations in a general sense, not simply improving the representational strength of the specific chasing relation – suggesting the utility of this type of instruction for improving general, not problem-specific, understanding of analogy. And interestingly, children in the speech-alone condition made significantly fewer choices of the distractor, whereas this decrease was marginal for children in the speech+gesture condition. Finally, while the relation between children's visual attention and performance did not change due to instruction, the relation between children's IC and performance was weakened following both types of instruction, suggesting that instruction was able to overcome the pre-existing association between children's IC and their analogical reasoning ability. We situate these results within the analogical reasoning and gesture learning literature.

Understanding the Relation Between Visual Attention, IC, and Behavioral Performance

As in previous work, our results suggest that visual attention patterns relate to children's ability to solve analogies (e.g., Starr et al., 2018), and that children's low IC hinders their performance (e.g., Richland et al., 2006). We also find support for a prediction made by Simms and colleagues (2018), that the Eriksen Flanker task is related to analogical reasoning ability in young children.

But more importantly, we move beyond considering the relation of behavioral performance with visual attention or IC, and consider how all three of these measures relate to one an-

other. Whereas researchers have inferred that improved reasoning skills *should* be associated with ability to direct attention away from superficial matches (e.g., Richland et al., 2006); Here, we provide direct evidence for this claim. We show that when children are solving analogical reasoning problems before having any instructional support, allocation of attention to the featural distractor is systematically related to analogical reasoning performance and marginally related to our measure of IC. We also find that children's IC is systematically related to their choice of the featural match prior to instruction, which provides direct evidence of a relation between children's IC and analogical reasoning ability.

These results have important implications for future research on the development of analogical reasoning ability. Researchers studying how individual differences impact the development of analogical reasoning have traditionally employed separate behavioral measures of IC like the flanker task, which are collected before or after children complete an analogical reasoning task. From a practical perspective, our work shows that measuring visual attention during problem solving could be used as a more direct proxy for assessing a child's IC than other behavioral measures. While replication of the current findings are warranted, future research could use visual attention as a behavioral measure of children's IC *during* tasks of analogical reasoning. This would be particularly informative prior to any kind of instructional intervention, where IC could serve as an index of children's solving strategy and reasoning ability. Our results are also interesting from a theoretical perspective. Whereas prior work suggests that IC is one contributing component to the development of children's analogical reasoning ability, *how* exactly IC functions during problem solving is unclear. Here, our results provide further support for the argument that visual attention strategies are related to children's IC.

However, when taken as a whole, our results suggest a more complicated story about the relation between IC (as measured by the Eriksen flanker task) and behavioral performance. Our measure of IC correlated with children's choice of the featural distractor at pretest, but this relation was *not* found at posttest. This suggests that either type of instruction, much like relational learning (Richland et al., 2010), can help to compensate for children's lack of IC at critical moments of development when IC is particularly important for determining analogical reasoning ability (e.g., Doumas, et al., 2018). The fact that a short instructional period can disrupt the relation between the flanker task and behavioral performance following instruction may indicate that the flanker task is capturing an aspect of IC that has more to do with cognitive control, rather than avoiding featural lures. The instruction may have acted as a guide to the goal of the task and provided a tool to structure children's thought process at posttest that was not available during pretest.

In contrast, whereas IC was systematically related to choice of the featural distractor at pretest but not at posttest, visual attention *did* remain systematically related to behavioral performance across time points. When children allocated more attention to distractor, they were more likely to choose the distractor. We suggest that these visual attention patterns may reflect the aspect of IC that acts as a gating mechanism and that either a longer intervention or maturation may be required to change these patterns. Additional work must be conducted to further tease apart these different aspects of IC, and how they relate to visual attention and behavioral performance.

Understanding the Impact of Gesture During Problem Solving

In addressing our second question, we found that when viewing instruction that incorporated gesture, children allocated their visual attention differently than when viewing instruction

without gesture. Gesture promoted the mature looking pattern that had been associated with higher IC at pretest – directing attention away from the featural distractor. We also found, in line with previous research on mathematical equivalence instruction (Wakefield et al., 2018b), that gesture instruction supported children's ability to follow along with spoken instruction. That is, children were more likely to attend to the relational items in the scenes as they were referenced in speech if they received speech+gesture instruction, compared to speech-alone instruction. This impact of gesture is likely due to its ability to direct visual attention, but also its ability to facilitate comparison processes. Previous work has demonstrated that gestures can physically embody links between representations, which can serve to disambiguate spoken instruction. For example, Wakefield and colleagues (2018b) found that when the connection between an instructor's words and the physical environment is unclear during mathematical instruction, gesture is able to facilitate the link between words and what they map on to, and thereby, disambiguate spoken referents. Grassmann and Tomasello (2010) made similar conclusions: Pointing gestures help young children disambiguate complex or contradictory verbal instructions in order to retrieve an object. Additionally, previous research suggests that seeing gesture should help to alleviate limitations of processing instruction through an auditory channel alone: Learning is facilitated when information is presented through speech and gesture simultaneously, rather than sequentially (Congdon et al., 2017).

While we find that gesture supports effective visual attention patterns during the instructional session, we did not find that those visual attention patterns predict posttest performance or visual attention patterns, indicating that gesture can encourage looking patterns associated with high IC when it is being used, but that these effects are transient and do not create a lasting change in solving strategies. The finding that speech+gesture instruction facilitates learning to

the same degree as speech-alone instruction is not novel: Previous work within the gesture-for-learning literature has also found that there are contexts in which gesture does not always support learning (e.g., Congdon et al., 2018; Guarino & Wakefield, 2020; Post, Gog, Paas, & Zwaan, 2013; Wakefield & James, 2015). Both the form of the gesture itself (Dargue & Sweller, 2018a, 2018b) and the characteristics of the learner, such as their degree of prior knowledge within a given domain (mathematics: Congdon et al., 2018; grammar: Post et al., 2013; word learning: Wakefield & James, 2015), can impact whether or not gesture supports learning. For example, Post and colleagues (2013) found that when asking children to learn grammatical rules, only children with higher levels of general language skills benefited from the inclusion of gesture during training. They suggested that these findings were due to an expertise reversal effect, such that for children with lower levels of language skills the added cognitive effort needed to process gesture along with verbal instruction actually hurt their ability to learn. This is just one example in a line of work that has demonstrated that gesture's effects can depend on the context in which it is learned, including the characteristics of the learner.

However, these findings do diverge from much of the previous literature which finds that children benefit more from instruction with gesture than instruction with speech alone in a number of domains (e.g., mathematical equivalence: Congdon et al., 2017; Wakefield et al., 2018b, measurement: Congdon et al., 2018, word learning: Wakefield & James, 2015; Wakefield et al., 2018a, spatial-task domains: Chu & Kita, 2008; Valenzeno et al., 2003). Why did children show no advantage when instructed through speech+gesture versus speech-alone, especially given that gesture *did* encourage visual attention patterns that should be helpful for problem solving? The discrepancy between our current work and previous literature may be a consequence of 1) the

domain being taught and 2) how gesture both directs and constrains attention to the problem space, and how this relates to what a child must learn.

When children are instructed with speech+gesture versus speech-alone in mathematical equivalence tasks, children already have some degree of prior experience with arithmetic concepts that sets the stage for mastering math equivalence, such as the ability to add single-digit numbers together. What needs to be learned is the meaning of the equal sign. In other words, children have the basic tools that are necessary to support correct problem solving, but must overcome a fundamental misunderstanding of *how* to solve the problem. Similarly, in word learning studies, young children have already mastered the ability to map new words to referents and are proficient language learners; Gesture is there to help support the new connections between novel words and referents that they are asked to make during the study session. In contrast, when considering analogical reasoning, children have *not* fully acquired the cognitive abilities that are fundamental to solving analogies, struggling with immature IC and WM. They must also work against biases to focus on featural properties when learning (e.g., Murphy et al., 2021). It may be that in this case, gesture is *not* able to give children the extra boost that is often evident in the gesture-for-learning literature. The present study therefore extends the line of research addressing gesture's impact on learners to demonstrate that there are situational factors, such as the learning domain, that can impact whether or not using gesture along with spoken instruction benefits the learner beyond speech alone instruction.

Beyond the domain, we can consider a more nuanced difference in the current learning paradigm and those used in previous studies showing greater gains from gesture: How gesture interacts with what a child must learn is important – or unimportant – when solving a problem. One of the puzzling findings in the current study was that children in both conditions made simi-

lar, significant gains in accuracy from pre- to posttest, but children in the speech-alone condition also made significantly fewer choices of the distractor, whereas this result was in the same direction, but marginal, for children in the speech+gesture condition. Why was gesture not as helpful as speech alone in decreasing the choice of the distractor, when it actually *directed children away from* the featural distractor during instruction? It may be that this feature of gesture – its ability to direct visual attention – is actually a double-edged sword. In directing a child's focus, gesture also *constrains* this focus, which may not allow for necessary exploration of the problem space. In other words, in the context of a scene analogy problem, gesture may interfere with the likelihood a child will become familiar with the featural distractor and its irrelevance for solving a scene analogy problem. This may also explain why following ability did not predict posttest performance. In contrast, if we consider how gesture is used when teaching the concept of mathematical equivalence, what a child has to learn is inherent in the structure of the equation – there is no additional information present that may help a child make sense of equivalence, beyond the components of the equation. Gesture is great at highlighting these problem components, and so the ability to direct and constrain attention has no drawbacks in this learning context. Taken together, we suggest that gesture is great at orienting visual attention, but that that may not always be a positive: For some learning contexts, it may be necessary to explore the problem space in order to fully understand what is and is not relevant for successful problem solving. This possibility could be tested in future studies, asking whether highlighting the distractor with gesture, and incorporating *why* it is not important when solving analogy problems into both spoken and gestured instruction.

Returning to our previous discussion of IC, it may be that either form of instruction can promote better cognitive control for children who are trying to solve analogical reasoning prob-

lems – which would lead to the decoupling of the relation between the Eriksen Flanker measure of IC and behavioral performance at posttest. Beyond the potential impact of instruction on children's cognitive control, it may be that both forms of instruction were able to strengthen children's relational knowledge representations. Previous computational simulations of analogical reasoning development have found that IC is not only related to one's ability to manage distraction, but also their ability to form new relational representations (Doumas et al., 2018; Morrison et al., 2011). Doumas and colleagues (2018) tested a model which finds through simulations that higher IC may allow for increased growth in relational knowledge representations, which supports more effective relational reasoning. Therefore, it may be that both forms of instruction in the present study were able to facilitate building relational representations of the task strategy by reducing IC demands to some extent. While the association between IC and building representations, or rule-learning, is well-established (e.g., Blackwell, Chatham, Wiseheart, & Munakata, 2014; Chevalier & Blaye, 2008; Diamond, Kirkham, & Amso, 2002; Egner & Hirsh, 2005), future work with this paradigm should further investigate whether training knowledge representations reduces children and adults' IC demands during the reasoning process.

However, despite the successful decoupling of IC and analogical reasoning performance, the gating mechanism component of IC is likely more maturational and thereby more difficult to affect. While gesture instruction may be able to direct attention effectively during the training session, perhaps children cannot overcome the maturational limitations of the gating mechanism of IC to sustain these advantageous patterns of visual attention when instructional support is no longer provided at posttest. It may be that children who are older than those in the present study and have that gating mechanism in place, but do not understand what rule should be applied when gating information into WM, may benefit *more* from instruction than those in the current

study. And, it may be that older children *would* show a larger benefit from gesture instruction than speech alone instruction. This would align with previous work that suggests gesture is a more advantageous tool for children with a certain level of prior knowledge or ability related to the to-be-learned concept (Congdon et al., 2018; Wakefield & James, 2015). Looking across a wider age range, future work could consider this possibility.

Conclusion

Because analogical reasoning is important for a host of skills necessary for future academic achievement and career success (e.g., innovation, creativity, inductive reasoning; Gentner, 2010), we aimed to better understand *why* children struggle with this ability and *how* we can support early development of these skills. Specifically, we aimed to better understand how visual attention might relate to IC, a factor that has been posited to contribute to the protracted development of analogical reasoning, and to ask whether gesture could support analogical reasoning ability. Our results extend previously established associations between IC and children's analogical reasoning by identifying a relation between IC and visual attention measures. Surprisingly, we found that whereas gesture is effective for directing visual attention in a beneficial way, children show equal gains in accuracy after speech-alone and speech+gesture instruction, and surprisingly, they show a significant decrease in choice of distractor after speech-alone instruction, whereas this change is marginal following speech+gesture instruction. These findings suggest that our instruction did have an impact on learning, but also raise important questions about the general claim that gesture will boost learning above and beyond speech-alone instruction, and highlight a need to further consider how aspects like domain and the way gesture interacts with what must be learned affect the utility of gesture as a teaching tool. It may be that the benefit of gesture depends on additional factors that were not tested here, such as individual variations

among children (e.g., familiarity with tasks of analogy or degree of relevant cognitive abilities) or situational variations in which the gesture is used (e.g., structure of the task or the task domain). Further, the finding that both analogical reasoning *and* the relation between IC and analogical reasoning ability is malleable to instruction provides insight into the developmental process of relational reasoning, revealing that it is not only driven by either maturation or knowledge acquisition, but also may be impacted by directed socialization.

CHAPTER THREE

TEACHING ANALOGICAL REASONING WITH CO-SPEECH GESTURE SHOWS CHILDREN WHERE TO LOOK, BUT ONLY BOOSTS LEARNING FOR SOME

Introduction

In Study 1 we found that although instruction incorporating speech and gesture facilitates different visual attention patterns than instruction through speech alone, children did not show the expected behavioral pattern. Specifically, while all children showed improvements in performance from pre- to posttest, gesture instruction did not lead to better performance than speech alone instruction. It is possible that these unanticipated findings may be attributed to the individual differences between children, such as their age or cognitive maturity: Previous work shows that children have difficulties solving analogical reasoning problems when they are younger, presumably because of maturational limitations of IC and WM (e.g., Doumas et al., 2018; Simms et al., 2018). It may be that young children cannot overcome these limitations to be successful when instructional support is no longer provided at posttest. Furthermore, while gesture has been found to support learning across a variety of domains, the mechanism by which gesture supports children's learning, and how individual differences between children impacts the effectiveness of incorporating gesture into instruction, are not fully understood. Therefore, the goal of Study 2 is to ask whether the impact of gesture instruction varies depending on children's domain relevant cognitive capacities.

Gesture Supports Learning By Disambiguating the Spoken Instruction

One way gesture is thought to help children learn is by grounding and disambiguating the meaning of spoken instruction (e.g., Alibali & Nathan, 2007; Ping & Goldin-Meadow, 2008). When learning a new concept, children may struggle to understand the meaning of spoken instruction and fail to see connections between a teacher's speech and their use of supportive materials like equations, figures, or diagrams. Gestures facilitate connections between spoken language and these physical supports by directing attention to key components of a problem being taught or providing a visual depiction of an abstract concept through hand shape or movement trajectory (e.g., Altmann & Kamide, 1999; Huettig, Rommers, & Meyer, 2011; McNeill, 1992; Wakefield et al., 2018b). For example, when being taught the concept of mathematical equivalence – the idea that two sides of an equation are equal to one another (e.g., 2+5+8 = _+8) – eye tracking results show that children follow along with spoken instruction more effectively if it is accompanied by gesture than if the concept is explained through speech alone. Importantly, children's ability to follow along with spoken instruction incorporating gesture predicts their ability to correctly solve mathematical equivalence problems beyond instruction (Wakefield et al., 2018b).

However, incorporating gesture may not support all children's understanding of spoken instruction to the same extent. Although prior work suggests that gesture supports children's learning, there are nuances to when gesture is beneficial: Children's pre-existing knowledge related to a domain – which we will refer to as their *cognitive profile* – can impact whether they learn from gesture instruction. For example, Wakefield and James (2015) taught children the concept of a palindrome (i.e., a word that reads the same forward and backward) through speech-alone or speech+gesture instruction. They considered whether the impact of gesture was affected

by children's relevant cognitive profile – in this case, their phonological ability, as the task relied heavily on understanding how sounds in words fit together. Children with high phonological ability benefitted more from speech+gesture instruction than speech-alone instruction, but children with low phonological ability did not show this advantage, suggesting that children need some degree of pre-existing knowledge within the domain to utilize gesture. In this case, the authors argued that gesture could not clarify spoken instruction unless children had a certain level of phonological awareness.

Although not considered by Wakefield and James, there may also be a developmental point when children are on the brink of understanding a concept and have a sufficiently developed cognitive profile that they need just a small boost from instruction to master a concept. In this case, incorporating gesture into instruction might not be any more powerful than spoken instruction alone. There may be a 'sweet spot' where children have enough foundational knowledge and cognitive abilities related to a concept that gesture can clarify spoken instruction and boost their learning, while children far below or above this developmental point do not show an advantage when learning through gesture.

In the present study, to better understand how gesture can support children's understanding of spoken instruction and whether the benefit of teaching though speech and gesture depends on differences in cognitive profile, we explore the impact of gesture in analogical reasoning. For the purpose of the present study, analogical reasoning is a useful testbed because it is a domain that requires disambiguating complex verbal information, and because the relevant cognitive profile for solving analogies shows protracted development across early childhood (e.g., Richland et al., 2006; Starr et al., 2018; Thibaut & French, 2016; Thibaut et al., 2010). When asked to solve

analogy problems, children cannot overcome their tendency to focus on surface features, rather than relational information (Gentner, 1988), until they are 9 or 11 years of age (Richland et al., 2006).

Because incorporating gesture in instruction can direct children's visual attention effectively to key components of a problem in other domains, such as mathematics instruction (Wakefield et al., 2018b), gesture should also be able to facilitate effective visual attention in problems of analogy. As seen in Study 1, this was found to be the case; Gesture was able to orient attention towards the relationally important information in the scene pairs during instruction. However, beyond gesture's ability to direct attention, it should be able to clearly indicate, and disambiguate, which items a teacher is referring to when providing spoken instruction, so that children focus on items and relations relevant for successful solving and do not attend to irrelevant items. When considering a scene analogy, a teacher is likely to align the important relations through speech, stating that the boy is chasing the girl, and the dog is chasing the cat. In theory, this type of statement, which highlights structural similarities between contexts, should orient children's attention to the items involved in the relation of chasing, and, thereby, facilitate an analogical comparison (e.g., Gentner, 1983, 2010; Markman & Gentner, 1993; Namy & Gentner, 2002). However, when a featural match is present, this spoken instruction may leave some ambiguity in terms of *which* boy is being discussed (i.e., the boy in the chasing relation and the featural match). Children may focus their attention on one or both boys, and miss the important connections being drawn between the relations in the source and target scenes. Indeed, we know from eye tracking studies that children who incorrectly solve analogical reasoning problems tend to focus their visual attention on the featural match, and ignore relational information (Glady,

French, & Thibaut, 2017; Guarino & Wakefield., 2020; Thibaut & French, 2016; Starr et al., 2018). Instruction that incorporates gesture may help young children understand which boy is relevant to the task and direct their attention away from irrelevant featural matches.

Gesture's Effect May Depend on the Learner's Cognitive Profile

But will gesture instruction provide the same boost to all children who struggle to solve analogical reasoning problems? The determining factor may be a child's cognitive profile relevant to analogical reasoning ability, comprised of effective IC and WM. Inhibitory control allows an individual to inhibit more salient, featural match responses, and select a less salient, but correct, relational match (e.g., Richland et al., 2006; Viskontas et al., 2004). Working memory allows an individual to simultaneously process multiple contexts and pieces of information present in an analogy (e.g. Gick & Holyoak, 1980; Halford, 1993; Simms et al., 2018). Due to the protracted development of these cognitive capacities, analogical reasoning similarly develops gradually over time, with initial stages presenting in children as young as 3-5 years old and maturing into adolescence (e.g., Alexander et al., 1987; Goswami & Brown, 1989; Rattermann & Gentner, 1998). In the case of a scene analogy, Richland and colleagues (2006) find that children have difficulty ignoring featural matches in favor of relational matches until they are 9-11-years-old, with children showing an increase in successful problem solving between the ages of 3 and 11, as children's cognitive profiles develop.

With this protracted development of cognitive profile in mind, we might expect differences in the effectiveness of gesture instruction. For very young children their IC and WM may be so limited that they may not be able to capitalize on gesture's ability to index spoken instruction to referents in a scene analogy, and therefore, gesture may not be helpful for disambiguating

complex spoken instruction. However, for slightly older children, we may find that gesture provides the exact boost they need: They may have the cognitive profile in place to benefit from instruction, and gesture may give them an extra boost by literally pointing them in the right direction to help them make sense of spoken instruction. For even older children with high IC and WM capacity, who typically demonstrate near-adult like ability on problems of analogy, receiving spoken instruction, even without gesture, may be enough support for understanding the structure of analogies.

Present Study

We test these predictions in the present study. To do this, we compare how children across a wide age range (4-11-year-olds) solve scene analogy problems before or after speech alone or speech and gesture instruction while monitoring their visual attention with eye tracking. Using a wide age range will allow us to understand how cognitive profile contributes to the effectiveness of gesture instruction. Using eye tracking will allow us to understand how gesture aids in disambiguation of spoken instruction meant to refer to an item within a relation, that could instead be linked to a featural match. Through this approach, we will address three questions: 1) Do children benefit differently from speech alone versus speech and gesture instruction on analogical reasoning based on their age (as a proxy for cognitive profile)? 2) Can we find evidence that gesture instruction helps disambiguate spoken instruction, and does this depend on age? 3) Do looking patterns associated with type of instruction impact whether children at different ages learn from instruction? Results of this study will add to our general understanding of the mechanisms by which children learn and explore the nuances of when gesture may or may not help beyond spoken instruction. And by focusing on analogical reasoning, we also explore the

utility of gesture instruction in a domain that is important for academic success that has been understudied in the gesture-for-learning literature.

Furthermore, this study aims to elucidate the findings of Study 1, where perhaps children did not show a benefit of speech+gesture instruction beyond speech-alone instruction either because they lacked the sufficient cognitive profile or because the length of the task was not conducive for demonstrating learning. It may be that the children used in Study 1 could not overcome the maturation limitations that come with their young age, and cannot extend the understanding gained during instruction to subsequent problem solving when instruction is no longer immediately supportive. Slightly older children who still do not have mature analogical reasoning ability, but have a slightly more foundational cognitive profile, may be able to demonstrate improvements after instruction and learn from gesture. Alternatively, the length of the task (i.e., use of 24 pre-posttest trials and 4 training trials) used in Study 1 may have caused fatigue, where learning at posttest was less likely to be evident. A shortened version of the task used in Study 1 may impact children's learning differently.

Method

Participants

Children between the ages of 4 and 11 years old ($N = 323$; 159 females) participated in the present study during a visit to a science museum[1]. Children were randomly assigned to one of two conditions ($n_{speech\text{-}alone} = 160$; $n_{speech+gesture} = 163$), with a target of ~ 20 children per age group

[1] Although we did not collect demographic information from individuals, our sample was representative of the general profile of museum visitors. According to museum reports based on short surveys with museum visitors, visitors to the museum represent a number of different racial and ethnic backgrounds (70% White, 10% Hispanic, 6% African American, 6% Asian, 5% Other, < 1% Native American, Native Hawaiian), and are also diverse in socioeconomic status, based on self- report measures of perceived socioeconomic status (13% lower or lower-middle class, 54% middle class, 33% upper middle or upper class) and parent or guardian's highest level of formal education (1% < high school diploma, 18% high school diploma, 16% associates degree, 35% bachelor's degree, 21% master's degree, 7% PhD, or other terminal professional degree, 3% not reporting).

in each condition. An additional 62 children participated in the study but were excluded from analyses for eye tracker malfunction (n = 20), parental involvement (n = 7), language barrier (n = 2), lack of response from participant (n = 7), poor eye tracking (n = 3), and experimenter error (n = 23). Two participants decided they did not want to continue before being assigned a condition. Informed consent was obtained from a parent or guardian of each participant, and verbal assent was obtained from children. Children participated individually in one 3-5 minute experimental session and received stickers as compensation.

Materials

Warm-up examples. Children were shown two scenes depicting relations occurring between items. For example, a scene showed one animal (e.g., elephant) reading to another animal (e.g. rabbit), and another scene showed an animal (e.g., duck) on top of another animal (e.g., cow). Instruction was provided that highlighted the relation of interest (i.e., *patterns* of 'reading' and 'on top of'). These trials served to familiarize children with our use of the term *pattern* and how items can be *relationally associated*.

Pre- and Post-Instruction Stimuli. Children were presented with printed copies of scene analogies similar to those used in Study 1. Stimuli were bound in a binder and displayed in front of the child on a small easel, with one pair of scenes presented at a time. Scenes depicted the relation of *chasing* occurring between items (i.e., animals or people).

Instruction Stimuli. Similar to pre- and post-instruction trials, printed instruction stimuli included two scenes in which a chasing relation was depicted in both scenes, and a featural match was located in the target scene. Unlike pre- and post-instruction trials, no item was circled in the instruction stimuli.

Eye Tracker. Eye tracking data were collected via corneal reflection using a Tobii Pro Glasses 2. Tobii software was used to perform a 1-point calibration procedure. This step was followed by the collection and integration of gaze data using Tobii Pro Lab (Tobii Technology, Sweden). Data were extracted on the level of individual fixations as defined by Tobii Pro Lab software—an algorithm that determined if two points of gaze data are within a preset minimum distance from one another for a minimum of 100 msec, allowing for the exclusion of eye position information during saccades. After extraction, fixations were manually mapped by research assistants. Individual fixations were classified as either oriented towards one of the items of interest within the scenes (e.g., to the item chasing in the source scene, to the item being chased in the source scene, to the featural match, etc.), other areas around the items within the scenes, or the space surrounding the scenes. Research assistants assigned each fixation to an area of interest (AOI), based on its location (e.g., if a fixation was located on or within the immediate area surrounding the featural match, it was manually mapped as a featural match fixation).

Procedure

Children participated individually at a table in a corner of the museum floor. Children were told they were going to play a picture game while wearing eye tracking glasses. After a brief explanation that the purpose of the glasses is to 'help us see what you see', an experimenter fitted them with the glasses. Children were seated approximately 40 cm in front of the printed stimuli next to an experimenter. The printed stimuli were displayed in a binder mounted on an easel. This allowed the experimenter to quickly flip between stimuli and gesture to the stimuli during instruction trials if a child was assigned to the speech+gesture condition. It also ensured proper eye tracking -- children could see the stimuli directly in front of them, and did not have to

look down towards the table, which would have disrupted our ability to capture their visual attention via the eye glasses. Children's position was calibrated and adjusted if necessary, and they were asked to remain as still as possible during the rest of the game while eye tracking data were collected.

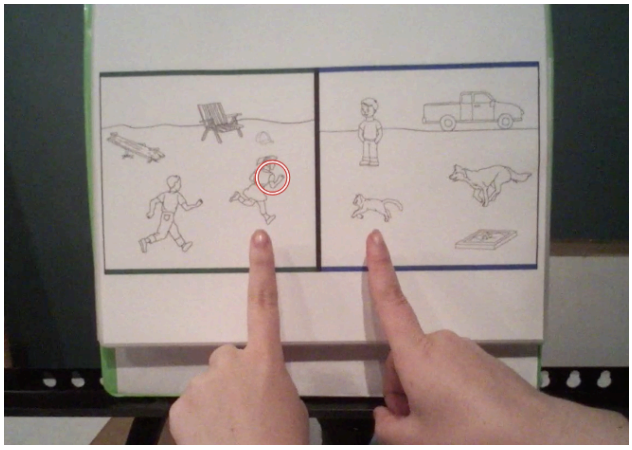
First, the experimenter explained the relational pattern in the two warm-up trials, meant to help promote relational thinking (see Materials for details). Next, children completed one pre-instruction trial. After orienting children to the two scenes presented simultaneously (e.g., one side has blue edges and one has green edges), children were asked, "*Which thing in the picture with the blue edges is in the same part of the pattern as the circled thing in the picture with the green edges?*" An item in the source scene (e.g. in Figure 2A page 27, blue edges) was circled and had a corresponding relational item and featural match in the target scene (e.g. in Figure 2A page 27, green edges). The task was self-paced, but if no response was given after a few seconds, the children were re-prompted by the experimenter.

Following the pre-instruction trial children were asked to pay attention to two instruction trials to learn about the pattern in the pictures (see Figure 6). Children were randomly assigned to receive *speech+gesture* instruction or *speech-alone* instruction provided by the experimenter. Scenes in both conditions were accompanied by spoken instruction that was identical to the audio recorded instruction used in Study 1, and the gestures in the speech+gesture condition were identical to those used in Study 1. The ambiguity of this instruction occurs when the boy in the source scene is referenced, because there is also a featurally similar boy in the target scene (i.e., the featural match). When the boy is mentioned in speech it may be difficult for children to reconcile *which* boy is being discussed: the one in the relation of chasing or the featural match. This

confusion or ambiguity could contribute to difficulty identifying the relational structural in an analogy problem.

Finally, a post-instruction trial was administered after children viewed the instructional trials, with an identical prompt and procedure as used during the pre-instruction trial.

Figure 6. Example of children's view during a speech+gesture training trial. The red circle represents the location of one fixation.



Measures of Visual Attention

Measure of Attention during Pre- and Post-Instruction Trials. Visual attention during pre- and post-instruction trials was quantified by generating areas of interest (AOIs) that represent different portions of the participant's field of view using Tobii Pro Lab. There were 11 AOIs in total. The AOIs encompassed regions within the scene pairs and areas in the field of view that were outside of the scene analogy. This included an AOI for each of the items in the scenes (items in chasing relations, featural match, neutral items), AOIs for when the participant fixated on the experimenter, on the experimenter's gesture, and on their own hands, and an AOI for looking elsewhere in the museum. Proportion of time spent looking to each AOI was then calculated by dividing the time looking to an AOI during a trial by the total time looking during a trial. For the sake of the present analyses, we focused on the AOI representing the featural match.

Children's ability to avoid featural matches is one of the key issues children overcome as they develop successful analogical reasoning. By assessing visual attention to the featural match we can assess whether gesture is more effective than speech alone for driving attention away from irrelevant featural components.

Measures of Attention during Instruction. Attention during instruction was quantified in two ways: 1) children's ability to synchronize their visual attention with spoken instruction and 2) 'check-ins' with the featural match during ambiguous spoken instruction.

Following Score. Following a similar procedure to Study 1, we calculated a 'following score' for each instruction trial. Children could receive a score of 0 to 4 on each training trial, and scores were averaged across the two training trials to generate an overall following score for each child. The average following score was used in analyses.

Check-in Score. Check-ins with the featural match are instances when the item that is perceptually similar to the featural match is referenced in speech and simultaneously fixated on by the child. In each instruction trial, there were two time segments during which a check-in could occur. For example, in the instruction trial depicting a boy chasing a girl in the source scene and a featural match boy in the target scene, the two relevant time segments occur when the experimenter said either '*The boy is chasing the girl'* or *'The boy is in the same part of the pattern as the dog because they are both chasing.'* For each segment, a child would receive a score of 1 if they looked to the featural match boy in the target scene rather than the boy in the source scene. Children would receive a score of 2 for a given instruction trial if they looked at the featural match boy during both time segments in which the boy in the relation was mentioned. Thus, whereas a score of 4 is possible for following score, a score of 2 is possible for check-in score.

Check-in scores from the two instruction trials were averaged to generate an overall check-in score for each child. The average 'check in' score was used in analyses.

Results

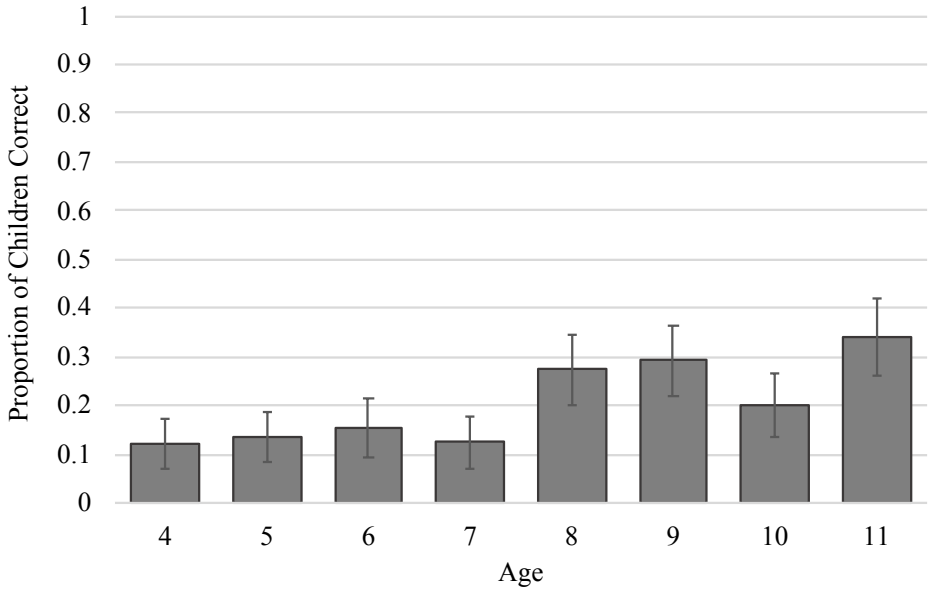
All analyses were conducted using R Studio (version 1.1.456), supported by R version 3.6.0. Analyses relied on the *stats* package, which allows for ANOVA and regression modeling (R Core Team, 2017). When running binomial generalized linear regression models assessing the impact of condition on accuracy or choice of the featural match at pre- and post-instruction, the speech-alone condition was set as the baseline condition and compared against the speech+gesture condition. For analyses of visual attention, which did not use a binomial outcome, generalized linear regression models were used. Again, speech-alone was set as the reference level for these analyses.

Before addressing our main questions of interest, we wanted to establish (1) that there were no significant performance differences pre-instruction between children who had been randomly assigned to the speech-alone versus speech+ gesture condition – we found that there were not: Both across all children and within age groups, there were no condition differences between pre-instruction accuracy or choice of the featural match (all $ps > .1$), and (2) that age could serve as a proxy for cognitive profile. To do this, we asked whether children's ability to solve analogical reasoning problems could be predicted by age and visual attention before instruction. Because previous work has shown that children are more likely to succeed on analogical reasoning problems when their IC and WM improve (e.g., Doumas et al., 2018; Simms et al., 2018), and a marginal relation exists between children's IC and looking to the featural match, such that lower

IC predicts more looking (Study 1), we reasoned that if age was predictive of these measures this would suggest that age can serve as a proxy for cognitive profile.

While only 20% of children correctly answered the pre-instruction trial, there was a main effect of age when predicting accuracy, such that older children were more likely to answer the problem correctly than younger children (Figure 7, $\beta = 0.18$, $SE = 0.06$, $t = 2.89$, $p = .004$), replicating previous work (e.g., Richland et al., 2006). And, as with previous studies using scene analogy problems, we found that the most common error children made was to choose the featural match – 64% of children made this type of error. In terms of visual attention, we assessed whether children's proportion looking to the featural match before instruction predicted their performance, as this is a key looking pattern associated with making featural errors (e.g., Study 1; Thibaut et al., 2010; Thibaut & French, 2016). On average, children who correctly answered the pre-instruction trial allocated less of their attention to the featural match ($M = 0.12$, $SD = 0.08$) than children who made featural errors ($M = 0.14$, $SD = 0.08$). Models predicting accuracy by visual attention to the featural match showed that proportion looking to the featural match was negatively related with accuracy ($\beta = -0.00$, $SE = 0.00$, $t = -2.22$, $p = .026$) and positively related with featural errors ($\beta = 0.00$, $SE = 0.00$, $t = 3.51$, $p < .001$). In sum, these results not only replicate previous work finding that prior to instruction children who are older and attend less to the featural match more successfully solve scene analogy problems, but also provide support for considering age as a proxy for cognitive profile.

Figure 7. Proportion of children within each age correct on the pre-instruction trial.



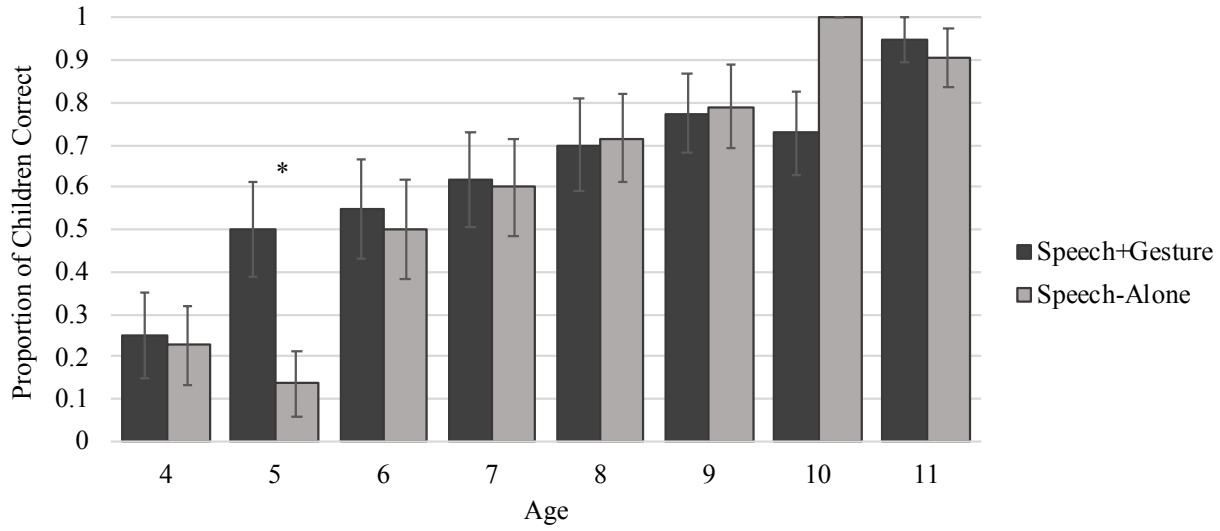
Impact of Age and Instruction on Children's Analogical Reasoning Ability

To understand how speech-alone versus speech+gesture instruction affected children's performance on the post-instruction trial, we limited the remainder of our analyses to children who incorrectly answered the pre-instruction trial (speech alone: n = 124; speech+gesture: n = 133) – importantly, a similar number of children were excluded from both experimental groups. Our first main question was whether the impact of gesture instruction on children's analogical reasoning is dependent on their cognitive profile (measured by age). Overall, more children in the speech+gesture condition correctly answered the post-instruction trial than children in the speech-alone condition (speech+gesture: 63% vs. speech-alone: 59%). But, from Figure 8 it is clear that performance is also dependent on age, and when considering performance binned by age, we see that the difference between conditions appears most pronounced for 5-year-olds. To determine whether these patterns were statistically significant, we constructed a generalized line-

ar model with accuracy (0, 1) as the dependent measure, and age, condition (speech alone, speech+gesture), and an interaction between age and condition as predictors of interest. In line with Figure 8, the model revealed a main effect of age, suggesting that older children performed better after instruction than younger children ($\beta = 0.62$, $SE = 0.12$, $t = 5.35$, $p < .001$), and a trending main effect of condition, suggesting that children improved marginally more after speech+gesture instruction than speech-alone instruction ($\beta = 1.82$, $SE = 1.06$, $t = 1.72$, $p = .085$).

However, these results should be considered within the context of a marginal interaction between age and condition ($\beta = -0.25$, $SE = 0.15$, $t = -1.69$, $p = .092$), where post hoc analyses indicate that only 5-year-old children demonstrate a benefit for speech+gesture compared to speech-alone ($\beta = 1.75$, $SE = 0.89$, $t = 1.97$, $p = .048$), and for all other children, there was not an effect of condition ($ps > .1$). Although this interaction was only marginally significant, this is likely due to the consideration of such a wide age range, with most age groups showing a clear lack of difference in response to instruction condition. The presence of an interaction aligns with the a priori hypothesis that gesture may only boost learning beyond speech-alone instruction at certain ages. Given previous work within the analogical reasoning literature that shows 5-year-olds demonstrate greater difficulty with problems incorporating featural matches than older children (e.g., Richland et al., 2006; Simms et al., 2018), it makes sense that gesture would provide these children the most benefit.

Figure 8. Proportion of children within each age correct on post-instructional trial separated by condition.



Gesture's Effect on Visual Attention during Instruction

Gesture instruction has previously been shown to help children follow along with spoken instruction and facilitate performance on subsequent assessments (Wakefield et al., 2018b). To understand how visual attention might play a role in the marginal behavioral effects of gesture on children's post-instruction performance, we next asked how condition and age influenced children's visual attention during instruction. Here, we used two measures of visual attention: following score and featural match check-in score. Children's following score is an index of whether they looked at relevant referents of the problem (i.e. items involved in the relation of chasing) when the referents were mentioned in spoken instruction. Children's featural match check-in score is an index of whether children attended to the featural match when the instructor's speech was meant to reference an item within a chasing relation, but was ambiguous. Without understanding the context of the analogy, children could associate the spoken referent with either an item within a relevant chasing relation (the item the instructor meant to reference) *or* the featural

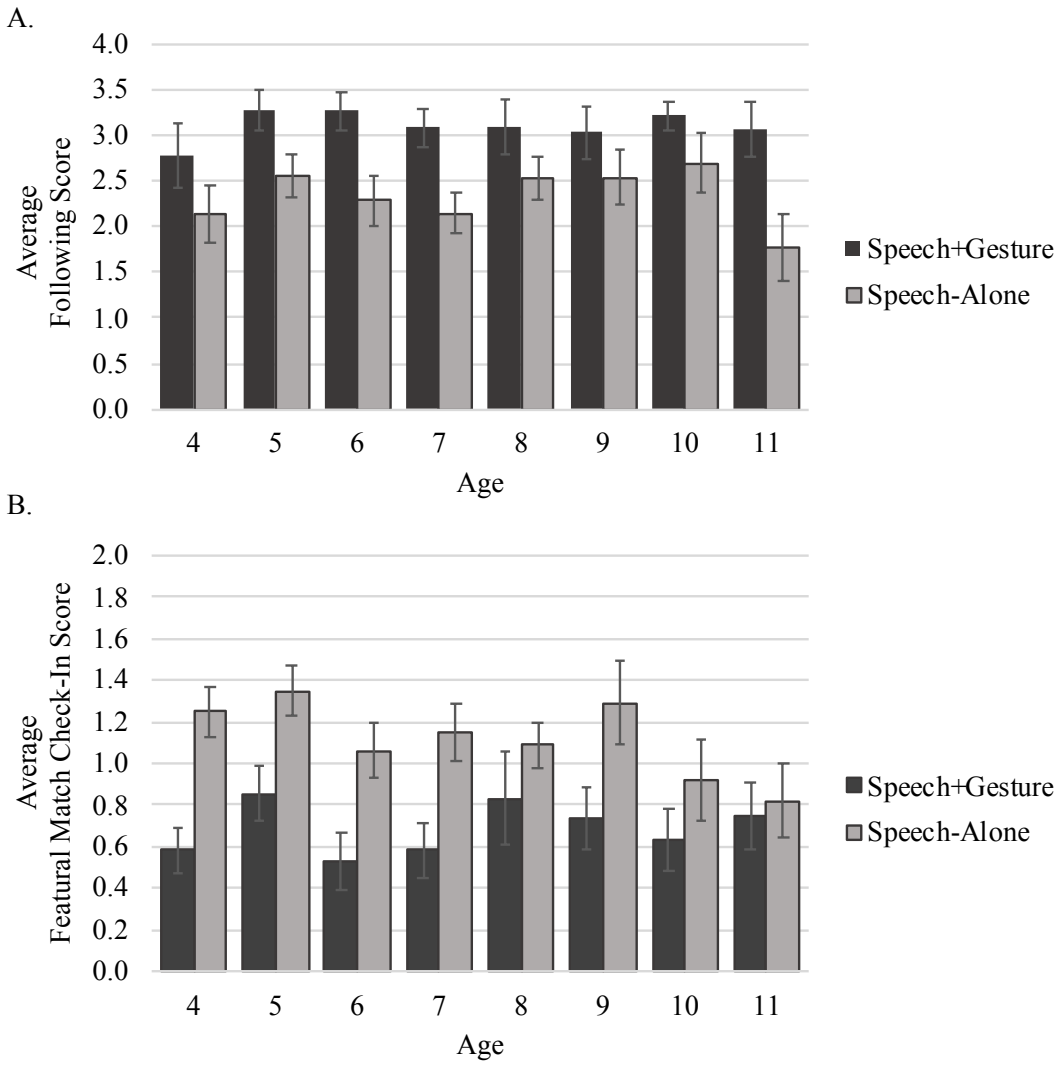
match to that item (an item that is irrelevant to the analogy). Attending to the featural match may disrupt a child's ability to effectively learn from instruction because it detracts from children's ability to process how the items within the two chasing relations are aligned.

On average, children followed along more successfully with spoken instruction if they were taught through speech+gesture (M = 3.08 out of a possible score of 4, SD = 1.10) than through speech-alone (M = 2.20, SD = 1.06). Figure 9A shows following score separated by age and condition and suggests that gesture supports effective following along with instruction for all children. Using a generalized linear model with following score as the dependent measure and age, condition (speech-alone, speech+gesture), and an interaction between age and condition as the predictor of interest, we found a main effect of condition, confirming that speech+gesture instruction supported more effective following than speech-alone instruction (β = 1.51, SE = 0.26, t = 5.81, $p < .001$). We also found no main effect of age (β = 0.02, SE = 0.06, t = 0.28, p = .783) and no interaction between age and condition (β = 0.12, SE = 0.12, t = 0.25, p = .806) suggesting that gesture is a cue that can organize visual attention regardless of a child's age.

Our second measure of visual attention during instruction was how children attended to the featural match, the key component of an analogy that draws children's attention away from the relational information (e.g., Guarino et al., 2019; Thibaut et al., 2010; Thibaut & French, 2016). Specifically, we asked whether children attended to the featural match during the time intervals when the spoken instruction was ambiguous as to whether the instructor was referring to an item within a relation or the featural match to that item outside of the relation (e.g., *Which boy is being referred to: the boy in the relation of chasing or the featural match boy?*). Because the lure of featural matches are at the root of young chilldren's difficulties with problems of

analogy, the most ambiguous portion of the instruction is when the item that is involved in the relation of chasing and perceptually similar to the featural match is discussed in speech.

Figure 9. A. Average check-in scores split by age and condition. B. Average following scores split by age and condition.



On average, children checked-in more with the featural match if they were in the speech-alone condition (M = 1.29 out of a possible score of 2, SD = 0.48) than in the speech+gesture condition (M = 0.71, SD = 0.57). But again, the amount of difference between condition seems to differ by age (Figure 9B). Using a generalized linear model with check-in score as the de-

pendent measure, and age, condition (speech-alone and speech+gesture), and an interaction between age and condition as the predictors, we found a main effect of condition, such that speech+gesture instruction facilitates fewer check-ins than speech-alone instruction (β = -1.77, SE = 0.47, t = -3.76, p < .001), and a main effect of age, such that older children check-in with the featural match less than younger children regardless of the type of instruction received (β = -0.11, SE = 0.05, t = -2.36, p = .019). These effects should be interpreted within the context of a trending interaction between condition and age (β = 0.00, SE = 0.06, t = 1.95, p = .052). Post hoc analyses indicate that this trending interaction results from a developmental shift between younger and older children (Table 1): Generally, older children are less likely to show a significant difference in check-in score across conditions, suggesting that they can make use of either speech-alone or speech+gesture instruction to avoid the featural match. In contrast, younger children's visual attention is oriented away from the featural match more effectively by speech+gesture than the speech-alone instruction. This suggests that younger children use the added support of gesture to disambiguate speech and orient their attention away from featural matches.

Table 1. Post hoc analyses for testing condition effects predicting featural match check-ins

Age	Beta (SE)	*p* value
4 year-olds	-1.33 (0.32)	**< .001**
5 year-olds	-0.92 (0.35)	**.012**
6 year-olds	-1.07 (0.38)	**.008**
7 year-olds	-1.13 (0.37)	**.005**
8 year-olds	-0.51 (0.44)	.259

9 year-olds	-1.11 (0.47)	**.024**
10 year-olds	-0.58 (0.46)	.216
11 year-olds	-0.14 (0.46)	.770

In sum, the main effect of condition for following score suggests that gesture is effective for directing all children's attention to the referents of spoken instruction. However, when considering the ambiguous portion of instruction, we see differences across age in the relative effectiveness of instruction. For older children, the alignment provided in spoken instruction, "*See, the boy is chasing the girl, and the dog is chasing the cat*" is enough context to recognize that when the instructor refers to the '*boy chasing the girl'* that the boy being referenced is the boy in the chasing relation, not the featural match that is outside of the relation: There is no added benefit of gesture for disambiguating speech. However, for the younger children, we see that gesture *does* have an effect. Children are less likely to look to the featural match when they receive speech and gesture instruction, compared to speech alone instruction. This suggests that gesture is helping disambiguate spoken instruction for these younger children.

Impact of Visual Attention during Instruction on Children's Analogical Reasoning

Having established that gesture does impact visual attention during instruction, whether this is for all children (following) or only children of particular ages (featural match check-in), we ask whether these patterns of visual attention can explain our behavioral results – that overall speech+gesture seems to marginally improve performance compared to speech alone, but that

this effect is driven by 5-year-old children, who show significantly better performance following speech+gesture instruction compared to speech-alone instruction.

To understand the relation between following along during instruction and performance on the post-instruction trial, we asked whether trial accuracy (0, 1) was predicted by following score. Age was not included in the model, as we found that it was not a relevant predictor of following. Our model revealed that following score was not predictive of accuracy (β = 0.07, *SE* = 0.06, t = 1.08, p = .280). This suggests that even though gesture helps children follow along with spoken instruction, this organization of visual attention does not contribute to its learning effects in the case of scene analogies.

To understand the relation between checking in with the featural match during instruction and performance on the post-instruction trial, we took into account our finding that, in general, younger children checked in less with the featural match when they received speech+gesture instruction than speech-alone instruction, but older children did not show this difference. This distinctly different pattern of results between younger and older children motivated the use of a median split by age (see Wakefield, Novack, & Goldin-Meadow, 2017 for a similar approach): We constructed two models to ask whether check-ins during instruction were predictive of performance on the post-instruction trial for older (8-11 years) and younger (4-7 years) children separately. Here, we found that, whereas older children's check-ins with the featural match did not significantly predict their accuracy at post-instruction (β = -0.12, *SE* = 0.18, t = -0.66, p = .512), younger children's check-ins with the featural math *were* predictive of their performance on the post-instruction trial: Check-ins were negatively related to successful problem solving (β = -0.45, *SE* = 0.18, t = -2.58, p = .009). This suggests that the ability of gesture instruction to di-

rect attention away from the featural match and disambiguate the meaning of an instructor's speech is the critical factor impacting analogical understanding for younger children.

Discussion

The goals of the present study were to explore whether the impact of adding gesture to spoken instruction on analogical reasoning depends on children's cognitive profile, and to use eye tracking to further understand how gesture might facilitate learning by disambiguating spoken instruction. Our behavioral results suggest a marginal benefit of gesture instruction over speech alone, but only 5-year-old children showed a distinct advantage from speech+gesture instruction when solving the post-instruction trial. This suggests that age – which we demonstrated was a good proxy for cognitive profile based on the relation between performance measures, visual attention, age, and in keeping with previous literature – does impact the utility of gesture for supporting analogical reasoning ability. To understand how disambiguation of speech may play a role in these results, we turned to eye tracking. We found evidence that gesture helps children follow along with spoken instruction, but that this was not predictive of successful problem solving post instruction. Rather, check-in score – visual attention towards the featural match at the point in instruction that was most ambiguous – was negatively predictive of post instruction success for younger children, but not for older children. This lends support to previous arguments that at the root of children's struggle with analogical reasoning is an inability to ignore featural, or superficial, matches in favor of relational matches, and that looking to the featural match is associated with making these types of errors (e.g., Guarino et al., 2019; Thibaut et al., 2010; Thibaut & French, 2016). Although more work must be done to fully explore the impact of gesture instruction for analogical reasoning, these results suggest that one way gesture may help

learning in this domain is through directing visual attention in a way that clarifies spoken instruction, but how much of a boost children get depends on their cognitive profile.

Gesture Benefits Some Children More Than Others

Our results suggest that in the case of analogical reasoning, gesture's ability to disambiguate speech may be particularly useful for 5-year-old children who have the foundational cognitive abilities in place to benefit from gesture during instruction. Five-year-old children may be at a pivotal time in development of analogical reasoning ability: While they have a limited cognitive profile and immature analogical reasoning, their IC and WM capacity are developed to the point that they can utilize the added support gesture provides. This finding that prior knowledge and ability impacts the utility of gesture corroborates other work in the gesture-for-learning literature. Children need some degree of prior knowledge within a domain that serves as a foundation that gesture instruction can build from (Congdon et al., 2018; Wakefield & James, 2015).

Importantly, our eye tracking data suggest what the added benefit of gesture might be: 5-year-old children showed an increased ability to follow along with instruction and less check-ins with the featural match when they learned through speech and gesture instruction versus speech alone instruction. Thus, the argument could be made that gesture is helping organize children's visual attention in relation to spoken instruction and clarifying ambiguous instruction. But, only check-ins predicted success on the post instruction trial. Considering this in relation to previous work with eye tracking, this may seem puzzling. Wakefield and colleagues (2018b) found that following along with spoken instruction *did* predict subsequent performance in the case of mathematical equivalence. However, in their measure of following, spoken instruction was ambiguous; Whereas in the present study, the general measure of following encompassed spoken in-

struction that was predominately not inherently difficult for children to decipher because the majority of items referenced in speech could only be associated with one unique item in a scene. In contrast, the speech during the featural match check-in measure *was* ambiguous, and is thus more analogous to the measure of following used by Wakefield et al. (2018b). In both of these cases, gesture is effective at clarifying parts of spoken instruction that are ambiguous, yet critical, for learning. Taken together, results from the current study and previous work suggest that gesture's power to disambiguate spoken instruction is an important mechanism by which gesture shapes learning. And in the case of analogical reasoning, gesture can help children overcome one of the most challenging aspects of problem solving: Clarifying for these children which items are in the relation of chasing and critical for solving the analogy, by helping them avoid the lure of a featural match.

While 5-year-olds may be in the developmental 'sweet spot' to benefit from gesture instruction, why does incorporating gesture not benefit all children equally? For all other children, those younger and older than 5 years, there was not a significant benefit of speech and gesture, compared to speech alone instruction, on post-instruction performance. It makes sense that older children (8-11-year-olds) demonstrated learning after both types of instruction: These children seemingly have all the necessary cognitive abilities and prior knowledge needed to utilize either type of instruction. Even though they struggled prior to instruction, their more developed IC and WM allowed them to learn even from speech-alone instruction, and the addition of gesture is not necessary for learning the task. This is evidenced by the lack of difference between the number of check-ins with the featural match in the speech+gesture versus speech-alone conditions. Likely because they had the capacity to hold more information in WM, they were able to consider the

instructor's alignment of the chasing relations and recognize which items were being referenced during instruction based on spoken instruction alone, and did not need gesture to organize their visual attention and help them make sense of instruction.

On the other end of the age range, the youngest children, 4-year-old children, may not have a sufficient cognitive profile in place to benefit more from gesture instruction than speech alone instruction. While gesture supports effective visual attention during instruction for these children, their IC and WM may be too underdeveloped to extend their understanding beyond the moment, when the support of gesture is no longer immediately present. Thus, even though they looked to the featural match less in the gesture condition, they could not process the multiple relations mentioned in spoken instruction effectively.

Interestingly, 6- and 7-year-old children did not perform similarly to 5-year-old children or older children. While their visual attention was more effectively guided by a combination of speech and gesture instruction, as seen with their younger peers, they did not show the added benefit of gesture post instruction. The non-significant difference between conditions at post-instruction performance for these children may speak to their ability to disambiguate the instructions to some extent when only speech was provided. That is, these children may be able to disambiguate the instructions even with speech alone to a greater extent than 4- or 5-year-olds, but not as effectively as older children. And because they have slightly more mature cognitive profiles (i.e., more developed IC and WM) than younger children, they may be better equipped to extend their understanding gained during instruction to post-instruction solving. Together, these results reflect that children's cognitive profile makes a difference for whether gesture facilitates learning above and beyond speech alone instruction.

Addressing Potential Limitations and Future Directions

While this work makes strides towards understanding the nuances of gesture's effects on learning, there are potential limitations that should be addressed. First, we suggest that age can serve as a proxy for a child's cognitive profile without having independent measures of IC and WM. Although collecting independent measures of IC and WM would have been ideal, previous work using scene analogies has established that these executive functions correlate with children's age (5-11-years-old: Simms et al., 2018) and with their analogical reasoning ability (WM: Simms et al., 2018; IC: Study 1), *and* that children's visual attention is correlated with performance and IC (Study 1). Specifically, IC, measured using the Erikson Flanker task, is positively correlated with accuracy and attention to relationally similar items prior to instruction, and negatively correlated with choosing the featural match and attention to the featural match. Therefore, while it may be advantageous in future work to collect direct measures of children's cognitive profile, here, we find the same relation between age, visual attention patterns, and analogical reasoning ability that has been documented in previous work. We are therefore confident that, motivated by previous work, age is associated with cognitive profile.

A second potential limitation is the length of our intervention, which consisted of one pre-instruction trial, two instruction trials, and one post-instruction trial. We designed the study based on previous gesture-for-learning literature showing children *can* benefit from a short intervention (Church et al., 2004; Rowe, Silverman, & Mullan, 2013; Valenzeno, et al., 2003). For example, Church and colleagues (2004) tested children's knowledge of three types of Piagetian conservation (water, length, number) using one question about each type of conservation before and after they watched one instructional video about conservation that either incorporated speech

and representational gestures or speech alone. Similarly in the analogical reasoning literature, Gentner and colleagues (2016) tested how well children can analogically compare separate contexts after a short intervention. They first exposed children to one pair of model skyscrapers that varied in degree of alignment based on experimental condition, and then asked them build a structure as tall as possible that was 'strong' and repair a structure so it was 'strong'. Through successful comparison of the two model skyscrapers children could identify that a diagonal brace helps make a building 'stronger'. In the present study, we did find an effect of gesture instruction, above and beyond that of speech alone instruction, for children at a pivotal point in their analogical reasoning development. This suggests that once again, gesture can impact performance in a short period of time. However, it would be interesting to conduct future work lengthening the period of instruction, as this may allow children more opportunity to benefit from instruction, especially younger children who may need more examples to support their learning.

Finally, while not a limitation, the current work represents a starting, not an ending point, motivating additional questions to answer. For example, similar work using the test-bed of analogical reasoning should consider even younger children. The children in this study likely all had an underdeveloped, but nevertheless present, relevant cognitive profile to support the rudimentary stages of analogical reasoning (e.g., Davidson, Amso, Anderson, & Diamond, 2006). Even 4-year-olds have been shown to have some degree of IC and WM that allow them to make very simple comparisons – one of the basic building blocks for mature analogical reasoning (e.g., Davidson et al., 2006). To more fully understand the impact of gesture on children with little to no relevant cognitive skills, one could extend and adapt this task to incorporate 2- or 3-year-olds, given that some suggest children younger than 4-years-old have rudimentary relational reasoning

capabilities (e.g., Ferry et al., 2015; Goswami & Brown, 1989; Rattermann & Gentner, 1998). The expectation would be that younger children, including 4-year-olds, would not benefit from gesture more than speech alone, and would strengthen the conclusions drawn from the present data.

Additionally, the impact of gesture is not only nuanced in terms of children's current cognitive profile, but many other contextual or situational factors have been cited as playing a role in the effect on learning. For example, the advantage of speech+gesture compared to speech-alone instruction is not always evident in immediate measures at post-instruction, but rather in follow-up measures, from 24 hours (Cook et al., 2013) to 4 weeks (Congdon et al., 2017) after initial training. The one-trial post-instruction assessment may have limited the evaluation of learning.

Conclusion

In sum, the results of the present study extend our understanding of how gesture instruction impacts learning to the domain of analogical reasoning, while providing further insight into how gesture can help disambiguate spoken instruction and how individual differences in a child's cognitive profile impacts the utility of gesture. These findings have important implications for designing teaching methods to support analogical reasoning, but also using gesture as a teaching tool more broadly. Because analogical reasoning shows such a protracted development, due to a slowly developing cognitive profile, it seems that only at certain points will gesture help children more than speech only instruction. Recognizing when this tool can be used could lead to faster growth in a skill that is at the root of a wide range of cognitive skills, such as innovation and creativity (for review see Gentner & Smith 2013). More broadly, this work speaks to one of

the reasons *why* gesture helps learning, but also emphasizes that individual differences influence the impact gesture can have. Across Studies 1 and 2 we can see that the impact of gesture is dependent on a number of factors, including the cognitive capacities that a child brings to an instructional session. And when comparing this work to previous research in other domains, it seems the impact of gesture can depend on the type of material being taught. Future work should continue to delve into the mechanisms by which gesture shapes learning, including situational factors of the context in which gesture is used, and consider a child's cognitive state as an important piece of this puzzle.

CHAPTER FOUR

WHEN IS GESTURE BENEFICAL FOR LEARNING? A META-ANALYSIS INTO THE NUANCES OF GESTURE'S EFFECTS ON LEARNERS

Introduction

Studies 1 and 2 build on existing literature by addressing whether gesture supports children's learning of analogical reasoning. In general, the existing literature suggests that gesture supports learning across a wide range of contextual conditions and in a variety of instructional environments. However, the findings of Studies 1 and 2 suggest that this is not always the case: All children do not benefit from gesture equally. In Study 2, we found that gesture benefitted 5-year-old children above and beyond speech alone instruction, and that children older and younger than 5 years did not show additional benefits from gesture. Importantly, the findings of Studies 1 and 2 tell an inconsistent story about the effects of gesture, where Study 1 found that 4- and 5-year-olds did not show an added benefit from gesture. This is not the first time contradictory or non-significant effects of gesture have been found (e.g., Beattie & Shovelton, 1999; Driskell & Radtke, 2003), suggesting that speech+gesture communication is not *always* more effective than speech-alone. The impact of gesture instruction may depend on a number of factors, such as the characteristics of the task (e.g., content of the accompanying spoken message: Driskell & Radtke, 2003, task difficulty: McNeil et al., 2000; timing of assessment: Congdon et al., 2017; Cook et al., 2013), population characteristics of the learner (e.g., age: McNeil et al., 2000, primary language: Sueyoshi & Hardison, 2005), or even characteristics of the gesture itself (e.g., redundancy

of the gesture with accompanying speech: Hostetter, 2011, form of the gesture: Dargue & Sweller, 2018a). The goal of Study 3 is to understand the variability across the gesture-for-learning literature, and identify under which contextual variations of the learning environment is gesture most beneficial for learners.

Systematic Evaluations of Gesture's Impact on Communication and Comprehension

The idea that the impact of gesture on learning depends on situational or contextual factors is not new: Kendon (1994) considered this question in the first systematic evaluation of gesture's role in communication. Although Kendon concluded that gestures benefit listeners' understanding of spoken material, he posited that this is not likely to be universally true, but rather there is variability in *how* and *when* listeners benefit from gesture. Since then, two meta-analyses have been performed to address what factors moderate the impact of gesture on communication and comprehension[1].

Hostetter (2011) synthesized 35 years of work by performing the first meta-analytic review of literature investigating the effects of gesture for communication. Hostetter not only asked whether gesture benefits communication, but also examined potential moderators that may account for variability found across individual studies. Although it was concluded that gestures provide a significant, moderate benefit to communication, this effect was moderated by three factors: whether the gesture depicted motor actions or abstract topics, whether the gesture was completely redundant or not with the accompanying speech, and the age of the learner. Specifically, gestures that illustrate motor actions and are non-redundant with speech are more commu-

[1] The scope of these previous meta-analyses differ from each other, as well as from the present meta-analysis. Hostetter's (2011) investigation was conceptualized as an investigation into the effect of gesture on communication, whereas Dargue and colleagues' (2019) investigation was conceptualized as an investigation into the effect of gesture on comprehension. The present meta-analysis is an investigation into the effect of gesture on learning. Although there is obvious overlap between the scope of these three meta-analyses, these distinctions result in slightly different samples of studies and different study implications.

nicative, and children benefit more from gestures than adults. A number of other potential moderators did not account for variability in the field: the degree to which gesture is spontaneous, the language proficiency of the listener (i.e., nonnative speakers or those with cognitive impairments), the degree to which gesture benefits the speaker versus the listener, whether gesture serves to capture the listener's attention, whether benefits are above and beyond seeing the speaker's face and mouth movements, and whether gesture benefits comprehension, memory, and learning equally.

Nearly a decade later, Dargue, Sweller, & Jones (2019) further explored under what conditions gesture benefits comprehension by performing a meta-analysis that accounted for nearly an additional decade of literature. They examined whether variability across the field was due to a range of potential moderators, by both attempting to replicate Hostetter's findings and exploring an additional set of potential moderators. Specifically, they assessed whether the impact of gesture depends on gesture type used during communication (deictic, iconic, beat, or metaphoric: see McNeill, 1992 for a review of these distinctions), whether the gesture was produced or observed by the learner, how comprehension was measured, redundancy of the gesture with accompanying speech, and age of the learner. The only factor that significantly moderated comprehension was who produced the gesture, with more successful comprehension occurring when the learner produces the gestures themselves rather than observing others' gestures. In contrast to Hostetter (2011), neither redundancy of gesture or listener's age significantly predicted the effect of gesture. Together, these two meta-analyses show that gesture is beneficial for communication and comprehension, but make it clear that there are nuances to these impacts: The benefit of ges-

ture is dependent on a variety of factors, including, but likely not limited to, how the gesture is implemented during communication.

Hostetter (2011) and Dargue et al. (2019) made great strides towards synthesizing the gesture-for-learning literature, but there are still important questions to ask. First, to better understand the nuances of gesture's effect on learning, there are additional potential moderators to consider. In the present study, we ask whether three additional moderators matter, where these potential moderators are methodological variations across the gesture-for-learning literature: 1) The type of assessment – Are individuals being asked to memorize facts with the help of gesture, or learn new concepts? 2) The content domain being learned – Are individuals being asked to learn and perform tasks of mathematics, spatial reasoning, language learning, narrative comprehension, or science? 3) The timing of assessment – Are individuals being tested immediately after instruction, or is there a delay before testing? Finding that any of these factors moderate the effect of gesture is an important step in understanding *how* and *when* gesture benefits learning. Second, we aim to resolve two discrepancies among the findings of these previous meta-analyses; Whereas Hostetter found effects of 1) learner's age and 2) redundancy of gesture, Dargue et al. (2019) did not find that either factor moderated gesture's effects. Beyond addressing these novel moderators and discrepancies, we aim to replicate two additional findings of Dargue et al. (2019); We will test the effect of 1) gesture type and 2) whether gesture is produced or observed by the learner. These are two key methodological factors that vary across the literature. Thus, it is important to understand how they impact the effect of gesture for learning, and in particular, if when combined with the aforementioned methodological variations, do they differentially impact learning. In sum, with the present study we ask whether the effect of gesture on learning is im-

pacted by these seven potential moderators. Thus, we aim to make sense of the variability across the gesture-for-learning literature. Next, we will review the existing literature related to these seven factors.

Additional Potential Moderators to Consider

Researchers have asked how gesture impacts learning across a wide range of domains, using a variety of tasks and measures. Typically, these tasks either ask participants to memorize facts or learn new concepts, which can be distinguished as measures of fact-learning and conceptual-learning, respectively. Here, we consider *fact-learning tasks* measures of rote memory, where participants are asked to memorize and re-produce pieces of information (e.g., perform a language or word learning task: Kelly, McDevitt, & Esch, 2009; Wakefield et al., 2018a, or recall a narrative: Dargue & Sweller, 2018b; Dargue & Sweller, 2020). For example, Kelly and colleagues (2009) assessed the role of gesture while learning a nonnative language. They provided native English-speakers with a brief training on Japanese verbs with and without gestures, and found that the greatest word learning occurred when training was accompanied by redundant imagistic gestures. However, Kelly and colleagues (2014) found that the benefits of gesture do not apply across all language learning tasks: Using co-speech gestures to highlight the syllable structure of Japanese words did not benefit word learning, suggesting that gestures may not be advantageous for learning phonetic distinctions at the syllable level (Kelly Hirata, Manansala, & Huang, 2014). Similarly, Dargue and colleagues (2018a; 2020) find that participants' ability to recall narratives is not blanketly benefited by just any gesture, but the benefit of gesture depends on its form and its relation to the verbal communication.

Whereas fact-learning tasks assess ability to recall concrete, often isolated, pieces of information, *conceptual-learning tasks* measure ability to learn and apply rules or strategies for solving problems; participants are asked to use some form of inductive or deductive reasoning to solve a problem (e.g., mathematical equivalence: e.g., Congdon et al., 2017; Cook et al., 2013; conservation tasks: Church et al., 2004; Ping & Goldin-Meadow, 2008). To successfully solve conceptual-learning tasks participants need to have a flexible understanding of the rule or concept being taught and apply this understanding after instruction. For example, Cook and colleagues (2013) provided children with instruction explaining how to solve mathematical equivalence problems (e.g., 2 + 4 + 6 = 2 + __), where they were asked to solve for the missing addend to make the two sides equal. This tested children's ability to extend the rules underlying mathematical equivalence to a new problems after instruction. In this case, children who saw instruction with gesture were more successful at posttest than those who only saw instruction with speech. Although several studies find similar significant effects of gesture for extending learning (e.g., Congdon et al., 2017; Cook et al. 2013; Koumoutsakis, Church, Alibali, Singer, & Ayman-Nolley, 2016; Novack et al., 2014), others do not find evidence of this benefit of gesture above and beyond spoken instruction alone (e.g., Dahl & Ludvigsen, 2014; Krauss, Dushay, Chen, & Rauscher, 1995).

When synthesizing across the literature, although gesture may be beneficial for both fact- and conceptual-learning, gesture's benefits may be more apparent in tasks of rule learning that require application across contexts**.** Because of gesture's ability to maintain its form across contexts and its ability to support generalizable rule-learning (e.g., Novack et al., 2014; Wakefield et

al., 2018b), gesture's effects may be in general stronger for conceptual-learning compared to fact-learning.

Beyond, this distinction between fact- and conceptual-learning tasks, previous work has explored the impact of gesture on learning in a variety of domains. The general consensus is that gesture supports learning across domains (e.g., mathematics: Cook et al., 2013; Congdon et al., 2017; Singer & Goldin-Meadow, 2005, symmetry: Valenzeno, et al., 2003, Piagetian conservation: Church et al., 2004; Ping & Goldin-Meadow, 2008, word learning: McGregor, Rohlfing, Bean, & Marschner, 2009; Rowe et al., 2013; Wakefield et al., 2018a, mental rotation: Levine, Goldin-Meadow, Carlson, & Hemani-Lopez, 2018; Ping, Ratliff, Hickey, & Levine, 2011, moral reasoning: Beaudoin-Ryan & Goldin-Meadow, 2014). Incorporating gesture during instruction has been found to support STEM learning, including biology (e.g., Colliot & Jamet, 2018), chemistry (e.g., Steiff, Lira, & Scopelitis, 2016), physics (e.g., Carlson et al., 2014), and mathematics, from relatively simple concepts like math equivalence (e.g., Singer & Goldin-Meadow, 2005) to more complex topics of ANOVA (e.g., Rueckert, Church, Avila, & Trejo, 2017). Gesture has also been shown to support tasks of language comprehension, including language learning tasks of tone and pitch categorization (e.g., Morett & Chang, 2015; Zhen, Van Hedger, Heald, Goldin-Meadow, Tian, 2019) and word or vowel learning (e.g., Flack & Horst, 2017; Kelly et al., 2009), to comprehension tasks of narrative or story recall (e.g., Beattie & Shovelton, 1999; Llanes-Coromina, Vilà-Giménez, Kushch, Borràs-Comes, & Prieto, 2018; Macoun & Sweller, 2018). Furthermore, gesture has been shown to support spatial reasoning, including tasks of conservation (e.g., Ping & Goldin-Meadow, 2008), comprehension of palindromes (e.g., Wakefield

& James, 2015), to navigation and route recall (e.g., Austin & Sweller, 2014; van Wermeskerken, Fijan, Eielts, & Pouw, 2016).

However, even with the many studies that suggest gesture benefits learning across domains, this is not always the case. For example, non-significant findings of gesture have been found when learning scientific concepts, including understanding how to solve a water distribution paradigm (Ouwehand, van Gog, & Paas, 2015), the mechanics of a lever (Pouw, van Gog, Zwaan, & Paas, 2016), or the physics of lightening (Davis & Vincent, 2019), to name a few. Similarly, the effects of gesture are not always evident at posttest measures of mathematics learning (e.g., Congdon et al., 2017; Yeo, Ledesma, Nathan, Alibali, & Church, 2017), language learning (e.g., Hirata & Kelly, 2010; Mavilidi, Okely, Chandler, Cliff, & Paas, 2015), narrative comprehension (e.g., Dahl & Ludvigsen, 2014; Cutica & Cucciarelli, 2008), or spatial reasoning (e.g., Austin, Sweller, & Van Bergen, 2018), including analogical reasoning (Study 1).

In line with the assumptions of gesture's effect for fact- vs conceptual-learning, we may also anticipate that gesture is more beneficial for domains that require generalization of rule-learning. For example, gesture may be more beneficial for domains such as mathematics and sciences, rather than language comprehension tasks, that often require more rote memorization. However, as with other factors explored here, demonstrating that gesture is important for a variety of domains is just as important for educators.

The third novel potential moderator that we explore in this meta-analysis is the timing of assessment. Many studies assess learning either immediately following an instructional session or after a delay. The majority of studies discussed so far in this chapter assess learning immediately after a training session. However, often the positive effects of gesture are more prominent

at a follow-up measure, from 24 hours (Cook et al., 2013) to 4 weeks (Congdon et al., 2017) after initial training, rather than during initial measures of performance. Cook and colleagues (2013) found that whereas speech+gesture instruction supports better post-training performance at mathematical equivalence than speech alone, the difference between the two types of instruction was more prominent after 24 hours. Whereas children in the speech-alone condition performed consistently over time following instruction, children in the speech+gesture condition continued to improve. In more recent work, Congdon and colleagues (2017) found that children who viewed speech+gesture did not perform better at immediate post-training measures than those who viewed speech-alone, but there was a significant impact of gesture by 24 hours and an even greater impact by 4 weeks after initial training.

In light of these findings, we assume that whereas gesture may or may not demonstrate a general effect of learning immediately following instruction, the effect of gesture may become more prominent over time. This could be due to consolidation of learning after training or perhaps gestures are more memorable over time when instructional supports are farther removed from the measure of learning.

Resolving Discrepancies in Previous Meta-Analyses

A secondary aim of the present study was to resolve discrepancies between the findings of the two previous meta-analyses: Hostetter (2011) and Dargue et al., (2019). First, we ask whether the effect of gesture on learning is impacted by the redundancy of gesture with accompanying speech. Numerous studies suggest that we notice information conveyed in non-redundant gestures and readily incorporate that information into our understanding of the speaker's message (Broaders & Goldin-Meadow, 2010; Goldin-Meadow & Sandhofer, 1999; McNeill,

Cassell, & McCullough, 1994), which can help clarify ambiguous speech (Kelly, 2001). However, in comparison to a lack of gestures, the use of redundant gestures has also been found to be beneficial for speech comprehension (Dargue & Sweller, 2018b). While Dargue and colleagues' (2019) meta-analysis did not find a differential effect of redundancy across the field, Hostetter (2011) found that non-redundant gestures have a larger effect on communication than redundant gestures. If gesture is helpful because it can provide additional information to speech and is easily integrated with accompanying speech (Kendon, 1986; McNeill, 1992), we anticipate that the redundancy of gesture will moderate the impact of gesture. However, it may also be the case that any use of gesture, redundant or not, is advantageous compared to instruction without gesture, and therefore, there may not be a significant effect of redundancy.

Second, we ask whether the effect of gesture depends on the learner's age. Categorizing studies into five age groups (preschool children, primary school children, adolescents, young adults, and older adults), Dargue and colleagues (2019) found a beneficial effect of gesture on comprehension for all groups except adolescents[2]. However, learner's age did not significantly moderate gesture's effects. In contrast, Hostetter (2011) categorized studies into two groups based on the sample age, and found that gesture benefits children (participants 12 years old or younger) more than adults (all other studies). This advantage of gesture for children is in line with previous work suggesting that gestures can help ground abstract or ambiguous concepts (e.g., Alibali & Nathan, 2007; Ping & Goldin-Meadow, 2008), can orient children's attention (e.g., Wakefield et al., 2018b), and do not tax children's limited cognitive processing capacity (e.g., Goldin-Meadow et al., 2001; Ping & Goldin-Meadow, 2008). If gesture is particularly

[2] Although the five age groups were used in their analyses, only 2 studies were categorized as *older adult* samples and thus, the results were not reported (Valentine et al., 2010).

helpful in clarifying ambiguous speech, it may be a useful tool for instructing children, who likely encounter more ambiguous content during instruction than adults.

However, previous work has also found non-significant effects of gesture for children, such that all children do not benefit equally from instruction with gesture (e.g., Broaders, Cook, Mitchell, & Goldin-Meadow, 2007; Church et al., 2004; Congdon et al., 2018; Wakefield & James, 2015). For example, McNeil and colleagues (2000) found that, compared to kindergarten-age children, preschool-age children's language comprehension benefited more from reinforcing gestures. Church, Kelly, & Lynch (2000**)** suggest a U-shaped curve across development, that represents the behavioral impact of gesture on young children, older children, and adults, where gesture helps some children more than others. Similar contradictions occur in the field when assessing the impact of gesture for adults. Although adults can use others' gestures to understand the intended message (e.g., Sueyoshi & Hardison, 2005), and producing gestures can facilitate encoding of spatial information and retrieval of rote-memory tasks (Goldin-Meadow et al., 2001; So, Shum, & Wong, 2015), non-significant effects of gesture are also evident in the field (e.g., Beattie & Shovelton, 1999; Driskell & Radtke, 2003). Therefore, it is likely that the benefits of gesture for children are greater, if not equal to, the benefits for adults, yet not all children may show a distinct advantage for speech+gesture compared to speech-alone. For example, as seen in Study 2, past a certain point in development children can benefit from instruction with and without gesture.

The discrepancies between these two meta-analyses could be due to a number of factors: The two analyses not only use different parameters for defining their moderators and sample of

studies, but Dargue and colleagues (2019) also had access to an additional decade of gesture-for-learning literature.

Attempts to Replicate Previous Meta-Analysis Findings

And finally, there are two theoretically important factors that vary across the gesture-for-learning literature that need to be further explored. Dargue and colleagues (2019) found that whereas gesture type does not moderate gesture's effects for comprehension, the ways in which gesture is experienced during instruction (produced or observed by the learner) does have differential effects. Specifically, Dargue and colleagues found that when gestures were produced by the learner themselves they were better able to comprehend spoken instruction compared to when the learner observed others' gestures. Gesture type and experience of gesture (produced vs observed) are two methodological factors that vary widely across the literature, such that being able to replicate Dargue's findings would provide additional evidence that gesture's effect is differentially (or not) impacted by these characteristics of the learning context.

As discussed in Chapter 1, McNeill (1992) classified gestures in four ways: iconic, metaphoric, deictic, and beat gestures, each varying in their representational nature. Briefly, representational gestures are hand movements that represent information through their form or trajectory. This category of gestures encompasses iconic, metaphoric, and deictic gestures, which represent a concrete action or object, an abstract idea, or point to a referent, respectively. In contrast, beat gestures are rhythmic movements of the hands that accompany speech (McNeil et al., 2000) that have no semantic relation to the content of the speech. Rather these gestures function to emphasize components of speech by highlighting a word or phrase to focus a listener's attention on important information (e.g., Biau & Soto-Faraco, 2013; Holle et al., 2012).

Previous work has found beneficial effects of each of these gesture types. For example, requiring children to perform deictic pointing gestures while learning a new concept, such as mathematical equivalence, helped children retain the knowledge they had gained during instruction (e.g., Cook et al., 2008). Similarly, asking children to observe beat gestures that highlight prominent prosody during narrative recall tasks helped children remember more words when exposed to prominence in both speech and gesture, rather than prominence in speech alone (Llanes-Coromina et al., 2018). Previous research further suggests that the semantic relatedness of a gesture with accompanying speech differentially impacts gesture's effect (Dargue & Sweller, 2018a). Specifically, the more semantically related a gesture is with speech, the more beneficial it may be to comprehension and learning (Straube, Green, Weis, Chatterjee, & Kircher, 2009). To support this suggestion, several studies have found that observing iconic gestures during instruction are more beneficial for learning compared to observing no gestures. (e.g., Beattie & Shovelton, 1999; Aussems & Kita, 2019; Church, Garber, & Rogalski, 2007). However, the benefit of iconic gestures is not found consistently across the literature. For example, Macedonia and colleagues (2019) found that learners perform equally well after instruction with and without iconic gestures (e.g., Macedonia, Repetto, Ischebeck, & Mueller, 2019; Kelly & Lee, 2012). Furthermore, although the research on the impact of metaphoric gestures is limited compared to the research on iconic gestures, both significant and non-significant effects of metaphoric gestures have been found (e.g., Yuan, Gonzalez-Fuente, Baills, & Preito, 2019; Zheng, Hirata, & Kelly, 2018). The difference between these studies may be the semantic relatedness between the gesture and speech; Perhaps non-significant findings are a consequence of the gesture being insufficiently semantically related to the content of the accompanying speech.

Dargue and colleagues (2019) found that although there was no difference between gesture types, all types of gestures were found to benefit comprehension. In the present meta-analysis we aim to replicate this finding with a different sample of studies. Although we may replicate Dargue's findings, there is reason to anticipate that gesture's impact depends on the gesture type used during instruction. Previous work points to gesture's semantic relatedness with speech as the driving cause for gesture's benefits. Therefore, gestures that are tightly coupled with speech may be more beneficial than gestures that contain less semantically important information. If this is the case, iconic and metaphoric gestures may provide an added advantage over beat gestures, or even perhaps deictic gestures, that are less semantically related to speech than iconic or metaphoric gestures. Importantly, a lack of difference between gesture types still has implications for the field – indicating that despite its form, incorporating gesture during instruction is beneficial for learners.

The final moderator of interest to the present meta-analysis is the experience of gesture. Dargue and colleagues (2019) found that both observing and producing gestures was beneficial for comprehension, but producing gestures has an added benefit. This finding is line with previous research. For example, Goldin-Meadow and colleagues (2012) found that young children asked to produce a gesture during a mental transformation task outperformed children who were asked to only observe an experimenter perform the gesture. Producing gestures is said to benefit comprehension and learning because it reduces the learner's cognitive load and allows for more resources to be allocated to the task at hand (Cook, Yip, & Goldin-Meadow, 2010). However, the research is ambiguous; Although numerous studies have found that producing gestures along with speech benefits learning (e.g., mental rotation: Chu & Kita, 2011), mathematics: Goldin-

Meadow, Cook, & Mitchell, 2009), other studies have found no benefit of producing gestures (e.g., Alibali, Spencer, Knox, & Kita, 2011; Lajevardi, Narang, Marcus, & Ayres, 2017).

The difference between observing and producing gestures may be driven by the role gesture plays in these two experiences of learning. Perhaps observing others' gestures during instruction functions differently for the learner than producing gestures. That is, processing both the visual and auditory information provided by an instructor may put strain on the learner's cognitive load, whereas off-loading cognitive effort onto one's own hands during learning may free up resources for better comprehension. With the present meta-analyses we want to provide further evidence for these suggestions. Although the present analysis cannot speak to the underlying mechanisms, we aim to either replicate and extend these findings with a different sample, or provide contrary evidence which would call for future investigations of this question. If we replicate Dargue's findings we can lend further support to the claim that producing gestures during learning provides an added benefit.

Present Study

Whereas numerous studies conclude that gesture accompanying speech instruction is beneficial for children and adults' learning, others do not find significant effects of gesture. This suggests that although the consensus is that gesture is useful for communication and comprehension, there is variability to *when* gesture is beneficial. Perhaps individual differences across children or situational factors in the context in which gesture is used affects its impact on the learner. The results of Studies 1 and 2 are an example of conflicting findings within a particular domain: Study 1 suggests that gesture supports analogical reasoning during training, but that benefit does

not extend to post-training measures, and Study 2 suggests that the benefits of gesture instruction are dependent on the learner's age and associated cognitive profile.

The aim of the current meta-analysis is to understand the variability across the gesture-for-learning literature. To build on the previous attempts to synthesize across the literature and identify factors that predict whether or not gesture facilitates learning, we will: (1) consider theoretically important factors that have been neglected previously, (2) resolve discrepancies in the findings of the two previous meta-analyses (Dargue et al., 2019; Hostetter, 2011), and (3) attempt to replicate previous findings of Dargue et al. (2019). By replicating previous findings using a sample that includes studies in both previous meta-analyses and those not included in previous work, we can provide further support for the role of particular methodological variations on the effect of gesture. In general, we expect significant variability in the sample due to methodological variations across the gesture-for-learning literature. This hypothesis is supported by both previous meta-analyses, which find that across studies gesture benefits communication and comprehension.

In summary, the current meta-analysis aims to synthesize existing work on the effects of gesture for learning to understand under what conditions gestures are beneficial. This meta-analysis addresses the following four questions:

1. Across studies, does the presence of gesture accompanying spoken instruction benefit learning to a greater extent than when the spoken instruction is not accompanied by gesture?
2. Across studies, is speech+gesture instruction, compared to speech-alone instruction, beneficial for:

a. Learning of fact or conceptual material?
b. Learning of mathematics, spatial reasoning, language, narrative comprehension, sciences, or other types of material?
c. Learning that occurs immediately following instruction or after a delay?
d. Learning when gesture provides additional or redundant information to the accompanying speech?
e. Children or adults' learning?
f. Learning when gesture is beat, iconic, metaphoric, deictic, or mixed?
g. Learning when gesture is observed or produced by the learner?

3. Do the factors listed above significantly moderate the effect that gesture has on learning?
4. When considering these factors together, does some combination of them predict the effect gesture has on learning?

Method

The protocol for this systematic review was registered with OSF Registries online (DOI: 10.17605/OSF.IO/29X6Z). The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standards were followed.

Selection Criteria

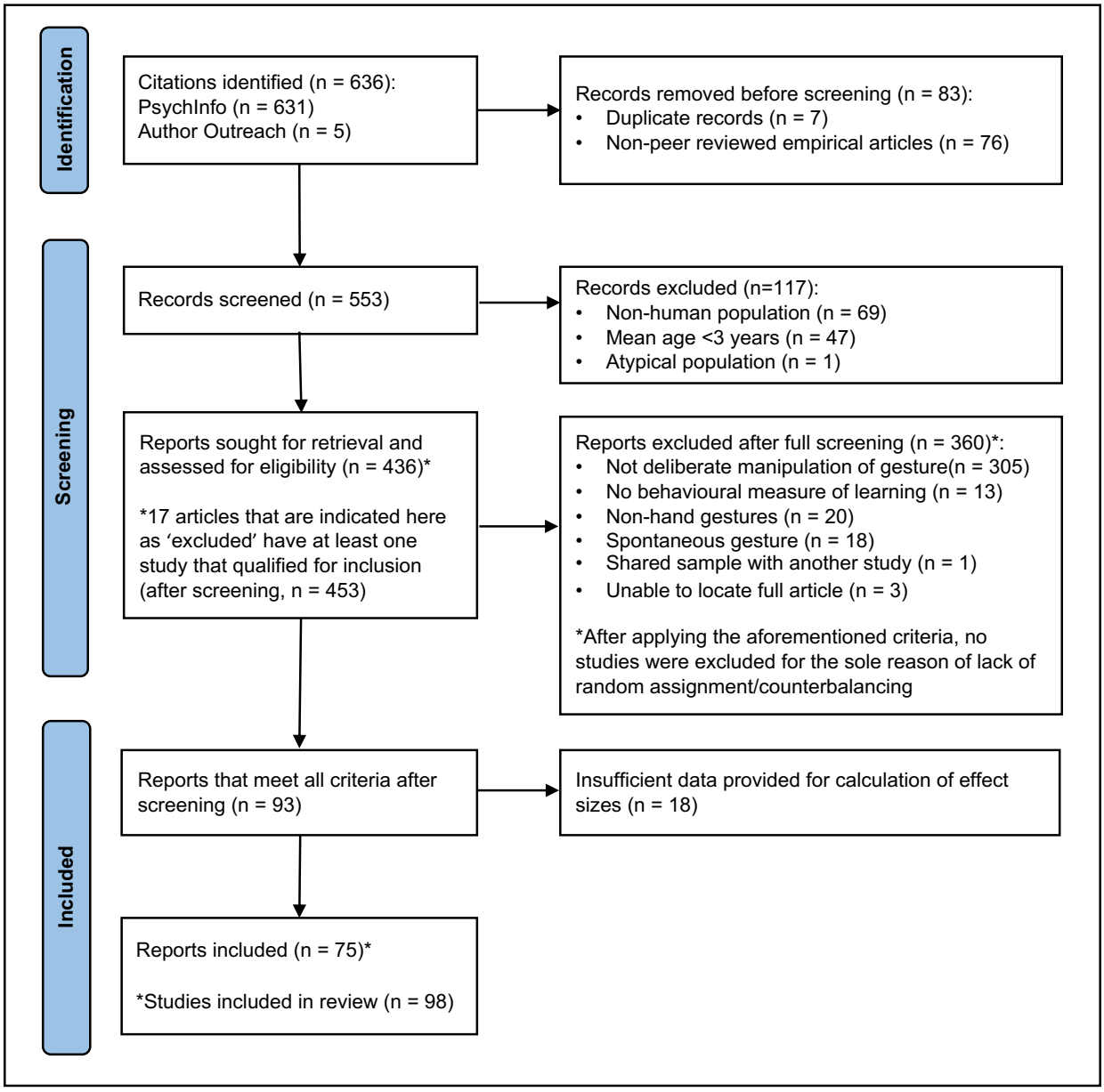
Search and review strategy. Relevant studies were identified through a search of PsycINFO database. The key word *gesture* was crossed with *learn**, NOT *infan*,* NOT *disab*,* NOT *autism,* NOT *damage,* NOT *apraxia, and* NOT *aphasia*. This set of key word limiters helped ensure that studies using samples with developmental abnormalities were excluded from the current review. Only peer-reviewed, published articles were considered to reduce bias and

poor study quality associated with the inclusion of non-peer-reviewed data (e.g., conference proceedings or dissertations). The primary author performed the initial eligibility assessment of each article. In the event of ambiguity, articles were assessed by both authors to determine whether a given study was eligible for inclusion or not. By strictly adhering to these eligibility criteria, we reduced the possible impact of poor-quality empirical studies on the results (see Figure 10 for PRISMA flowchart).

Inclusion criteria. To be included in the analysis, studies had to meet nine criteria. The first three criteria were used during the initial database search to review the title and abstract. If the title and abstract met the three initial screening criteria, a further screening of the full text was conducted considering the remaining six criteria. For articles in which inclusion for further screening could not be determined from the title and abstract, the full text was reviewed to address the remaining criteria. The final sample of studies meets these nine criteria.

For the initial screening, the article had to include a human sample with a mean age over 3 years. This accounted for two exclusion criteria: non-human sample and sample with mean age < 3 years. While prior research has examined the impact of gesture on infant and nonhuman species (e.g., Agnetta, Hare, & Tomasello, 2000; Capone & McGregor, 2005; Rowe, Ozçaliskan, & Goldin-Meadow, 2008), these functions of gesture are beyond the scope of the current investigation. Additionally, studies were excluded that used atypical populations. Because including samples with developmental or acquired disorders could confound the results, we limited our analysis to typically developing populations or populations that did not have neurotypical histories.

Figure 10. PRISMA flowchart of screening process for article selection



After a review of titles and abstracts, the resulting articles were further screened using six additional criteria. First, studies were excluded that measured or manipulated spontaneous use of gesture. That is, often the presentation of gesture to the learner is not controlled, such that the individual presenting the instruction is not told to provide their instruction (speech or gesture) in

a particular way. For example, Driskell & Radtke (2003) asked one set of participants to convey clues about pictures on cards to another set of participants who were asked to guess the identity of the picture. The first set of participants were only given instructions as to whether or not they could gesture when communicating clues, but not told *how to use* gesture or what to say verbally. Because the use of gesture is not controlled in this type of study, any comparison to studies in which gesture is scripted would be uninterpretable. Whereas studies that provide spontaneous gesture cannot account for variability across participants' experiences, studies that provide scripted gesture ensure that all participants receive similar or identical instruction.

Second, the availability of speaker or learner's gesture had to be deliberately manipulated so that there was one condition in which speech and gesture were produced together (treatment condition) and a second condition in which speech was presented without gesture (control condition). This allows for the examination of only effect sizes that contrast the effectiveness of speech with gesture to the effectiveness of speech alone. Importantly, conditions need to be identical except for the addition of gesture in the treatment condition. In general, this criterion excluded three types of studies. Observational studies were excluded because the inclusion of gesture was not deliberately manipulated. Studies that compared the effectiveness of different types of gestures (e.g., beat vs iconic gestures) were excluded if they did not include a control condition in which speech was presented without gesture. And lastly, studies that compared two conditions manipulating some factor other than gesture (e.g., speech with gestures vs speech with pictures) were excluded because there was not a control condition identical to the treatment condition aside from the inclusion of gesture.

Third, the treatment condition had to include hand gestures. Thus, studies that manipulated the presence of articulatory or facial gestures would be excluded (e.g., facial gestures provided through video: Hardison, 2003, or images of facial gestures provided on cards when producing a phoneme: Boyer & Ehri, 2011).

And finally[3], the dependent variable had to represent a behavioral measure of learning (e.g., gains from pretest to posttest). This criterion excluded studies for two primary reasons: studies that examined neurophysiological measures (e.g., electroencephalography or functional Magnetic Resonance Imaging [fMRI]) in the absence of a behavioral measure, and studies that examined the use of referential gestures in the absence of learning measures.

Assessment of Study Quality

The quality of included studies was assessed to identify any potential sources of bias that may impact the results of this systematic review and the quality of the existing research investigating the impact of gesture on learning. For the current review, the Critical Appraisal Skills Programme Checklist (CASP)[4] was used to assess the quality of studies. Specifically, the 11-item Randomized Control Trial (RCT) Standard Checklist was used for studies implementing a between-subjects design. There are many checklists used in previous literature for assessing quality of study design, but this measure is the most appropriate checklist for the current sample of studies using between-subjects designs. Although none of the studies in the current sample are true randomized control trial studies, in that they do not adhere to blinded random assignment,

[3] Random assignment to condition was also an exclusionary criterion. However, after other exclusionary criteria were applied, no studies were excluded for the this reason alone.

[4] The CASP is a part of the Oxford Centre for Triple Value Healthcare Ltd. (3V) portfolio, which provides learning and development resources to support the development of critical appraisal skills (http://www.casp-uk.net/; Singh, 2013; Taylor et al., 2000).

they do randomly assign participants to conditions to ensure matched samples or to fill a quota for each condition.

For the sample of studies using within-subjects designs we adapted the RCT Checklist to match the structure of these studies. To implement this checklist, the phrasing of the questionnaire had to be altered so that it was appropriate for a within- vs between-subjects design. For example, questions about whether participants were randomly assigned to conditions on the original version of the RCT checklist were adapted to consider whether condition presentation was counterbalanced or randomized. One item of the original version of the checklist was specific to RCT (i.e., Apart from the experimental intervention, did each study group receive the same level of care?). This question was not applicable and was removed from the checklist. Therefore, for the sample of within-subjects designs, studies were assessed against 10 of the 11 scale items.

Thus, two tools were implemented – one for between-subjects designs and one for within-subjects designs. For each tool, a global quality score was calculated, with higher scores denoting higher-quality studies. To assess the quality of studies implementing a between-subjects design (n = 71) the 11-item RCT Standard checklist was used, and to assess the quality of studies implementing a within-subjects design (n = 27) the adapted version of the RCT checklist was used. A global quality score of 22 and 20 points, respectively, could be allocated to each article, with 2 points being allocated for each item addressed by the article, 1 point for items that might have been addressed (but were unclear), and 0 points for each item not addressed.

To describe the risk of bias across articles included in this systematic review (consistent with the PRISMA statement), the National Health and Medical Research Council (NHMRC)

Levels of Evidence Guidelines (Merlin, Weston, & Tooher, 2009[5]; available at https://www.mja.com.au/) was used. The NHMRC was a convenient way of summarizing study designs according to their generally perceived capacity to minimize or eliminate bias in the effect being measured.

Data Extraction

Data relating to the study (e.g., author, publication year), sample characteristics (e.g., sample age, number of participants), study characteristics (e.g., between- or within-subjects design, type of gesture, domain, timing of assessment; detailed below), and result characteristics (e.g., means, standard deviations, *F*-, *t*-, or *p*-statistics as related to measures of learning) were extracted from each study. If any study reported conducting analyses on relevant variables but did not publish the data, the corresponding author was contacted and the data were requested for inclusion. Authors were contacted for data at the study level, meaning data was requested for individual studies within articles. Authors were contacted for 57 studies (33 authors). Relevant data was received for 36 studies and were subsequently included in the present meta-analysis.

Each study was classified according to the following study characteristics to identify whether significant variation in effect sizes could be explained by these variables. That is, we extracted information relevant to determining the role of these characteristics as potential moderators. Characteristics of each study were categorized as characteristics of the learner, of the task, and of the gesture.

[5] These guidelines are the gold standard for judging "levels of evidence" for the purposes of health technology assessment and clinical practice guideline development (Merlin et al., 2009).

Characteristics of the learner. Age of learner was categorized in two ways to compare the present meta-analysis with the previous meta-analyses. First, for a comparison to Hostetter (2011), learners 12 years or younger were categorized as *children* and all other learners as *adults*. Second, for a comparison to Dargue and colleagues (2019), studies were classified into five categories based sample age and description: *preschool* children, *elementary* school children, *adolescents* (students in high school), *young adults* (university students and adults over 18 years of age), and *older adults* (adults between 60-80 years old).

Characteristics of the gesture. The type of gesture used during instruction was classified as one of five categories following the classification system developed by McNeill (1992): beat, iconic, metaphoric, deictic, or mixed. *Beat* gestures are movements emphasizing the prosody or structure of speech without conveying semantic information. *Iconic* gestures are movements that provide information semantically related to the content in the accompanying speech. These gestures often depict a concrete action, event, or object with the motion or shape of the hands or the placement of something in space when the referent is not physically present. *Metaphoric* gestures can also provide information semantically related to speech, but are movements that represent an abstract or concrete metaphor for a concept. These gestures can indicate a spatial location for an abstract idea or even depict the structure or shape of a sound (e.g., a pitch, length of vowels, or movement of breath). *Deictic* gestures are pointing gestures that indicate objects, events, or directions. And finally studies that use a combination of different types of gestures were categorized as *mixed* gestures.

Studies were further categorized dependent on how gestures were experienced by the learner: whether the gestures were produced by the learner themselves during training or pro-

duced by the experimenter or instructor. Studies were categorized as *produced* if the learner themselves produced the gesture in the treatment condition. Often when the manipulation of gesture was produced by the learner it was first illustrated by an experimenter for the learner to mimic. Studies were categorized as *observed* if the learner only watched the experimenter or instructor produce gestures through live or pre-recorded video instruction.

Studies were further categorized according to whether or not the gestures used in the treatment condition were redundant with the accompanying speech. Studies were categorized as *redundant* if the gestures and speech provided the same information. For example, during a lesson on symmetry, Valenzeno and colleagues (2003) used deictic and tracing gestures to delineate the center of a symmetrical shape to facilitate comparisons between the two sides of the shape. In this case, gestures mirrored the spoken instruction provided. Studies were categorized as *non-redundant* if the gestures added additional information to the speech. For example, during a lesson on mathematical equivalence, Wakefield and colleagues (2018b) used deictic and iconic gestures to indicate that two sides of an equation are equal and how to solve for a missing addend (e.g., 5 + 6 + 3 = __ + 3). However, as the instructor said '*I want to make one side equal to the other side*', rather than pointing at all numbers on each side of the equation, they specifically pointed to the two numbers that can be added to find the missing addend.

Characteristics of the task. Studies were classified as according to the type of learning used at assessment. Studies were categorized as *fact-learning* if the assessment was a measure of rote memory and categorized as *conceptual-learning* if the assessment was a measure of one's ability to learn and apply rules or strategies for solving problems (for additional discussion of these learning types refer to page 86). In additional to type of learning, studies were further clas-

sified according to the domain of the task that was trained and learned. Studies were categorized as *mathematics* (e.g., math equivalence, fractions, geometry), *language learning* (e.g., grammar, word learning), *spatial reasoning* (e.g., conservation, symmetry, mental rotation, navigation), *narrative comprehension* (e.g., event, story, or phrase recall), *science* (e.g., biology, medicine, chemistry, physics, environmental science), and *other* (e.g., singing, history, computers, software).

And lastly, studies were categorized according to the timing of the assessment. Studies were categorized as *immediate* if the measure of learning was a posttest immediately following the training or instruction (e.g., occurring on the same day as the training) and categorized as *delay* if there was an interval of at least 24 hours between training and posttest (e.g., 24 hours, 1 month). Typically studies that incorporated delayed assessments also included immediate posttest assessments. Therefore, it is important to note that direct comparisons between these two categories (e.g., asking if timing of assessment moderates the effect of gesture) are likely using similar samples and results must be interpreted with caution.

See Appendix A for the full list of studies retrieved, along with each article's classification according to the criteria listed above.

Reliability

All coding and data extraction was done by one coder. To assess reliability of data extraction, a second coder independently coded 19 randomly selected samples (approximately 19% of the data). Cohen's κ was used to evaluate agreement on codes related to study characteristics as these classifications are tested as potential moderators. For age group (both age categorization schemes), $\kappa = 1.00$, $p < .005$, for type of learning, $\kappa = 0.86$, $p < .005$, for domain, $\kappa = 1.00$, p

< .005, for type of gesture, $\kappa = 0.79$, $p < .005$, for timing of assessment, $\kappa = 1.00$, $p < .005$, for redundancy of gesture, $\kappa = 0.77$, $p < .001$, and experience of gesture, $\kappa = 1.00$, $p < .005$. In all cases, the codes of the original coder were used.

Estimation of Effect Sizes

Calculating effect sizes. Cohen's *d*, a standardized estimate of effect size, was calculated for each unique sample in each study. Whenever possible, Cohen's *d* was computed directly from relevant means and standard deviations by taking the difference between means and dividing the difference by the pooled standard deviation (*SD*). In the case where means and standard errors (*SEs*) were given, *SEs* were transformed into *SDs* so that the pooled *SD* could be calculated.

Study designs were categorized as between- or within-subjects designs. This information was recorded to ensure the appropriate method was followed for estimating study effect sizes. For between-subjects studies that did not report relevant means and *SDs* or *SEs*, Cohen's *d* was calculated directly from *t, F, β, or* X^2 values using formulae suggested by Cumming (2012), Lipsey and Wilson (2001), and Rosenthal (1995). Some between-subjects studies reported the mean proportion of participants who met some performance criteria in each condition (e.g., the proportion of participants who learned or succeed on the task). In these cases, Cohen's *d* was calculated by converting each proportion to a standardized *z* score by using the probit transformation recommended by Lipsey and Wilson (2001) and then taking the difference between these two *z* scores. For studies that did not provide the necessary values to calculate Cohen's *d* or did not provide values specific to the comparison of interest, we attempted to contact the authors. For example, occasionally studies include more than the two conditions of interest to the present

meta-analysis, yet only reported ANOVA values examining all conditions in the study without post hoc comparisons between pairs of conditions. In these cases, it was necessary to contact authors for these details.

Effect sizes for within-subjects designs, however, cannot be accurately estimated from the aforementioned values without also knowing the *SD* of the difference scores between conditions and the correlation between scores of the two conditions (Dunlap, Cortina, Vaslow, & Burke, 1996). This information was not provided in any of the within-subjects designs under consideration. Although the *SD* of difference scores and the correlation can be calculated from the raw by-participant data, only one within-subjects study provided the raw data within the article. Thus, we attempted to contact the authors of the remaining studies to obtain the relevant raw data. Studies that did not provide the necessary values to calculate an effect size or those for which authors did not respond were excluded from further analysis[6].

After calculating Cohen's *d* for all samples, an adjustment recommended by Hedges (1981) was applied to calculate unbiased estimates of Cohen's *d*. This adjustment is particularly important when sample sizes are small, such that there is an increased risk of overestimation (Cumming, 2012). Unbiased estimates of Cohen's *d* on average will not underestimate or overestimate the parameter. Finally, estimates of *SE* around unbiased Cohen's *d* were calculated separately for between- and within-subjects studies, using the formulae recommended by Lipsey and Wilson (2001).

Selecting Conditions of Interest within Studies. As mentioned, some studies contained more than two relevant conditions. In all cases, the control condition that most closely matched

[6] Of the 33 authors contacted (for 57 samples), 23 responded and were able to provide the missing information for 36 samples.

the presentation of the speech and gesture condition was chosen. For example, a control condition in which speech is presented alone would be preferred over a speech condition with head nods, if the experimental condition only presented speech with gestures (Zheng et al., 2018). In some instances, multiple experimental conditions were included in the same study. Because including two comparisons from the same sample would violate the independence of observations, a decision had to be made for each study about *which* two conditions would be used to calculate an effect size. For example, if speech was presented with a matching gesture and separately with a mismatching gesture, the mismatching gesture condition would be favored. This was the case if and only if the mismatching gesture condition provided correct information and if the same information was provided verbally across conditions (e.g., Dargue & Sweller, 2018b; Singer & Goldin-Meadow, 2005). However, if the mismatching gesture provided inaccurate information or verbal content was not identical across conditions, then the matching gesture condition was favored. Similarly, if a study included two experimental conditions in which gestures were observed and produced by the learner, the latter was favored (e.g., Krönke, Mueller, Friederici, & Obrig, 2013; Stieff et al., 2016). And lastly, if multiple experimental conditions provided different forms of gesture, the gesture that provided more meaningful information to the instruction was chosen. For example, a condition in which speech was accompanied by an iconic or semantic gesture was chosen over a condition using an arbitrary gesture (e.g., Lüke & Ritterfeld, 2014).

Studies also often used multiple types of assessments. In these instances, the assessment that demonstrated the best understanding of the content was chosen. Often this was the more difficult task in which success would indicate better comprehension of the trained content. For example, free recall tasks were favored over recognition or cued recall tasks (e.g., Kushch, Igualа-

da, & Prieto, 2018). Further, where possible, samples were also split according to characteristics of interest. For example, samples were split between children and adults (e.g., Kartalkanat & Goksun, 2020) or between native and foreign language speakers (e.g., Dahl & Ludvigsen, 2014) if participant characteristics and results were provided in the article separately for the two groups.

Results

Overall, 75 articles (98 samples) were included in the current systematic review (see Appendix A). Publication dates ranged from 1995-2021. Before investigating our four questions, the quality of studies in the sample was evaluated against field standards. This helped determine our confidence level in the sample estimate and whether caution is warranted when interpreting the present findings due to any type of bias (e.g., low quality studies reduce confidence in meta-analytic findings).

Subsequently, analyses of effect sizes were conducted. We began by asking if the presence of gesture accompanying spoken instruction (speech+gesture instruction) benefits learning to a greater extent than when spoken instruction is not accompanied by gesture (speech-alone instruction). This was tested by estimating a mean effect size and testing its ability to represent the sample. During this analysis we also asked if there is significant variability of effect sizes within the sample. If our results indicate that in general speech+gesture instruction benefits learning above and beyond speech-alone instruction, and that there is significant variability of effect sizes within the sample, we can then explore if certain methodological variations across studies account for this variability. This step of the investigation is addressed in two ways. First, we asked whether across studies, is speech+gesture instruction compared to speech-alone instruction differentially beneficial for learning under various contextual conditions (i.e., methodo-

logical variations). Second, we asked if these methodological variations significantly moderate the effect gesture has on learning. We then performed an exploratory analysis to investigate whether combinations of these methodological variations predict gesture's effect.

Finally, we determined the degree to which the sample, and subsequent analyses, may be impacted by publication bias. Publication bias can be due to a variety of factors, such as reporting bias against studies with small sample sizes or authors declining to submit because results go against the prevailing understanding of a phenomenon. Because publication bias cannot be comprehensively assessed through a single method, multiple approaches were used (Rücker, Carpenter, & Schwarzer, 2011).

Assessment of Study Quality

Each study that met the nine abovementioned inclusion criteria were evaluated for study quality. Although the CASP checklist includes a set of 11 items (10 for within-subjects designs) this assessment will focus on five key items that directly relate to study quality in this sample. Whereas the additional items also speak to quality, these 5 items directly relate to experimental designs with a non-clinical focus. For between-subject designs, discussion of quality relates to: (1) the use of random assignment to conditions, (2) whether all participants who initially participated were accounted for at the conclusion of the study, (3) whether personnel were blind to the condition participants were assigned to, (4) whether groups were treated equally aside from the experimental manipulation, and (5) whether the dependent variable was clearly specified. For within-subjects designs, rather than considering assignment to condition (item 1), discussion of study quality considered whether counterbalancing occurred to prevent order effects. Similarly, for within-subjects designs, we did not consider similarity of groups (item 4) given the lack of

separate intervention groups. The sample majority used between-subjects designs (n = 71), with CASP RCT ratings ranging from 8 to 15 points (M = 12.65; SD = 1.56, max score of 22). The remaining studies used within-subjects designs (n = 27), with adapted CASP RCT ratings ranging from 9 to 13 points (M = 11.11; SD = 1.07, max score of 20).

The majority of between-subjects studies explicitly stated use of random assignment with the exception of 4 studies that specified some kind of pseudo-assignment and 6 studies that did not explicitly mention random allocation. Similarly, all but two within-subjects studies explicitly stated that some form of counterbalancing was used. All studies accounted for all participants who participated initially. Either no loses were mentioned and the sample size was the same at the start and conclusion of the study, or clear reasons were given for any exclusion of participants from analysis (e.g., technological failure, experimenter error, failure to respond, or poor performance at pretest). Groups were treated equally in all between-subjects studies and the dependent variable of interest was clearly defined in all studies. However, studies rarely specified whether study personnel were blinded to condition, or if study personnel were blinded to condition during qualitative data coding. Only one study stated that experimenters were blind to hypotheses, and 8 studies specified that experimenters were blind to condition assignment when coding qualitative data. On occasion, studies specified that the experimenter delivered the training or watched the instructional videos with participants. In such instances, the experimenter would not be blind to condition assignment. Taken together, the biggest concern is the limited discussion of whether personnel were blind to condition. Because it was unclear in the articles whether this protocol was followed or not, we cannot criticize a lack of blinding, but rather en-

courage future work to address whether personnel were blind to condition during testing or when assessing performance.

To assess the clinical importance of findings across all included studies, the NHMRC Levels of Evidence Guidelines (Merlin et al., 2009) were used. This measure allows each individual study assessed to be given a level, ranging from level 1 (high quality; i.e. systematic review of RCTs) to level 4 (low quality; i.e. case series with pre-test/post-test outcomes). All studies, both between- and within-subjects designs were awarded a level 3 rating. Between-subjects designs were awarded a level 3 rather than a level 2 rating: Even though the majority of studies state that random assignment was used, they did not specify *how* participants were randomly allocated to conditions or ensure a double-blind protocol where both the experimenter and participant are blind to condition. Similarly, often random allocation was only implemented until all conditions were of equal size. Within-subjects designs were awarded a level 3 rather than a level 2 rating for similar reasons, including a lack of experimenters blind to condition presentation. Taken together, on a scale from *high* to *very low*, the body of evidence receives a grade of *low* (Grade C; Merlin et al., 2009). This grade reflects that our confidence in the effect estimate is limited, and the true effect may be different from the estimate of the effect. Consequently, the results described below should be interpreted with caution.

Analyses of Effect Sizes

Data was analyzed using R studio (version 1.2.1335). Cohen's *d* effect sizes were calculated from 98 unique samples from 75 studies, given that some of the articles included more than one study. Multiple studies from one article were included if they used unique samples of participants across the studies. Therefore, we do not only have independent studies but also independ-

ent samples of participants. An additional 11 Cohen's *d* effect sizes were calculated for measures of learning after a delay. These delay measures were used to examine the effect of timepoint on gesture's effects. Analyses examining the effect of timepoint should be interpreted with caution given that measures of delay come from studies that also included measures of immediate learning. Therefore, the subsample of studies with delay measures used the same sample of participants as their immediate posttest measure counterparts. For testing of all other potential moderators only the sample of immediate measures of learning were used.

The size of the effects was interpreted using the guidelines suggested by Cohen (1988), where $d = .2$ indicates a small effect size, $d = .5$ indicates a medium effect size, and $d = .8$ indicates a large effect size. Sample sizes ranged from 11 to 323 participants ($M = 68.61$, $SD = 51.40$), with a total of 6744 participants represented in the complete analysis. The effect sizes ranged from – 0.97 to 2.60, and 83 of the 98 effect sizes were positive.

Does the presence of gesture accompanying spoken instruction benefit learning to a greater extent than when spoken instruction is not accompanied by gesture? Before calculating the mean effect size and testing its significance to address this question, it is necessary to adopt either a fixed-effects or random-effects model (a detailed discussion of model assumptions can be found elsewhere: Lipsey & Wilson, 2001; Rosenthal, 1984; Schmidt, Oh, & Hayes, 2009). Although the a priori hypothesis (i.e., assuming significant variation across studies due to methodological variations) suggests the adoption of a random-effects model, it is important to confirm this hypothesis before testing for the effect of potential moderators. Here, potential moderators represent 7 different methodological variations across the sample. Briefly, a fixed-effects model assumes that the only source of error present in a given effect size is due to subject-level sam-

pling in that study, and that only random variation exists between studies. In contrast, a random-effects model assumes there is additional error in each effect size due to methodological variations across studies as well as from other sources. Because a random-effects model takes systematic heterogeneity into consideration in the calculation of weights, this model type generates more appropriate pooled estimates and variances of the pooled estimate than a fixed-effects model.

A standard test of heterogeneity (Cochrane's test; DerSimonian & Laird, 1986) was used to test for the presence of heterogeneity in our sample. In the current sample, Cochran's Q = 386.82, $p < .0001$, suggesting that the variability among effect sizes was not due to sampling error alone, but to systematic differences between studies. Therefore, adoption of a random-effects model in the current analysis is supported by both the a priori theoretical hypothesis and by the empirical heterogeneity of the studies.

Whereas Cochran's test of heterogeneity (Q) indicates the presence or absence of heterogeneity in a sample of effect sizes, it does not indicate the extent of heterogeneity. The I^2 index quantifies the degree of heterogeneity within the sample of effect sizes. Therefore, beyond Cochran's test of heterogeneity, I^2 was used to indicate the percentage of heterogeneity present between studies, with I^2 = 25% indicating a small amount of heterogeneity, I^2 = 50% indicating a medium amount of heterogeneity, and I^2 = 75% indicating a large amount of heterogeneity (Higgins & Thompson, 2002; Higgins, Thompson, Deeks, & Altman, 2003). A large amount of heterogeneity remained in the current meta-analysis (I^2 = 75.0%), with the Cochran's Q analysis explaining 25.0% of the total between-study variation. That is, the unexplained between-study

variance in effect sizes was greater than the explained variance. This finding is unsurprising given that the a priori hypotheses assumes an impact of different methodologies.

Under the random-effects model, the weighted unbiased mean effect size was calculated as .49 (SE = .06; 95% CI = .39, .59)[7], and is significantly greater than zero ($z = 9.73$, $p < .0001$). This result suggests that across studies, gestures have a medium, beneficial effect on learning.

Across studies, is speech+gesture instruction, compared to speech-alone instruction, beneficial for learning under different circumstances? Given that the a priori hypothesis is supported, we further explored the impact of individual potential moderators on the variability in the sample. Using a series of 7 stratified meta-analyses, we determined if methodological variations in the sample account for differential effects of gesture on learning, and examine if speech+gesture is more beneficial than speech-alone instruction under certain circumstances (a similar analytic approached has been used previously, e.g., Dargue et al., 2019). For example, when considering *type of learning*, which has two subgroups (fact-learning, conceptual-learning), a stratified meta-analysis asks if speech+gesture instruction is more beneficial for fact-learning than speech-alone instruction, and separately, asks if speech+gesture instruction is more beneficial for conceptual-learning than speech-alone instruction. Although this analytic approach does not allow for comparison across subgroups within predictor (fact vs conceptual), we can determine whether each subgroup has an effect size that significantly differs from zero. A subgroup

[7] Sensitivity analyses were performed to test whether individual studies had a greater than average effect on the overall weighted mean effect size. Given that all adjusted mean effect sizes are within the omnibus random effects model's confidence interval, none of the studies have a greater than average effect on the sample. Further, eleven outliers were identified (d = -0.97, 1.24, 1.46, 1.48, 1.69, 1.73, 1.81, 1.90, 2.14, 2.40, and 2.60). To determine whether the outliers had a significant impact on the results, analyses were repeated excluding these studies. Because removing the outliers did not change the omnibus results (overall weighted mean = .35, $Q = 195.45$, $p < .001$), we opted to report the original results. Together, these analyses indicate that neither removing each study individually from the sample nor removing outliers as a group impacted the overall findings. All results from the sensitivity analysis and results excluding outliers are available on request.

mean effect size (e.g., for fact-learning studies) that is significantly differently from zero indicates that within this subgroup, the experimental manipulation (speech+gesture vs speech-alone) significantly predicts learning. Stratifying by predictor not only allows exploration into the impact of gesture within each subgroup but also addresses how much heterogeneity exists within each subgroup using Cochran's Q and I^2. Specifically, Cochran's Q indicates whether there is significant heterogeneity within subgroups, and I^2 indicates the degree of residual heterogeneity that is unaccounted for by the predictor in question. A large degree of residual heterogeneity (i.e., a large I^2) indicates that the majority of the variability in effect sizes in not due to the predictor in question, but rather due to other methodological variations across the sample (i.e., more within-subgroup variability rather than between-subgroup variability).

Testing the impact of novel factors. ***Type of learning.*** The first stratification explored whether gestures benefit learning of fact or conceptual content. Of the 98 samples included, 23 investigated gesture's effect on fact-learning and 75 investigated gesture's effect on conceptual-learning. A random effects model was used for each learning type strata (Table 2). Here, strata refers to each subgroup of a predictor. The weighted mean effect sizes for both fact- and conceptual-learning subgroups were significantly greater than zero, suggesting that the impact of speech+gesture is greater than speech-alone instruction for both learning types. Specifically, results suggest that instruction using gesture has a medium beneficial effect on both fact-learning (d = .46) and conceptual-learning (d = .48).

Stratifying by learning type indicates that within both subgroups there is heterogeneity unaccounted for by the experimental manipulation (see Table 2), such that within studies assessing fact-learning there is 85.10% heterogeneity and within studies assessing conceptual-

learning there is 41.35% heterogeneity. Specifically, if there is 41.35% residual heterogeneity in the conceptual-learning subgroup, then roughly 58.65% of the variance in the subgroup is explained by the effect of gesture on learning compared to speech alone. Compared to the total between-study variance accounted for in the overall meta-analysis (omnibus random effects I^2 = 75.0%), fact-learning explains *less* heterogeneity than the omnibus model, and conceptual-learning explains *more* heterogeneity than the omnibus model. In other words, within the subgroup of studies assessing conceptual-learning, there is *less* variability in effect sizes compared the sample as a whole. Descriptively, this suggests that there is more consistency in the effect of gesture in the conceptual-learning subgroup, than the fact-learning subgroup, despite all other possible methodological variations that exist in that subgroup.

Further, stratification analysis not only allows us to explore the impact of gesture within individual subgroups, but also explore the impact of the predictor on sample variability. That is, we can interpret the omnibus stratification statistics to determine the amount of heterogeneity in the sample accounted for by the predictor and whether there is significant variability both between and within subgroups. Although learning type accounts for a large amount of heterogeneity in the current sample, with the stratification analysis accounting for 94.50% of the total between-study variation (I^2 = 5.50%), there is still significant residual heterogeneity unaccounted within learning types (Q_{within} = 386.82, p <.0001). The stratification analysis indicates that using gesture is in general more beneficial across types of learning than not using gesture during instruction. However, there is not a significant difference in heterogeneity between learning types ($Q_{between}$ = 0.01, p = .92).

Although these results suggest there may be a difference in heterogeneity between fact- and conceptual-learning subgroups, this does not tell us if there is a significant difference between subgroup mean effect sizes. Following the review of the stratified meta-analyses results, a metaregression will explicitly test whether the effect of gesture is moderated by learning type, and if subgroup mean effect sizes significantly differ.

Table 2. Cochran's Q and stratification test results for type of learning

	Heterogeneity			Stratified Pooled Estimates				
Stratum	*Q*	*p*-value	I^2	*M*	*CI* (95%)	*z*-score	*p*-value	n
Fact-learning	347.56	<.0001	85.10%	.46	.35, .58	7.80	<.0001	75
Conceptual-learning	39.26	0.01	41.35%	.48	.25, .70	4.17	<.0001	23

Note. CI = confidence interval. *M* refers to the pooled value of Cohen's *d* (effect size).

Timepoint. The second stratification explored whether gesture's effect depends on assessment timepoint. Rather than using the sample of 98 studies, for this stratification a combined sample of immediate and delayed studies was used. Of the 109 samples included in this analysis, 98 investigated gesture's effect on learning at immediate posttest and 11 investigated gesture's effect on learning at a delayed posttest. A random effects model was used for each timepoint strata (Table 3). The weighted mean effect sizes for both immediate and delay subgroups were significantly greater than zero, suggesting that the impact of speech+gesture is greater than speech-alone instruction at both timepoints. Specifically, results suggest that instruction using gesture has a medium beneficial effect on learning at immediate posttest ($d = .47$) and instruction using gesture has a large beneficial effect on learning at delayed posttest ($d = .61$).

Stratifying by timepoint indicates that within both subgroups there is heterogeneity unaccounted for by the experimental manipulation (see Table 3), such that within studies assessing immediate learning there is 80.65% heterogeneity and within studies assessing delayed learning there is 81.01% heterogeneity. Compared to the total between-study variance accounted for in the omnibus meta-analysis (I^2 = 75.0%), both timepoints explain *less* heterogeneity than the omnibus random effects model. In other words, within the individual subgroups, there is *more* variability in effect sizes compared to the sample as a whole, suggesting that timepoint alone does not account for a significant portion of the variability in gesture's effects, and gesture's effects likely vary due to other methodological variations in the sample.

Although timepoint accounts for a large amount of heterogeneity in the current sample, with the stratification analysis accounting for 94.29% of the total between-study variation (I^2 = 5.71%), there is still significant residual heterogeneity unaccounted for within timepoints (Q_{within} = 432.01, p <.0001). The stratification analysis indicates that gesture is in general more beneficial for learning than not using gesture during instruction across timepoints. However, there is not a significant difference in heterogeneity between timepoints ($Q_{between}$ = 1.21, p = .27). The potential differences in gesture's effect across timepoints will be explored further with a metaregression, when we ask if timepoint moderates gesture's effect on learning.

Table 3. Cochran's Q and stratification test results for timepoint

	Heterogeneity			Stratified Pooled Estimates				
Stratum	*Q*	*p*-value	I^2	*M*	*CI* (95%)	*z*-score	*p*-value	n
Immediate	386.82	<.0001	80.65%	.47	.36, .57	8.83	<.0001	98
Delayed	45.1824	<.0001	81.01%	.61	.31, .91	3.97	<.0001	11

Note. CI = confidence interval. *M* refers to the pooled value of Cohen's *d* (effect size).

Domain. The third stratification explored whether gesture's effect depends on the type of content (i.e., different domains) being learned. The domains included in this analyses were language learning, mathematics, narrative comprehension, spatial reasoning, science, and *other*, indicating additional domains that did not fit into the other five domain categories. Of the 98 samples included, 30 investigated gesture's effect on language learning, 9 investigated gesture's effect on mathematics learning, 25 investigated gesture's effect on narrative comprehension, 15 investigated gesture's effect on spatial reasoning learning, 13 investigated gesture's effect on science learning, and 6 investigated gesture's effect on learning in other domains (a mixed category). A random effects model was used for each domain type strata (Table 4). For samples that investigated language learning, mathematics, narrative comprehension, spatial reasoning, and science, the weighted mean effect sizes were significantly greater than zero, suggesting that the impact of speech+gesture is greater than speech-alone instruction for these five subgroups. Specifically, results suggest that instruction using gesture has a medium beneficial effect on language learning (d = .52) and narrative comprehension (d = .54), and instruction using gesture has a small to medium beneficial effect on mathematics learning (d = .43), spatial reasoning learning (d = .41), and science learning (d = .42). However, for samples that investigated learning in *other* domains, the weighted mean effect size was not significantly different from zero, suggesting that speech+gesture instruction does not have an advantage over speech-alone instruction in this mixed category.

Stratifying by domain type indicates that within subgroups of language learning, narrative comprehension, and spatial reasoning there is heterogeneity unaccounted for by the experimental manipulation (Table 4), such that within studies of language learning there is 88.88% heteroge-

neity, within studies of narrative comprehension there is 86.29% heterogeneity, and within studies of spatial reasoning learning there is 76.25% heterogeneity. Compared to the total between-study variance accounted for in the omnibus meta-analysis (I^2 = 75.0%), these three domains explain *less* heterogeneity than the omnibus random effects model.

In contrast, there is *less* residual heterogeneity within samples of mathematics, science, and *other* (see Table 4), such that within studies of mathematics learning there is 24.32% heterogeneity, within studies of science learning there is 4.30% heterogeneity, and within studies of other there is 41.76% heterogeneity. A smaller I^2 indicates less residual heterogeneity in the subgroup, and, therefore, often corresponds to a non-significant Cochran's Q statistic, such that Cochran's Q indicates whether there is significant heterogeneity within the subgroup. Therefore, a non-significant Cochran's Q and a small I^2 indicate that within the strata of mathematics, science, and *other* there is some degree of homogeneity in these subgroups in which each domain accounts for a significant portion of the variance in the respective strata, above and beyond other methodological variations. As seen with conceptual-learning, descriptively, this suggests that perhaps there is more consistency in gesture's effect on learning in these subgroups, compared to the other domains investigated here.

Overall, although the domain predictor accounts for a large amount of heterogeneity in the current sample, with the stratification analysis accounting for 89.71% of the total between-study variation (I^2 = 10.29%), there is still significant residual heterogeneity unaccounted within domains (Q_{within} = 381.25, p <.0001). The stratification analysis indicates that using gesture is in general more beneficial for learning than not using gesture during instruction across domains (except *other*). However, there is not a significant difference in heterogeneity between domains

($Q_{between}$ = 3.28, p = .66). The potential difference in gesture's effect across domains will be explored further with a metaregression, when we ask if domain moderates gesture's effect.

Table 4. Cochran's Q and stratification test results for domain

Stratum	Heterogeneity			Stratified Pooled Estimates				
	Q	*p*-value	I^2	*M*	*CI* (95%)	*z*-score	*p*-value	n
Language	175.77	<.0001	88.88%	.52	.33, . 71	5.29	<.0001	30
Mathematics	8.89	0.351	24.32%	.44	.08, 79	2.37	0.018	9
Narrative Comp.	126.14	<.0001	86.29%	.54	.34, .75	5.12	<.0001	25
Spatial Reasoning	46.09	<.0001	76.25%	.41	.13, .70	2.83	0.005	15
Science	14.39	0.276	4.30%	.42	.12, .72	2.72	0.007	13
Other	9.96	0.126	41.76%	.26	-.11, .63	1.37	.171	6

Note. CI = confidence interval. *M* refers to the pooled value of Cohen's *d* (effect size).

Resolving discrepancies. *Redundancy.* The fourth stratification explored whether gestures that provide redundant or additional (non-redundant) information to speech benefit learning. Of the 98 samples included, 92 investigated the effect of redundant gestures on learning and 6 investigated the effect of non-redundant gestures on learning. A random effects model was used for each strata (Table 5). Examination of weighted mean effect sizes indicates that the impact of speech+gesture is greater than speech-alone for both subgroups, with non-redundant gestures showing a trending effect of instruction (p = .051). Results suggest that both redundant (d = .47) and non-redundant (d = .44) gestures have a medium beneficial effect on learning.

Stratifying by the information provided in gestures indicates that within both subgroups there is heterogeneity unaccounted for by the experimental manipulation. However, there seems to be less residual heterogeneity within the non-redundant strata, compared to the redundant strata (See Table 5). Specifically, within studies using redundant gestures there is 82.10% heteroge-

neity and within studies using non-redundant gestures there is 23.15% heterogeneity[8]. Compared to the total between-study variance accounted for in the omnibus meta-analysis (I^2 = 75.0%), redundant gestures explain *less* heterogeneity than the omnibus random effects model. The non-significant Cochran's *Q* and smaller I^2 associated with non-redundant gestures suggests there is some degree of homogeneity in this subgroup, above and beyond other methodological variations.

Overall, although the information provided by gestures accounts for a large amount of heterogeneity in the current sample, with the stratification analysis accounting for 94.56% of the total between-study variation (I^2 = 5.44%), there is still significant residual heterogeneity unaccounted for within redundant and non-redundant subgroups (Q_{within} = 386.73, p <.0001). The stratification analysis indicates that using gesture is in general more beneficial for learning than not using gesture during instruction for both forms of gesture. However, there is not a significant difference in heterogeneity between redundant and non-redundant gestures ($Q_{between}$ = 0.06, p = .80). The potential difference in gesture's effect across redundancy will be explored further with a metaregression.

Table 5. Cochran's Q and stratification test results for redundancy

	Heterogeneity			Stratified Pooled Estimates				
Stratum	*Q*	*p*-value	I^2	*M*	*CI* (95%)	*z*-score	*p*-value	n
Non-redundant	6.03	0.303	23.15%	.44	-.003, .88	1.95	.051	6
Redundant	380.71	<.0001	82.10%	.47	.36, .57	8.63	<.0001	92

Note. CI = confidence interval. *M* refers to the pooled value of Cohen's *d* (effect size).

[8] Given that I^2 is influenced by the number of studies included within an analysis (von Hippel, 2015), I^2 often over or underestimating the amount of systematic variance when less than seven studies are included.

Learner's age group. The fifth stratification explored whether gesture's benefit depends on the learners' age. Of the 98 samples included, 36 investigated gesture's effect on children's learning and 62 investigated gesture's effect on adults' learning. A random effects model was used for each learning type strata (Table 6). The weighted mean effect sizes for both subgroups were significantly greater than zero, suggesting that the impact of speech+gesture is greater than speech-alone instruction for both age groups. Results suggest that instruction using gesture has a medium beneficial effect on both children (d = .55) and adults' learning (d = .42).

Stratifying by age group indicates that within both age groups there is heterogeneity unaccounted for by the experimental manipulation (see Table 6), such that within studies assessing children's learning there is 76.17% heterogeneity and within studies of assessing adult's learning there is 82.38% heterogeneity. Compared to the total between-study variance accounted for in the omnibus meta-analysis (I^2 = 75.0%), both age group strata explain *less* heterogeneity than the omnibus random effects model.

Although age group accounts for a large amount of heterogeneity in the current sample, with the stratification analysis accounting for 94.89% of the total between-study variation (I^2 = 5.11%), there is still significant residual heterogeneity unaccounted within age groups (Q_{within} = 384.66, p <.0001). The stratification analysis indicates that using gesture is in general more beneficial for learning than not using gesture during instruction across age groups. However, there is not a significant difference in heterogeneity between age groups ($Q_{between}$ = 2.10, p = .15). The potential difference in gesture's effect across redundancy will be explored further with a metaregression.

Table 6. Cochran's Q and stratification test results for learner's age group

	Heterogeneity			Stratified Pooled Estimates				
Stratum	*Q*	*p*-value	I^2	*M*	*CI* (95%)	*z*-score	*p*-value	n
Child	124.27	<.0001	76.17%	.55	.38, .73	6.26	<.0001	36
Adult	260.38	<.0001	82.38%	.42	.29, .54	6.39	<.0001	62

Note. CI = confidence interval. *M* refers to the pooled value of Cohen's *d* (effect size).

Attempts to replicate previous findings. ***Gesture type.*** The sixth stratification explored whether gesture's effect depends on its form. The types of gestures included in this analyses were beat, deictic, iconic, metaphoric, and mixed gestures, indicating the use of multiple gesture types in one study. Of the 98 samples included, 3 investigated the effect of beat gestures, 17 investigated the effect of deictic gestures, 40 investigated the effect of iconic gestures, 10 investigated the effect of metaphoric gestures, and 28 investigated the effect of mixed gestures. A random effects model was used for each gesture type strata (Table 7). For samples that investigated deictic, iconic, metaphoric, and mixed gestures, the weighted mean effect size was significantly greater than zero, suggesting that the impact of speech+gesture is greater than speech-alone for these four gesture types. Specifically, results suggest that instruction using deictic (d = .45) and iconic (d = .50) gesture has a medium beneficial effect on learning, instruction using metaphoric gesture has a large (d = .76) beneficial effect on learning, and instruction using mixed gesture has a small (d = .35) beneficial effect on learning. However, for samples that investigated beat gestures, the weighted mean effect size was not significantly different from zero, suggesting that instruction using beat gestures has a nonsignificant effect on learning.

Stratifying by gesture type indicates that within types there is heterogeneity unaccounted for by the experimental manipulation (see Table 7), such that within studies using deictic ges-

tures there is 46.81% heterogeneity, within studies of using iconic gestures there is 80.40% heterogeneity, within studies using metaphoric gestures there is 85.22% heterogeneity, and within studies using mixed gestures there is 85.10% heterogeneity. Compared to the total between-study variance accounted for in the omnibus meta-analysis (I^2 = 75.0%), all gesture types explain *less* heterogeneity than the omnibus random effects model, with the exception of deictic gestures. Descriptively, this suggests that there is some degree of consistency in the effect of deictic gestures on learning, despite all other possible methodological variations. Only 3 studies were included in the current meta-analysis that investigated the effect of beat gestures on learning. Consequently, the reported I^2 values for beat gestures cannot be meaningfully interpreted (von Hippel, 2015).

Although gesture type accounts for a large amount of heterogeneity in the current sample, with the stratification analysis accounting for 95.14% of the total between-study variation (I^2 = 4.86%), there is still significant residual heterogeneity unaccounted within gesture types (Q_{within} = 370.99, p <.0001). The stratification analysis indicates that using gesture is in general more beneficial for learning than not using gesture during instruction across all gesture types (except beat gestures). However, there is not a significant difference in heterogeneity between gesture types ($Q_{between}$ = 4.94, p = .29). The potential difference in gesture's effect across gesture types will be explored further with a metaregression.

Table 7. Cochran's Q and stratification test results for gesture type

	Heterogeneity			Stratified Pooled Estimates				
Stratum	*Q*	*p*-value	I^2	*M*	*CI* (95%)	*z*-score	*p*-value	n
Beat	4.64	0.098	56.45%	.39	-.13, .91	1.45	0.146	3
Deictic	32.20	0.009	46.81%	.45	.18, .71	3.34	.001	17

Iconic	153.99	<.0001	80.40%	.50	.34, .66	6.19	<.0001	40
Meta-phoric	51.56	<.0001	85.22%	.76	.40, .1.11	4.17	<.0001	10
Mixed	85.10	<.0001	85.10%	.35	.17, .54	3.72	<.0001	28

Note. CI = confidence interval. *M* refers to the pooled value of Cohen's *d* (effect size).

Experience of gesture. The seventh stratification explored whether gesture's benefit is dependent on how the gesture is experienced (observed or produced by the learner). Of the 98 samples included, 79 investigated the effect of observed gestures on learning and 19 investigated the effect of produced gestures on learning. A random effects model was used for each strata (Table 8). The weighted mean effect sizes for both subgroups were significantly greater than zero, suggesting that the impact of speech+gesture is greater than speech-alone instruction for both experiences of gesture. Specifically, results suggest that observing gestures has a small to medium beneficial effect on learning (d = .41) and producing gestures has a large beneficial effect on learning (d = .67).

Stratifying by experience of gesture indicates that within both groups there is heterogeneity unaccounted for by the experimental manipulation (see Table 8), such that within studies assessing observed gestures there is 77.51% heterogeneity and within studies of assessing produced gestures is 85.88% heterogeneity. Compared to the total between-study variance accounted for in the omnibus meta-analysis (I^2 = 75.0%), both strata explain *less* heterogeneity than the omnibus random effects model.

Although experience of gesture accounts for a large amount of heterogeneity in the current sample, with the stratification analysis accounting for 99.25% of the total between-study variation (I^2 = 0.75%), there is still residual heterogeneity unaccounted within subgroups (Q_{within}

= 372.01, p <.0001). Importantly, there is significant heterogeneity between experiences of gesture ($Q_{between}$ = 4.64, p = .03). When taken into consideration with the weighted mean effect sizes of each strata, these findings suggest that when gestures are produced by the learner, the benefit of speech+gesture instruction, compared to speech alone, is more pronounced than when gestures are observed.

Table 8. Cochran's Q and stratification test results for experience of gesture

	Heterogeneity			Stratified Pooled Estimates				
Stratum	*Q*	*p*-value	I^2	*M*	*CI* (95%)	*z*-score	*p*-value	n
Observed	276.79	<.0001	77.51%	.41	.30, .52	7.19	<.0001	79
Produced	95.23	<.0001	85.88%	.67	.44, .89	5.81	<.0001	19

Note. CI = confidence interval. *M* refers to the pooled value of Cohen's *d* (effect size).

Across the stratified meta-analyses reported here, the pooled effect sizes indicate variability in the benefits of gesture, from small to large benefits. Furthermore, heterogeneity was considerable. As a result, metaregression analyses were used to determine whether differences in heterogeneity between subgroups were significant and whether methodological variations across studies impact the degree of heterogeneity in the sample.

Moderator Analyses

Specifically, with this next set of analyses, we asked if these 7 methodological variations significantly moderate the effect gesture has on learning. In order to compare across levels of an individual predictor, we used a series of metaregressions – one metaregression for each predictor to ask if there is a meaningful difference between levels of a predictor in gesture's effect on learning. For example, a metaregression that indicates a significant difference between *types of learning* would suggest that the impact of gesture on learning is different across the two learning

types. As with the other analyses, the presence and degree of heterogeneity is interpreted to indicate the degree to which each predictor accounts for variability in the sample. A total of 7 predictors were analyzed. Five predictors had two levels: type of learning, redundancy, experience of gesture, age, and timing of assessment. One predictor had five levels: gesture type, and one predictor had six levels: domain. These variables were dummy coded with the reference category as the subgroup of studies with the smallest effect size. The results of the metaregression analyses are shown in Table 9. All results that involved multiple comparisons (i.e., domain, gesture type) were Bonferroni adjusted for the number of comparisons.

Table 9. Metaregression test results for moderator variables

Comparison	β	SE_β	p-value
Conceptual vs fact	.01	.12	0.924
Redundant vs non-redundant	.05	.21	0.799
Produced vs observed	.27	.12	0.031*
Child vs adults	.15	.11	0.147
Delay vs immediate	.17	.16	0.272
Beat vs mixed	.00	.29	0.997
Iconic vs mixed	.16	.12	0.199
Deictic vs mixed	.07	.16	0.644
Metaphoric vs mixed	.40	.19	0.038
Language learning vs other	.31	.21	0.147
Mathematics vs other	.20	.26	0.431
Narrative comprehension vs other	.34	.22	0.116
Spatial reasoning vs other	.21	.24	0.380
Sciences vs other	.18	.24	0.459

* Denotes significant moderator at $p < .05$, or after Bonferroni adjustment.

Note. The second predictor listed in each comparison is the reference category for the respective predictor.

For 6 of the 7 predictors, category level did not significantly predict the size of gesture's effect. These were type of learning, redundancy of gestures, age of learner, timing of assessment, type of gesture, and domain. Specifically, whether samples explored the effect of gesture on fact- or conceptual-learning, whether samples used redundant or non-redundant gestures, whether samples used children or adults, or whether samples were tested immediately following training or after a delay did not significantly predict the size of the effect that gesture had on learning. Similarly, whether the type of gesture used during instruction was iconic, deictic, beat, metaphoric, or mixed did not significantly predict the size of the effect against a Bonferroni adjusted α of .013, and whether the content trained with gesture was language learning, mathematics, narrative comprehension, spatial reasoning, science, or categorized as *other* did not significantly predict the size of the effect against a Bonferroni adjusted α of .01. In contrast, samples that investigated the effect of producing gestures on learning found a larger effect of gesture than samples that investigated the effect of observing gestures on learning. Together, the metaregression and stratification analyses suggest that for experience of gesture, one subgroup explains more heterogeneity than the other. For all other predictors, there is not a significant difference in explained heterogeneity across subgroups.

As a final investigation with metaregression models, we further explored the impact of learner's age. A primary aim of the present meta-analysis was to resolve discrepancies among the previous two meta-analyses exploring the impact of gesture. Whereas the use of two age groups (children and adults) replicates the method used by Hostetter (2011), the present findings do not replicate Hostetter's findings. That is, although Hostetter found that children benefit more than adults from speech+gesture instruction compared to speech-alone, in the present sample,

age was not a significant moderator and adults and children benefit from gesture in similar ways. However, this is in line with Dargue and colleagues' (2019) findings. Therefore, we wanted to further test if we could replicate their findings by using similar age groups. Dargue split their sample across five age groups (preschool, elementary, adolescents, young adults, and older adults). Because only one study in the present sample had participants who were *older adults* (age 60-80), that study was aggregated into the adult category, resulting in 4 age groups[9]. Results cannot be meaningfully interpreted for subgroups with less than 2 studies (Valentine, Pigott, & Rothstein, 2010). By running one additional metaregression model using random-effect weights, we support our previous finding that age does not moderate gesture's effect in this sample. Specifically, whether samples were preschool children, elementary children, adolescents, or adults did not significantly predict the size of gesture's effect on learning.

Multi-Model Inference

Now that we understand the impact of individual predictors and their subgroups, we performed an exploratory analysis to investigate whether combinations of these methodological variations predict gesture's effect on learning. A multi-model inference analysis provides a comprehensive look at which factors are important for predicting differences in effect sizes (*metafor* package in R, Viechtbauer, 2010). By assessing the fit of models with all possible combinations of predictors, the multi-model inference determines the best fitting models – whether that is a model with only a pair of predictors, a set of predictors, or all predictors.

---

[9] Studies were categorized by age group according to the mean age of participants in the study, or based on the specified participant age range if mean age was not reported. Samples were classified as *preschool* if the sample was 5 years or younger. Samples were classified as *elementary* if the sample was 6-11.99 years or a description of the sample indicated elementary age children (e.g., primary school, 5^{th}-6^{th} grade). Samples were classified as *adolescents* if the sample was 12-17.99 years or a description of the sample indicated adolescents (e.g., 7^{th}-8^{th} grade) . Samples were classified as *adults* if the sample was 18+ years or a description of the sample indicated adults (e.g., undergraduates, university students).

The outcome variable was the unbiased effect size of each study and the model used random-effect weights. Six predictors were analyzed. Timepoint was not considered in this model as that would require a sample including both the immediate and delayed studies. 64 possible models[10] were fitted. Average importance of each predictor across all models was calculated (see Table 10). Results suggest that the 6 predictors vary in the degree to which they predict a difference in effect sizes, and whether studies assess the effect of observed or produced gestures on learners is the most important predictor to the model, which is in line with the previous findings in this study. Corrected Akaike information criterion (AICc) was used to compare and identify the best fitting models out of all possible models. AICc is an estimator of prediction error and thereby estimates the relative quality of statistical models for a given set of data. In other words, it estimates the quality of each model, relative to the other models. Small AICcs indicate the best balance of model fit. See Table 11 for the five best fitting models with the smallest AICcs. The *experience of gesture* predictor (produced vs observed) is present in all five models, suggesting that this factor accounts for significant variation in the sample of effect sizes. The age group factor is present in three of these models, suggesting that it also accounts for a relatively large portion of variance in this sample.

Table 10. Average importance of each predictor across all models

Predictor	Predictor Importance
Experience of gesture	0.76
Age group	0.50
Redundancy	0.33
Type of learning	0.32

[10] This is the total number of possible combinations of the 6 predictors, without all predictors having to be in the model at once (i.e., 2^6). Each model had include a minimum of 2 predictors.

Type of gesture	0.17
Domain	0.06

Table 11. Model selection: Five best fitting models

Intercept	Age of learner	Type of learning	Experience of gesture	Redundancy of gesture	*df*	AICc
X	X		X		4	176.4
X			X		3	176.9
X	X		X	X	5	177.8
X	X	X	X		5	177.9
X		X	X		4	178.3

Note. The smaller the AICc, the better the model fits the sample.

Publication Bias

The data were examined for publication bias. Publication bias in a meta-analysis is often due to the greater likelihood of significant findings being published more often than non-significant findings, and therefore, more likely to be discovered in the search process and included in a meta-analysis. Specifically, studies that do not reach $p < .05$ or .01 are less likely to be published (Dickersin, 1997; Ioannidis, 1998). To understand if this accounts for publication bias in the present sample, a contour-enhanced funnel plot was used. This plot overlays contours that represent different levels of statistical significance on a traditional funnel plot (Peters, Sutton, Jones, Abrams, & Rushton, 2008). In the absence of publication bias or small study effects, the funnel plot should appear symmetrical. Similarly, when sampling error is the sole source of variance in the distribution of effect sizes, a funnel plot tends to be symmetrical. A contour-enhanced funnel plot of the effect size against the standard error of each study was generated (see Figure 11A). Inspection of the funnel plot revealed a pattern of effects that were asymmetrically distrib-

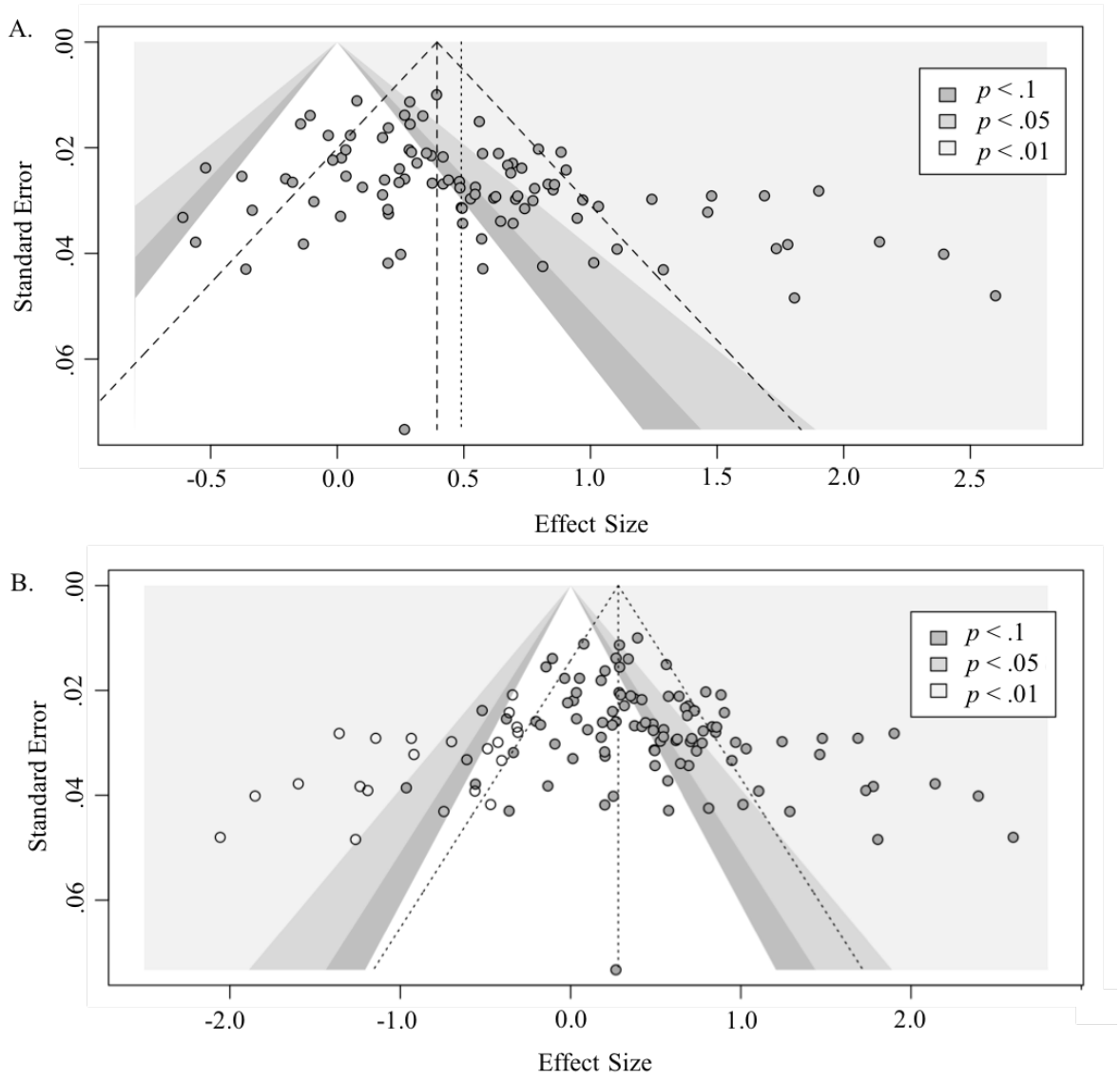
uted with studies appearing to be missing from areas of low significance, suggesting possible positive overestimation of the overall effect size and that asymmetry may be due to publication bias.

Given the visual asymmetry of the funnel plot, the trim-and-fill method (Duval & Tweedie, 2000) was used to impute and estimate the effect sizes of any identified missing studies, and subsequently, adjust the mean effect size as though there was no presence of publication bias (i.e., as though the funnel plot were symmetrical). See Figure 11B for the contour-enhanced funnel plot with imputed effect sizes. The trim-and-fill method identified 21 studies that would be needed to have a symmetrical distribution, and the corrected mean effect size was calculated as .28 (95% *CI* = .17, .39). Compared to the .49 mean effect size for the current sample of studies, this suggests that initial results were overestimated potentially due to publication bias.

Beyond the trim-and-fill method, Egger's test of asymmetry was conducted on both the omnibus model using the full sample and the separate stratified analyses using the sample subgroups to further assess whether publication bias was present in the current meta-analysis (Egger, Smith, Schneider, & Minder, 1997). In total, twenty Egger's tests were performed. Tests were significant for some analyses, which may indicate the potential for publication bias. Therefore, statistically significant findings should be interpreted in that context. Table 12 provides details of analyses with significant Egger's test results[11].

[11] A full list of Egger's test results including nonsignificant analyses is available upon request.

Figure 11. A. Contour-enhanced funnel plot. B. Contour-enhanced funnel plot after applying trim-and-fill method



There was no evidence of publication bias on Egger's tests for strata that investigated beat gestures, metaphoric gestures, deictic gestures, mixed gestures, non-redundant gestures, mathematics, sciences, and the *other* domain category (*ps* > .05). Because Egger's test may lack statistical power to detect bias when the number of studies is small (i.e., $k < 10$), tests of strata with small bins (i.e., gesture type of *beat*, domain of *mathematics*, and domain of *other*) should

be interpreted with caution. Overall, the evidence for publication bias in the current meta-analysis represents a limitation of our findings given the possibility that our effects sizes may be overestimated in some cases.

Table 12. Significant Egger's Tests Results

Stratum	intercept	95% CI	*t*	*p*-value
Iconic gestures	2.88	1.33, -4.42	3.65	<.0001
Fact-learning tasks	2.16	0.83, -3.49	3.81	.002
Conceptual-learning tasks	1.87	0.60, -3.14	2.89	.009
Redundant gestures	2.09	1.01, -3.16	3.81	<.0001
Observed gestures	1.71	0.65, -2.77	3.17	.002
Produced gestures	4.11	0.93, -7.29	2.54	.021
Narrative comprehension	2.53	0.38, -4.68	2.30	.031
Spatial reasoning	2.34	0.29, -4.42	2.24	.044
Language learning	3.37	0.86, -5.87	2.64	.013
Child learners	2.78	1.42, -4.17	3.99	<.0001
Adult learners	1.52	0.06, -2.97	2.04	.045
Overall	2.05	1.02, -3.09	3.89	<.0001

Discussion

The aim of the present meta-analysis was to build on previous work synthesizing the gesture-for-learning literature and further explore the impact of situational characteristics of the learning environment on gesture's effects. We first asked whether across studies learners benefit more from gesture+speech instruction compared to speech alone instruction and found that they do. Next, we asked if situational factors or methodological variations across the sample of studies differentially impact gesture's effect on learning. The final aim was more exploratory in nature – we considered whether some combination of factors predict gesture's effect. The results of

this synthesis have implications for research investigating the role gesture plays in learning, by asking when gestures accompanying verbal instruction are more beneficial than speech alone instruction.

Gesture's Effect on Learning

In line with the a priori hypothesis, across all studies in the current sample there is evidence that gestures significantly benefit learning. The unbiased effect size, .49, is significantly greater than zero and can be interpreted as a medium beneficial effect, suggesting that learners perform better after instruction that incorporates speech and gesture components, rather than speech alone. This result is also in line with both previous meta-analyses: Hostetter (2011) obtained a mean unbiased effect size of .61 across 63 samples and Dargue et al. (2019) obtained .61 across 83 samples. Of the 98 samples in the present analysis, 85% percent demonstrated a positive effect size. However, only 46% of the sample reported a significant difference between speech+gesture and speech-alone conditions. Although there is variability in whether or not studies detect a significant impact of gesture, it appears that across studies gestures have a significant, beneficial effect on learning. The remaining questions investigated by the current meta-analysis explored a variety of factors that could potentially moderate the beneficial effect of gesture.

Exploring the Effect of Novel Potential Moderators

First we discuss the impact of three potential moderators that have not previously been investigated in meta-analyses. These factors take a step beyond considering characteristics of the learner (e.g., age) or of the gesture (e.g., redundancy or gesture type), to consider characteristics of the learning context (type of learning, domain, and timing of assessment).

The effect of learning type. Here, we considered *fact-learning tasks* as measures of one's ability to memorize and re-produce pieces of information (e.g., Dargue & Sweller, 2018a; Kelly et al., 2009) and *conceptual-learning tasks* as measures of one's ability to learn and apply rules or strategies for solving problems (e.g., Congdon et al., 2017; Novack et al., 2014). In line with previous research, the present meta-analysis finds that incorporating gesture in instruction is beneficial for both fact-learning (e.g., Austin & Sweller, 2014; Buisine & Martin, 2007) and conceptual-learning (e.g., Ping & Goldin-Meadow, 2008; Rueckert et al., 2017) tasks. Although learning type accounts for a large degree of heterogeneity in the current sample, there is still significant residual heterogeneity unaccounted for within learning types, suggesting that other methodological variations also impact gesture's effect. Furthermore, the effect sizes for these two groups of studies did not significantly differ from one another, suggesting that there is not a significant difference in gesture's effect between learning types. However, we did find that conceptual-learning explains more heterogeneity than the omnibus random effects model. This suggests that gesture's effect on learning is more consistent within the group of conceptual-learning studies, compared to the whole sample, and importantly compared to the group of fact-learning studies. Therefore, although the impact of these two types of learning do not significantly differ, the present results hint at perhaps a stronger relation between gesture's impact and conceptual learning, such that gesture may be particularly advantageous for tasks that require conceptual learning.

Fact- and conceptual-learning may capitalize on different aspects of gesture. That is, during tasks of fact-learning, gestures often provide a concrete or imagistic piece of information that is associated with a single concept. For example, during a word learning task, a novel word for

bird may be associated with an iconic gesture representing *flapping wings* – where there is a one-to-one association between a representational gesture and a concept presented in speech. During fact-learning, gestures can help clarify speech – understanding that a *flapping-wings* gesture represents a bird can facilitate comprehension of a novel word for *bird* if the gesture and word are presented simultaneously. In contrast, during tasks of conceptual learning, gestures often give more than a single piece of representational information. Rather gestures during conceptual-learning can help give a framework for generalizing schematics or rules to novel contexts or problems. These gestures can facilitate comprehension that extends beyond surface-level associations (e.g., flapping-wings gesture means bird) to help learners understand complex underlying rules that can be applied across contexts. In the case of analogical reasoning or mathematical equivalence, which are both cases of conceptual-learning, gestures do not simply represent the individual items in an analogy or math problem, but rather the underlying structure of the problem that can be used solve subsequent problems. In sum, any difference in gesture's effect between learning types may be due to how gesture functions during instruction and how the gesture is associated with the verbal components of instruction.

If this is the case, then gesture's impact may further depend on what the gesture adds to instruction, such that the contribution of gesture may vary between these types of learning which could drive a difference in gesture's effect between learning types. For example, characteristics of the gesture, such as gesture type or gesture's redundancy with speech, may vary in predictable ways across these learning types. From a closer look at the present sample, it seems that gesture type varies across these two types of learning. Interestingly, it appears that across fact-learning studies, 56% used iconic or metaphoric gestures, 11% used deictic gestures, and 31% used mixed

gestures, whereas across conceptual-learning studies, only 39% used iconic or metaphoric gestures and another 39% used deictic gestures[12]. Descriptively, it seems that fact-learning studies use a wider variety of gesture types, which may account for the greater degree of residual heterogeneity in this subsample. Additionally, the redundancy of gesture may vary systematically across these learning types. However, because few studies in the current sample used non-redundant gestures (n = 6), it is difficult to make conclusions about whether fact- or conceptual-learning studies tend to use redundant or non-redundant gestures. There may be variability within these subgroups due to gesture type or redundancy that we are not accounting for in the present study. Future research should further investigate whether gesture's effects differ across types of learning and explore how learning type interacts with other aspects of the learning environment.

The effect of domain. In addition to considering a distinction in the gesture-for-learning literature between fact- and conceptual-learning tasks, we also considered if the effect of gesture varies by domain. And although the general consensus is that gesture is beneficial across domains, there is variability in the literature, suggesting that this is not always the case. With this analysis we wanted to explore if there is a pattern to this variability – whether gesture's effects are more prominent for some domains than others. The present analysis reflects the variability in the field and suggests that gesture's effects are not explicitly moderated by domain. Aside from the *other* category, the weighted mean effect sizes in all other subgroups were significantly different from zero, suggesting that gestures are beneficial for narrative comprehension, language learning, mathematics, spatial reasoning, and science. However, there was still significant heterogeneity within these subgroups, suggesting that other contextual factors of the learning environment also impact gesture's effect. Interestingly, for studies that investigated mathematics and

[12] This distribution does not account for beat gestures or mixed gestures in the case of conceptual-learning studies.

science, there was *less* heterogeneity than in the omnibus test, suggesting that gesture's effects are perhaps more consistent in these subgroups, despite other methodological variations. However, this interpretation must be taken cautiously as there was not a significant difference in gesture's effect between domains.

As discussed, descriptively there seem to be differences in heterogeneity between mathematics and science, and narrative comprehension, language learning, and spatial reasoning. Because interpretation of these somewhat subtle differences is limited with the present analysis, future work should further explore if there is a distinction between these two groups that is driving differences in heterogeneity. For example, perhaps these two groups utilize gestures in different ways. That is, the latter group (narrative comprehension, language learning, and spatial reasoning) may include more methodological variations that, in turn, result in more variability in gesture's effects. In contrast, the former group (mathematics and sciences) may use gestures more consistently, and, in turn, there are more consistent effects of gesture on learning. Beyond using similar gesture types in mathematics and science (i.e., primarily deictic and mixed gestures in the present sample), gesture may function similarly across these two domains. For example, during mathematics or science instruction gestures may be particularly useful for drawing connections across problems or contexts. This aligns with the present sample: all mathematics studies and roughly half of the science studies were categorized as conceptual-learning tasks. As discussed previously, conceptual-learning is about understanding problem-structures or strategies that can be used across problems, and gesture is particularly useful for facilitating extension of learning across contexts (e.g., Novack et al., 2014; Wakefield et al., 2018b). This potential relation between conceptual-learning and the domains of mathematics and science is further supported by a

lesser degree of heterogeneity in each of these subgroups. That is, if these two domains frequently require conceptual-learning then perhaps some degree of consistency in the function of gesture accounts for the lesser heterogeneity in effect sizes.

In contrast, all narrative comprehension studies and a large proportion of the language learning and spatial reasoning studies were categorized as fact-learning tasks. Similar to the fact-learning subgroup, the increased heterogeneity in these domain subgroups may be due to variability in gesture's function during instruction. Future work should explore whether a relation between learning type and domain exists, and if this potential relation impacts gesture's effect because gesture functions differently across domains.

In sum, the present findings are in line with previous work suggesting that, in general, gesture is beneficial across domains, but it seems that other methodological variations across studies also impact gesture's effect on learning. Future research should explore the effect of gesture in additional domains: those not considered here and those with limited existing research. For example, the studies binned into the *other* category used here included a variety of domains, from singing (e.g., Liao & Davidson, 2016), technical or software programs (e.g., Baylor & Kim, 2009; Buisine & Martin, 2007), to history (e.g., Beege et al., 2020). Existing research on these domains in the gesture-for-learning literature is limited. Therefore, future meta-analyses in this field should either specifically investigate the impact of gesture by domain or capitalize on a larger breadth of studies.

The effect of timing of assessment. The third novel potential moderator that we explored in this meta-analysis is the timing of assessment: immediate vs delayed measures of learning. Previous work has found that the benefits of gesture are increasingly evident over time, such

that positive effects of gesture are more prominent at a follow-up measure (e.g., Congdon et al., 2017; Cook et al., 2013), compared to initial measures of performance. With the present sample, gesture benefited learning at both timepoints, yet there was not a significant difference in the impact of gesture between timepoints. Specifically, even though timepoint accounts for a large degree of heterogeneity in the current sample, there is still significant residual heterogeneity unaccounted for within timepoints. This indicates that other methodological variations impact the variability of effect sizes in these subgroups.

Across the 11 studies that used delayed learning measures there was a high degree of variability across the length of delay, varying from 24 hours to 6 months after training. Future work should investigate if gesture's effects depend on the length of delay, and whether there is a limit to gesture's effects or if they continue to grow over time. The variability in length of delay combined with a small sample size warrants cautious interpretation of these findings. Furthermore, it may be that at different timepoints gesture's effects are impacted by additional methodological variations across studies. For example, perhaps the benefit of redundant vs non-redundant gestures vary across timepoints. The benefit of non-redundant gestures may be more apparent at delayed measures because they provide additional information to speech in an another modality to support consolidation of information. Future research should ask this question with a larger sample and explore if other methodological variations across studies differentially impact gesture's effect over time.

Further, these results should be interpreted with caution, given that only 11 studies assessed learning at delay and, importantly, all of these studies shared samples with their immedi-

ate measure counterparts. Therefore, independent samples was not achieved for this particular analysis – which means that 11 samples are overrepresented.

Resolving Discrepancies

Beyond exploring a set of novel potential moderators, we aimed to resolve discrepancies found across the two previous meta-analyses in the gesture-for-learning literature. Specifically, whereas Hostetter (2011) found that a learner's age and redundancy of gesture with speech differentially impact the effect of gesture on communication, Dargue and colleagues (2019) did not find an effect of either factor. Importantly, both meta-analyses used very different samples and even used different categorizations of learner's age. Here, we attempted to resolve these differences by using a sample that incorporated studies from both meta-analyses and additional years of gesture-for-learning research.

The effect of learner's age. Hostetter (2011) suggests that the benefits of gesture for children are significantly greater than for adults. However, Dargue and colleagues (2019) did not find a difference in gesture's benefit across age. Importantly, the two meta-analyses categorized age in different ways: Hostetter used two age groups, categorizing studies with sample mean ages under 12 years as children and all other studies as adults, and Dargue categorized studies into four age groups for analysis (preschool, primary school, adolescents, and adults). Additionally, Dargue had access to nearly an additional decade of literature – Either one or both of the these factors could account for the conflicting age effects. Therefore, to resolve this discrepancy, the present meta-analyses first used the age categorization scheme used by Hostetter and asked if learner's age differentially impacts the effect of gesture on learning. Results suggest that alt-

hough there is not a difference in gesture's effect between children and adults, both age groups do benefit from the inclusion of gesture during instruction.

To further explore this discrepancy, we used the categorization scheme used by Dargue and colleagues to see if the lack of heterogeneity between age groups would remain when using more specific age categories. In line with Dargue's findings, results suggest that all age groups, with the exception of adolescents, benefit more from speech+gesture compared to speech alone instruction, but the benefit of gesture did not significantly differ across age groups. This suggests that there is no significant difference between effect sizes for preschool, primary school, adolescent, and adult studies in the current meta-analysis. Previously it has been suggested that the impact of gesture follows a U-shaped curve, such that children ages 7-8 years old and adults both benefit from gesture to a greater extent than children ages 9-10 years old (Church et al., 2000). The present results do not follow this pattern. The lack of beneficial effect of gestures for adolescents was surprising. However, given that only three studies had adolescent learners in this sample, it may be that this particular analysis was underpowered. Descriptively, of the three studies that used adolescent samples, two reported significant effects of gesture (Dahl & Ludvigsen, 2014; Pouw et al., 2016). Interestingly, Dahl and Ludvigsen (2014) only found positive effects of gesture for narrative comprehension when the sample was foreign language learners, rather than native English speakers. The limited adolescent research makes it difficult to conclude whether or not a U-shaped pattern exists, and whether additional learner characteristics, such as native language, impact gesture's effect. Further research is necessary to better understand whether adolescents benefit from gestures in the same way as other age groups.

Importantly, the present findings are in line with the results of Study 2, which found that the benefits of gesture are fairly consistent across a wide age range (with exception of 5-year-olds). Together, the findings of the present meta-analysis, Study 2, and Dargue's findings suggest that, in general, gesture is beneficial for learners. With the present study, we replicate Dargue's findings and suggest that when incorporating additional years of literature the impact of gesture on learning is, on average, beneficial for learners of all ages.

The effect of gesture's redundancy. When asking whether the information provided by gestures has an impact on communication, Hostetter's (2011) meta-analysis found that when gestures provide additional information to speech (non-redundant gestures) they support more effective comprehension than when gestures provide the same information as speech (redundant gestures). However, Dargue and colleagues (2019) asked a similar question with their meta-analysis and found that both redundant and non-redundant gestures benefit comprehension in similar ways. The present study replicates Dargue's findings, such that both forms of gesture have a medium beneficial effect on learning. This suggests that whether or not gestures provide identical information to accompanying speech they are more beneficial for learning than when gestures are not used during instruction. The results of the present analysis and Dargue et al. are in line with previous work; Both redundant and non-redundant gestures have beneficial effects on learning compared to speech alone instruction (e.g., Aussems & Kita, 2019; Church et al., 2004; Church et al., 2007; Singer & Goldin-Meadow, 2005). These findings are supported by previous work that has pitted the two forms of gesture against each other. For example, Austin and colleagues (2018) found that route recall was supported by gestures that convey additional information to speech (e.g., narrator points to the left and says, 'and then you turn') *and* by gestures

that provide identical information to speech about route directions (e.g., the narrator points left and says, 'and then you turn left').

When considering these results in relation to Hostetter's findings, the impact of gesture may not be as straightforward as to say that non-redundant gestures support learning above and beyond redundant gestures in all cases. It may be that while this is true in some circumstances, this is not always the outcome. The impact of gesture, redundant or not, may depend on other characteristics of the learning environment. Specifically, the advantage of non-redundant gestures may depend on the task, such that the additional information provided by these gestures may be more or less useful for different types of content. For example, non-redundant gestures may provide an added advantage for disambiguating poor-quality spoken messages during narrative recall tasks (e.g., Church et al., 2007). Perhaps during this type task gesture can provide added context for comprehension, but for spatial reasoning tasks when the verbal information provides enough context (e.g., Austin et al., 2018), non-redundant gestures may not provide an added benefit compared to redundant gestures.

Similarly, the impact of redundant gestures varies across the field: Although previous work has found that redundant speech and gesture instruction is more beneficial for learning than speech alone instruction (e.g., Aussems & Kita, 2019; Austin & Sweller, 2014), that is not always the case (Study 1 and 2, Holle, Obleser, Rueschemeyer, & Gunter, 2010). For example, Holle and colleagues (2010) found that whereas providing iconic non-redundant gestures benefits speech comprehension more so than providing no gestures, this was only the case when the speech was difficult to comprehend. Specifically, gestures were particularly beneficial when accompanied by multi-speaker 'babble' sounds (i.e., increased signal-to-noise ratio of speech). Fu-

ture research should explore if the impact of gesture redundancy varies across domains or by task characteristics, such as complexity of the spoken message. Unfortunately, the limited number of studies that explored non-redundant gestures in the present sample prohibits these comparisons.

The root of the discrepancies explored here may be traced back to the samples used. Not only did the present analysis and Dargue's meta-analysis have access to an additional decade of data compared to Hostetter's meta-analysis, but discrepancies across these studies could be due to the scope of the investigations. Specifically, Hostetter investigated the effect of gesture on communication, Dargue investigated gesture's effect on comprehension, and the present analysis investigated gesture's effect on learning. These are semantic differences that impact each analyse's inclusion criteria. That is, in the present analysis, we included studies that assess comprehension, but specifically comprehension of instructional material, and required studies to collect explicit learning measures. Hostetter's investigation took a broad approach at considering the various contexts in which gesture could have benefits for communication, including behavioral measures of comprehension, memory, and learning. Dargue and colleagues limited their sample to studies with behavioral measures of comprehension. In the present analysis, we focused our search to studies with behavioral learning measures or comparisons between pre-gesture to post-gesture performance. Of the present sample, 8 studies were included in both of the previous meta-analyses. Twenty-one additional studies in the present analysis were also in Dargue's sample and 2 additional studies were also in Hostetter's sample. Even though the present analysis used a larger sample than the others, it focused on studies of learning, which encompassed studies investigating the effects of gesture on communication and comprehension, but only if all inclusionary criteria were met.

Replicating Previous Meta-Analytic Findings

Finally, we aimed to replicate the findings of Dargue and colleagues (2019) to assess whether they can be generalized to a different sample of studies. This final set of potential moderators explored here were targeted because they are theoretically important questions and key methodological variations across the field. These methodological variations speak to the how gesture is experienced by the learner – that is, the form of the gesture (type of gesture) and whether the gesture is generated by themselves or an instructor (experience of gesture). Whereas many of the methodological variations explored in this analysis speak to characteristics of the learner (e.g., age) or the task (e.g., timing, type of learning, domain), these two factors speak to the experience of the learner during an instructional session, and are therefore important to consider when asking how and when gesture impacts learning.

The effect of gesture type. The results obtained suggest that all gesture types explored here, with the exception of beat gestures, benefit learning. Specifically, compared to speech only instruction, when studies use of iconic, metaphoric, deictic, and mixed gestures in addition to speech learners perform better. This is in line with previous research showing that iconic (e.g., Beattie & Shovelton, 1999), metaphoric (e.g., Yuan et al., 2019), and deictic gestures (e.g., Du & Zhang, 2019) support learning. Specifically, iconic and metaphoric gestures are said to benefit learner's comprehension because the content of the accompanying speech binds with the semantic information provided by the gesture (Straube et al., 2009). That is, the more semantically related speech and gesture are the more effective they are for comprehension and learning. Additionally, gestures benefit learning because of its ability to capture and direct attention and facilitate the link between words and visual referents (Richland et al., 2007; Wakefield et al., 2018b).

Providing both iconic and deictic gestures during instruction has been shown to help children synchronize their visual attention with spoken instruction (math equivalence: Wakefield et al., 2018b; analogical reasoning: Study 1 & 2), and that this synchronization predicts success at post-test (Wakefield et al., 2018b). The findings of the current meta-analysis suggest that both semantically related gestures and those that serve to direct attention both benefit learning, and that the impact of gesture does not differ significantly across gesture types.

It should be noted that similar to Dargue's findings, even though beat gestures were not more beneficial than no gestures, effect sizes for iconic, metaphoric, deictic, and mixed gestures did not significantly differ from beat gestures. Beat gestures may have differential effects compared to other gesture types due to its function during communication. That is, iconic and metaphoric gestures have representational properties that help to tie gestures semantically to speech, and deictic gestures are particularly beneficial for directing attention and grounding learning in the environment. In contrast, beat gestures do not contain semantically important content or help to form connections between visual referents and verbal instruction. Thus, although beat gestures certainly have their place in discourse, their impact on learning may be different from other gesture types. However, because there was limited beat gesture research in present sample of studies, the analysis of beat gestures may have been underpowered, and strong conclusions cannot be made. Within the present sample, the benefit of beat gestures varies, with two of the three studies showing beneficial effects of gesture and the other showing non-significant effects. Therefore, it remains unclear across studies whether or not observing beat gestures benefits learning**.**

Importantly, these results are in line with the findings of Dargue and colleagues (2019). Aside from type of gesture used across studies, there are many other methodological variations

that may be impact the effect of gesture – some of which were explored in the present analysis. Specifically, the impact of gesture type may depend on task characteristics. For example, perhaps certain gesture types are more or less beneficial for different types of content or types of learning. Because of the limited sample of studies using particular gesture types in the present analysis (e.g., beat gestures), future research should ask if gesture type interacts with other task characteristics to support learning.

The effect of how gesture is experienced. In line with previous research, the results of the current meta-analysis suggest that both observing (e.g., Aussems & Kita, 2019; Beege et al., 2020) and producing gestures (e.g., Stieff et al., 2016; Sweller, Shinooka-Phelan, & Austin, 2020) are beneficial for learning above and beyond speech alone instruction. However, when a learner produces the gestures themselves there is an added benefit compared to when a learner observes others' gestures during instruction. Plenty of past research also suggests that producing gestures has an added benefit (Goldin-Meadow et al., 2012; Stieff et al., 2016). For example, Stieff and colleagues (2016) found that when children were asked to reproduce gestures during a spatial reasoning task, they outperformed their counterparts who were asked to only watch gestures. This result was replicable across a large sample of studies: both the present meta-analysis and Dargue found that studies that investigated the production of gesture had significantly larger effect sizes than studies that investigated observation of gesture. This finding is supported by theories of embodied cognition that suggest cognitive representations are grounded in sensory-motor processes, and that grounding knowledge in one's own bodily experience (i.e., gestures) can aid learning (e.g., Barsalou, 2008; Paas & Sweller, 2012; Wilson, 2002). These theories propose that active learning through physical movements of one's own body is more effective than passive

learning through observing others' movements because active learning capitalizes on integrated sensory-motor processes. Pouw and colleagues (2014) suggest that gestures serve as external physical tools of our cognitive system that support and potentially replace cognitive processes. In support of these theoretical assumptions, Cherdieu, Palombi, Gerber, Troccaz, & Rochet-Capellan (2017) found that when learners imitate body movements in a lecture on forearm anatomy they outperform those who only observed video instruction. Interestingly, this effect was more apparent over time – suggesting that integration of motor actions and knowledge require consolidation and reactivation of learned content.

The finding that produced gestures are more effective than observed gestures is further supported by previous work that suggests when learners produce gestures themselves during learning experiences, gestures help reduce the cognitive load placed on the learner's WM. In turn, this helps mitigate limitations of WM by off-loading aspects of cognitive processing to the physical environment, or in this case, to the hands (e.g., Alibali & Nathan, 2012; Goldin-Meadow et al., 2001). In sum, the present meta-analysis is able to provide further support for the finding that although both producing and observing gestures are beneficial for learning compared to no gestures, producing gestures provides an added benefit to the learner.

Furthermore, the differential effects of producing and observing gestures may be due to additional methodological variations across studies. That is, other characteristics of the learning environment may coincide with whether gestures are produced or observed. For example, perhaps different gesture types (iconic, deictic, etc.) tend to be used with one experience of gesture more so than the other. An interaction between gesture type and experience of gesture would address whether there is a relation between these factors that impacts learning. However, given that

only 19 studies investigated produced gestures, the current sample is too underpowered for this investigation and would not provide meaningful results. And although numerous studies have investigated the effect of *observing* different types of gesture on learning (e.g., Austin et al., 2018; Aussem & Kita, 2019; Dargue & Sweller, 2018a; Macoun & Sweller, 2016; Morett & Chang, 2015), few studies have directly compared the impact of *producing* different types of gestures. Future research is needed to understand if there are differential effects of gesture types across producing or observing gestures and if there are gaps in the literature investigating an array of gesture types using both experiences of gesture. Additional interactions with other methodological variations, such as learner's age or domain, should also be explored with a larger sample of studies to allow for a better understanding of the differences between producing and observing gestures during learning.

Multi-Model Inference

For the final aim of the present meta-analysis, we considered whether some combination of factors predict the effect gesture has on learning using a multi-model inference analysis. This model allowed us to ask not only if one of the factors stands out from the rest in terms of predicting effect sizes, but also what combinations of factors best predict gesture's effect on learning. Across the 6 predictors used in this model, we find that whether studies assessed the effect of observed or produced gestures is the most important predictor, which supports our metaregression findings. Further, results indicate that across the top five best fitting models, age group and experience of gesture (produced vs observed gestures) are present in a majority of the models. Although this analytic approach cannot speak to interactions between variables, it does tell us that these two methodological variations are driving factors predicting gesture's effects. It seems

that when accounting for other predictors, a meaningful difference between children and adult learners becomes evident, whereas when only asking if age differentially impacts gesture's effect there is no difference between groups. This finding may speak to a potential interaction, perhaps between age group and experience of gesture, and potentially other predictors. Descriptively, it appears that although children benefit equally from both produced (mean unbiased effect size = .64) and observed (mean unbiased effect size = .59) gestures, adults benefit more from produced gestures (mean unbiased effect size = .94) compared to observed gestures (mean unbiased effect size = .36). However, a degree of caution is warranted in interpreting interactions from the multi-model inference as these were not explicitly tested, and in most of the best fitting models other predictors aside from these two were also included.

Aside from these predictors, type of learning (fact vs conceptual) and redundancy of gesture (redundant vs non-redundant) were also present in the top five models. These findings are in line with the metaregression findings, such that when looking at each moderator separately, there were hints of differences between subgroups. In sum, the possible conclusions that can be drawn from the multi-model inference are seemingly endless. Future research should explore these top fitting models more extensively and the possible interactions among the set of moderators tested here as well as additional methodological variations.

Limitations

One possible limitation of the present meta-analysis is that only published, peer reviewed studies were included in the final sample. Studies that have undergone the peer review process are more likely to be considered higher quality, whereas unpublished studies are of lower quality. Quality refers to the likelihood that unpublished studies have smaller effect sizes or do not find

condition differences (i.e., lack of beneficial effects of gesture in this case). Limiting the sample to peer-reviewed studies helps to eliminate possible bias associated with unpublished studies. However, despite this inclusion criteria, there was still evidence of possible publication bias, and potentially an overestimation of the overall effect size.

The present study made strides towards understanding how methodological variations in the learning environment impact gesture's effects and extended the findings of two previous syntheses of the gesture-for-learning literature. However, in the same way as the effects of publication bias are unknown, plenty of variance is left unexplained by the set of moderators tested here. The present analysis did not attempt to look at *all* potential moderators – as that is not possible, but we were able to answer some outstanding questions about when gesture's effects are beneficial for learners. Further, some degree of variability was likely due to the use of *mixed* and *other* categories. For example, the *mixed* gesture type category was comprised of studies that used a combination of gesture types, both gesture types explored in the present analysis as well as others. The *other* domain category included a range of domains that did not fit with the other domain categories, but likely all interact with gesture's effects in different ways. Similarly, even within categories such as *delayed* measures of learning, there was a wide range of delay lengths that certainly could account for variability in the sample.

Future work should not only explore additional moderators, but also possible relations among these, and other, methodological variations. As previously suggested, it seems that interactions between moderators may drive gesture's effect in unique and interesting ways – that not only have implications for gesture research, but also for educators who implement gesture in the classroom. Heterogeneity could very well be explained by other moderators not explored here –

we only began to scrape the surface of understanding gesture's role during learning. An additional limitation of the current analysis is the sample size. Although the sample size is larger than previous meta-analyses, with a larger sample, testing of interactions and additional moderators would be more plausible. With the current sample, some subgroups included very few studies, which made testing of interactions not possible.

Conclusion

The aim of the present meta-analysis was three-fold. First, we investigated a set of factors that had not been explored through a meta-analysis previously and provided evidence that gesture's benefits are consistent despite timing of assessment, type of learning, and domain. Second, we aimed to resolve discrepancies between previous meta-analyses examining gesture's effect on learning. The findings of Dargue and colleagues (2019) regarding the impact of learner's age and redundancy of gesture were replicated by the present analysis – suggesting that compared to Hostetter (2011), the present study and Dargue's study more accurately represent the existing body of research examining gesture's effect in learning environments. And finally, beyond resolving discrepancies, this work validated the impact of particular gesture characteristics on learning – finding that all gesture types are more beneficial than no gesture and producing gestures provides an added benefit compared to observing gestures. However, it is important to keep in mind that a large degree of heterogeneity remained in the sample suggesting that no one factor alone accounted for a significant portion of the variability in effect sizes.

The present study provides evidence that, in general, instruction that incorporates speech and gesture is beneficial for learning above and beyond instruction with only speech, replicating the omnibus finding of both previous meta-analyses. When considering a handful of methodo-

logical factors that vary across the gesture-for-learning literature, as outlined above, we find that gestures are beneficial under a variety of circumstances and that in general these methodological variations do not explicitly moderator gesture's benefit for learners. Only one factor tested here – whether the gestures are produced or observed by the learner – significantly predicted whether gestures provided an added benefit above speech alone instruction. Specifically, while both producing and observing gestures supports learning, producing gestures gives an extra boost in performance following instruction.

In sum, the story of gesture's effects is complicated; Additional methodological variations should be explored in future research, as well as the potential impact of interactions between factors in determining gesture's effect on learning. Demonstrating that gesture is important across contextual variations of the learning environment is just as important for educators as understanding the conditions in which gesture is most advantageous.

CHAPTER FIVE

GENERAL DISCUSSION

Summary of Main Findings

The purpose of this body of research was to advance our understanding of gesture's effect on learning. Across Studies 1 and 2, we investigated the effects of gesture in a domain that has not been considered in previous gesture-for-learning research: analogical reasoning. Specifically, we asked whether gesture can support children's analogical reasoning ability, and explored *why* and *how* gesture might support this type of reasoning. Study 3 took a holistic approach to understanding gesture's effect on learning by synthesizing across the gesture literature to investigate *under what* conditions speech+gesture instruction is most beneficial for learners. Following a review of the key findings across these studies, we will explore how Studies 1 and 2 add nuance to the gesture-for-learning literature and how Study 3 informs our findings from Studies 1 and 2, as well as our broader understanding of gesture's role and function during instruction.

As discussed, gesture has been an effective instructional tool for a variety of topics, including mathematics (Cook et al., 2013; Singer & Goldin-Meadow, 2005), word learning (e.g., Rowe et al., 2013; Wakefield et al., 2018a), and Piagetian conservation (Church et al., 2004; Ping & Goldin-Meadow, 2008), to name a few. Previous research suggests that gesture is an effective instructional tool because it directs visual attention toward key components of instructional content (e.g., Rohlfing et al., 2012; Wakefield et al., 2018b) and facilitates links between words and what they map onto (Richland et al., 2007; Wakefield et al., 2018b). As previously

discussed, both of these functions help gesture facilitate language comprehension (Hostetter, 2011; Kendon, 1994), particularly when speech is ambiguous (Thompson & Massaro, 1986). For these reasons, we predicted that gesture would be an effective tool for supporting children's analogical reasoning ability; It should orient attention towards relational comparisons, and, thereby, facilitate comprehension of ambiguous speech that accompanies analogy problems. Thus, the features of gesture that make it an effective teaching tool in other domains suggest that is has the potential to facilitate analogical reasoning ability.

In line with this prediction, gesture promoted effective visual attention strategies in Studies 1 and 2 by orienting children's attention away from the featural match and synchronizing attention with spoken instruction. However, Studies 1 and 2 suggest that incorporating gesture during instruction does not always benefit children's analogical reasoning to a greater extent than speech only instruction beyond training. The findings of Studies 1 and 2 tell an inconsistent story about the effects of gesture: Study 1 found that 4- and 5-year-olds benefited from both speech+gesture and speech-alone instruction, and Study 2 found that 5-year-old children demonstrated additional benefits from gesture, yet children older and younger than 5 years learned from both forms of instruction. Together, these results suggest that gesture instruction can support children's understanding of analogical reasoning, but children may only benefit from gesture above and beyond spoken instruction when they reach a certain level of cognitive development. These results add to our understanding of children's analogical reasoning and how we can support development of this foundational ability.

A secondary aim of Study 1 was to better understand how visual attention might relate to children's IC, a factor that has been posited to contribute to the protracted development of analogical reasoning (e.g., Begolli et al., 2018; Doumas et al., 2018; Morrison et al., 2011). Whereas

the role of IC in adult analogical reasoning has long been established (e.g., Morrison et al., 2004; Viskontas et al., 2004), our results expand upon previously established associations between IC and children's analogical reasoning by identifying a relation between IC and visual attention measures. This was the first time IC was not only associated with children's scene analogy performance, but also children's visual attention during problem solving. Eye-tracking data from Study 2 further suggests that gesture can disambiguate speech during instruction; Gesture oriented visual attention away from the featural match during the point of instruction that was most ambiguous, which was predictive of successful problem solving post-instruction.

Together, Studies 1 and 2 extend our understanding of gesture's function during instruction to the domain of analogical reasoning. These findings speak to one of the reasons *why* gesture supports learning (i.e., to disambiguate spoken instruction), but also suggest that individual differences across learners determine the impact gesture can have. Across Studies 1 and 2 it is evident that gesture's impact is dependent on a number of factors, including the cognitive capacities a child brings to an instructional session. And when compared to previous research in other domains, it seems gesture's impact may also depend on the type of material being taught. Although this is not the first time non-significant effects of gesture have been found (e.g., Beattie & Shovelton, 1999; Driskell & Radtke, 2003), these findings raise important questions about the general claim that gesture boosts learning beyond speech-alone instruction.

Because a wide array of methodological variations exists across the gesture-for-learning literature, Study 3 investigated whether the differential effects of gesture found in previous work are due to these methodological variations. These variations include contextual differences in the learning environment: characteristics of the task, the learner, and the gesture itself. Using a meta-

analytic approach we synthesized across the literature to ask *when* and *under what* conditions gesture is most beneficial for learning. Overall, in line with two previous meta-analyses in the field, our results suggest that gesture accompanying verbal instruction is beneficial for learning above and beyond verbal only instruction. A range of potential moderators were tested to see if they significantly predict gesture's effect. While there are several nuances to the these findings, the only factor that moderated the impact of gesture on learning was whether the gesture was produced or observed by the learner. Specifically, larger learning effects were found when gesture is produced by the learner rather than when gesture is only observed.

Study 3 not only helps us gain a better understanding of *under what* conditions gesture is most advantageous for learning but also provides additional insight into the findings of Studies 1 and 2. That is, Study 3 can help us understand how gesture functions during analogical reasoning instruction. These implications are discussed further in the remainder of this chapter. Broadly, the threads that connect these three studies can be divided into three aspects of the learning environment: learners' characteristics, task characteristics, and characteristics of the gesture.

The Impact of Learners' Characteristics on the Utility of Gesture Instruction

When considering the impact of learner's characteristics on the present studies, we will: 1) discuss the relation between learner's cognitive profile and gesture's effects and consider whether the impact of cognitive profile is due to gesture's ability to disambiguate speech, and 2) explore how the non-significant effect of age in Study 3 helps us understand the findings of Study 2 and consider potential sources of the variability in the effect of age.

Children's task-relevant cognitive profile determines gesture's effects, potentially because of gesture's ability to disambiguate speech. The surprising finding of Study 1 (i.e., the

lack of posttest difference in performance between conditions) suggests that perhaps a factor other than type of instruction was impacting children's performance and even comprehension of instruction. With Study 2 we investigated if the findings of Study 1 may be attributable to individual differences between children – that is, the degree of task-relevant cognitive abilities. We hypothesized that perhaps the children of Study 1 were not able to overcome their maturational limitations to be successful when instructional support was no longer present at posttest. In Study 2, age was treated as a proxy for children's cognitive profile, which, in this case, represented the cognitive abilities (e.g., IC and WM) that support improvements in analogical reasoning over development. Whereas the youngest children in the age range did not show a benefit after viewing either form of instruction, the oldest children benefited from both forms of instruction. It was posited that the youngest children did not have the rudimentary capabilities in place to make use of gesture instruction to support improved analogical reasoning, whereas the older children were able to learn from instruction with and without gesture.

Interestingly, 5-year-old children were at a unique timepoint in development of analogical reasoning, in that they were able to capitalize on the added support gesture provided during instruction. In the context of Study 1 and previously established associations between children's IC, analogical reasoning ability, and visual attention, Study 2 suggests that 5-year-old children have some cognitive abilities relevant to analogical reasoning in place to observe and learn from gesture instruction, and then apply that understanding to subsequent problem solving. Interestingly, this conflicts with the findings of Study 1, which did not find a posttest difference between conditions for 4-5-year-olds. If the variability in gesture's effects is, at least in part, due to cognitive profile or maturational limitations, then perhaps the differential effects at this age may be

due to the particular samples of children that were used in Studies 1 and 2. Future work should explore if the findings of Study 2 are replicable across analogy tasks and if length of intervention impacts gesture's effects, given that testing sessions of Study 2 were significantly shorter than that of Study 1.

Study 2 not only illustrates the protracted development of analogical reasoning ability, but also indicates that gesture's effects depend on children's task-relevant cognitive profile, and perhaps task variations, which will be explored later in this chapter. This is not the first time gesture's effects on learning have been found to depend on children's cognitive profile; Previous work has similarly found that prior knowledge within a domain serves as a foundation that gesture can build from, such that children need some degree of prior knowledge to learn from gesture (e.g., mathematics: Congdon et al., 2018; palindromes: Wakefield & James, 2015). For example, Wakefield and James (2015) find that incorporating gesture during instruction on the concept of palindromes only benefited children with a certain level of phonological awareness.

The differential effects of gesture due to cognitive profile may be driven by gesture's ability to disambiguate spoken instruction. That is, children with some degree of task-relevant cognitive profiles, but not entirely mature cognitive profiles, may only need gesture to clarify the particularly ambiguous parts of instruction. This assumption is supported by our eye-tracking findings; What we termed as *check-ins with the featural match* predicted posttest performance. When the item in the source relation that was featurally similar to the distractor was discussed verbally – which was the particularly ambiguous part of instruction – gesture oriented attention towards the source relation item and away from the distractor. Without gesture, 5-year-old children struggled to understand *which* featurally similar item was important to solving the problem.

But, with gesture, eye-tracking results suggest that these children better understood which items were relevant and irrelevant to solving the analogy. In sum, gesture's ability to disambiguate speech may be particularly useful for 5-year-old children who have the foundational cognitive abilities in place to benefit from gesture instruction.

A similar role of gesture has been found previously in other domains. Wakefield and colleagues (2018b) found that following along with spoken instruction predicted subsequent performance on tasks of mathematical equivalence. While children may have some degree of arithmetic concepts in place that set the stage for mathematical equivalence, they need to learn the meaning of the equals sign. In this case, spoken instruction highlighted two sides of an equation and indicated that a pair of numbers on one side can be added to find the missing addend on the other side. However, without the visual support of gesture, spoken instruction was ambiguous when referencing *which* numbers need to be added to find the missing addend and that the equals sign indicates numbers on *each* side of the equation should be added *separately* to solve the problem. Therefore, the findings of Study 2 and Wakefield and colleagues indicate that gesture has a unique ability to disambiguate spoken instruction when it is needed most for children's comprehension.

In sum, the differential effects of gesture due to cognitive profile may be rooted in gesture's ability to disambiguate speech, such that older learners who have more mature task-relevant cognitive profiles may not need gesture to disambiguate spoken instruction to the same extent as younger learners. For example, in the case of scene analogies, the ambiguity inherent in instruction will reduce over development, as cognitive profile matures. Taken together, the re-

sults of Studies 2 and 3 suggest that this function of gesture is an important mechanism by which gesture shapes learning.

Variability in the effect of learner's age may reflect differences in cognitive profile. Although the findings from Study 2 suggest that cognitive profile determines gesture's effects, these findings could similarly reflect an impact of children's age. If it is age that impacts gesture's effects, and not cognitive profile, then we would expect to see an effect of age in Study 3. However, we do not have an effect of age, which suggests that there is not one particular age, or in this case, period of development, in which gesture is most beneficial – rather it is likely that gesture benefits learners at a certain point in their knowledge state development relative to the task.

In addition to the non-significant effect of age in Study 3, significant heterogeneity remained in the sample when accounting for age. Given that task-relevant foundational knowledge varies across domains, and even across tasks within a particular domain, this variability is expected. That is, the heterogeneity in the sample, in particular within the subgroup of child studies, may be due to the vast variability in domains and tasks used in the sample of studies. For example, even within the domain of mathematics, although 8-10-year-olds (e.g. Singer & Goldin-Meadow, 2005; Wakefield et al., 2018b) may be 'ready' to learn concepts of mathematical equivalence, they may not be ready to learn more complex concepts of ANOVA or fractions. Importantly, in the present meta-analysis these mathematical concepts would have been binned together, and thus, result in further heterogeneity. Therefore, the cognitive profile relevant for learning each concept is going to be different, and presumably the age at which a learner is in this 'sweet spot' of cognitive maturity would vary by content.

Further, as seen in previous work, not all children arrive at these developmental points at the same time or when approaching a particular age. For example, Congdon and colleagues (2018) find that children who use more rudimentary pretest strategies associated with the mathematical concept of unit measurement did not benefit from producing gestures. Instead, their counterparts with more advanced conceptual knowledge of units did benefit from gesture during instruction. Importantly, all children were of the same age, but their relevant knowledge base varied – lending further support to the argument that age does not solely account for differences in gesture's effects, but rather children's task-relevant cognitive profile is a driving factor. Individual differences across children of similar ages may also account for the differential findings of 5-year-olds across Studies 1 and 2. Future work should further explore the role of individual differences, such as learners' cognitive profile, in determining gesture's effects and whether the impact of cognitive profile varies systematically by domain or task.

Additionally, researchers should consider other learner characteristics that may impact the effect of gesture for learning beyond cognitive profile or age. For example, the impact of children's language comprehension on learning from gesture varies across the literature. Hostetter (2011) asked if studies that include listeners from a special population with lower verbal proficiency vs more typical listeners, and if native vs nonnative speakers, differ in their ability to learn from gesture. Hostetter found that neither of these learner characteristics moderated the effect of gesture. However, Sueyoshi and Hardison (2005) found that the benefit of gesture for comprehension varied between native speakers (i.e., advanced learners of English as a second language) and nonnative speakers (i.e., low-intermediate learners). Specifically, whereas advanced learners showed better comprehension without gestures, the low-intermediate learners

showed better comprehension with gestures. Future work show explore the effect of learners' language comprehension and its potential impact on gesture's instructional effects – as this may help us further understand the role of cognitive profile in determining gesture's effects, as well as whether the root of gesture's ability to facilitate comprehension is its capacity to disambiguate speech.

Variability in Gesture's Effects due to Task Characteristics

As previously discussed there is reason to believe that variability in gesture's effects is not only impacted by characteristics of the learner, but also due to variations across tasks. When reviewing the impact of task characteristics on the present studies, we will review: 1) how the impact of domain may be rooted in an effect of cognitive profile, 2) how an effect of timepoint may become apparent when considering different types of learning, and 3) how fact- and conceptual-learning may capitalize on different aspects of gesture.

Learning is not simply moderated by domain; The story of gesture's effect is more nuanced. Study 3 finds that gesture supports learning in all domains examined and that gesture's effects are not moderated by domain, suggesting that learning effects are not dependent explicitly on domain. Rather, as discussed, gesture's effects depend on learner's task-relevant cognitive profile, which not only varies across domains, but also within domains, such that different tasks or problem types can require different forms of prior knowledge in order to benefit from gesture. Similarly, the ability of gesture to facilitate comprehension across domains may also depend on task characteristics, including the relation between speech and gesture during instruction. For example, the degree to which gesture helps to disambiguate speech may determine whether ges-

ture is beneficial – that is, the benefits of speech+gesture may be increasingly evident as instruction becomes more complex.

Previous work supports this assumption, finding that comprehension of complex messages is supported by reinforcing gestures, but not by speech alone, whereas simple messages were supported by both types of instruction (McNeil et al., 2000). McNeil and colleagues demonstrated this effect of gesture by comparing preschool-age children, for which the message was more complex, and kindergarten-age children, for which the message was considered simple. This suggests, that when messages are complex, and inherently more ambiguous, gesture is particularly helpful. This aligns with the findings of Study 2, indicating a 'sweet spot' in development, where a foundational knowledge state is necessary to learn from gesture. Therefore, rather than gesture's effect being explicitly different across domains, characteristics of the task used during instruction (e.g., the relation between speech and gesture to disambiguate verbal instruction) and even learner's cognitive profile contribute to variability found across the gesture-for-learning literature.

Although we did not find differential effects of timepoint, there is reason to believe they may exist under different conditions. The potential impact of timing of assessment was explored in Study 3 by asking whether, across the gesture-for-learning literature, gesture's effect varies in systematic ways dependent on the delay between instruction and assessment. The present meta-analysis found that speech+gesture instruction supports learning beyond speech alone instruction at both timepoints, but learning was not significantly different between timepoints.

Although Study 3 did not find a differential effect of timepoint, previous work has found that gesture's effects become increasingly evident over time even when condition differences do

not exist at immediate posttest measures. Previous research suggests that memory consolidation of learning contributes to these effects (Cherdieu et al. 2017; Cook et al., 2013; McGregor et al., 2009). Consolidation involves stabilization and enhancement of information in memory over time (Walker, 2005). Prior work suggests that performance improvements over time may be due to positive effects of sleep-dependent consolidation, such that consolidation occurs within the first 24 hours after learning (Diekelmann & Born, 2010; Margoliash & Fenn, 2008; McGaugh, 2000). During sleep, the neuronal networks involved in learning are reactivated (Ji & Wilson, 2007; Wilson & McNaughton, 1994) and memory traces are reorganized (e.g., Landmann et al., 2014). In the case of Studies 1 and 2, this may also extend to analogical reasoning, such that the performance of children in speech+gesture vs speech-alone may become increasingly distinct if a follow-up measure was used 24 hours or later after initial training. Future research should explore this possibility in the domain of analogical reasoning.

The lack of significant difference between timepoints in Study 3 may be due to other methodological variations across these subgroups that cause heterogeneity in the sample. For example, combined effects across methodological variations, such as timepoint and type of learning, may interact to impact gesture's effects. That is, the impact of gesture may change over time in different ways for fact- vs conceptual-learning. As discussed, previous research suggests that gesture may be particularly useful for consolidation of knowledge over time. The type of knowledge building required for conceptual learning (e.g., flexible rule or strategy learning) may benefit more from this kind of consolidation more so than fact-learning. This is not to say that fact-learning cannot also benefit from consolidation, but the process of reactivation and reorganization of memory traces may mimic the knowledge building process for conceptual-learning.

The lack of differential effects of gesture in Study 3 across timepoints may be due to variability caused by aggregating across different types of learning.

Additionally, as discussed, there is a wide range of delay lengths in the present sample that inevitably cause heterogeneity in effect sizes. Further, the analytic methods used in Study 3, in which the immediate and delay measures share samples is not ideal. Future work should address this possible limitation by maintaining independent samples, as well as exploring whether the effects of timing are more or less apparent for different types of learning or if length of delay moderates gesture's effects. Therefore, although Study 3 did not find differential effects of timepoint, there is reason to believe some systematicity in the effect of timepoint exists, but other characteristics of the learning environment, such as type of learning or length of delay, need to be considered.

Fact- and conceptual-learning may capitalize on different aspects of gesture. In addition to domain and timing of assessment, Study 3 categorized studies as fact-learning tasks (e.g., Dargue & Sweller, 2018a; Kelly et al., 2009) or conceptual-learning tasks (e.g., Congdon et al., 2017; Novack et al., 2014). The results of Study 3 allude to more consistency in gesture's effects for the conceptual-learning subgroup, compared to the fact-learning subgroup. Although this finding must be taken cautiously because type of learning did not significantly moderate gesture's effect in the larger sample, this result does suggest some degree of uniformity in gesture's effects across the conceptual-learning studies – something about this subgroup of studies results in less heterogeneity in effect sizes, compared to the fact-learning subgroup. This difference in heterogeneity may speak to how gesture functions during instruction for these two types of learning.

That is, beyond the potential relation with timing of assessment, these two types of learning may capitalize on different aspects of gesture. As discussed previously, for fact-learning, speech may be imagistic or situate concepts in space, such that the associated gestures may be one-dimensional and provide surface-level features of the message. In contrast, for conceptual-learning, gesture can provide the framework for generalizable schematics or rules for problem solving. This latter type of learning is different from fact-learning because comprehension must extend beyond a surface-level understanding to learn underlying principles or rules. Gesture may be particularly useful for disambiguation during conceptual-learning tasks, which may include more ambiguous speech than fact-learning tasks. As seen in Studies 1 and 2 which are tasks of conceptual-learning, and in previous work, gesture is particularly useful for directing attention and clarifying spoken instruction. It may be that these functions of gesture are more necessary for successful comprehension and generalization of learning in conceptual tasks, rather than in fact-learning tasks that do not require such flexible understanding or comprehension beyond the imagistic value of gesture. That is, the benefit of speech+gesture compared to speech alone instruction may be more apparent for conceptual-learning compared to fact-learning where the impact of gesture may be more variable. Therefore, any differences in variability of effect sizes between types of learning may be due to gesture's ability to overcome the ambiguity and complexity that is inherent in conceptual-learning tasks.

Beyond the situational characteristics mentioned here, a wide variety of contextual factors can vary in a learning environment, including, but not limited to, whether the instructor's gestures are scripted or spontaneous, the positioning of the instructor, their gestures, and the visual content, or even the frequency of instructor's gesture use. Furthermore, while it could not be

explored in the present meta-analysis, future work should also investigate if the length of delay between learning and assessment or the length of training impacts the benefits of gesture. Understanding how the factors discussed here and others impact the effect of gesture for learning has important implications for not only the gesture-for-learning literature, but also for educators in the classroom.

The Impact of Variations in Redundancy, Form, and Experience of Gesture on Learning

Beyond characteristics of the learner and the task, characteristics of the gesture itself are an important piece of the puzzle in understanding how gesture functions as a learning tool. Knowing how characteristics of gesture impact learning is important for understanding when to implement it in the classroom for the greatest benefits and how to support learning in a variety of contexts, with a variety of learners. As seen in Studies 1 and 2, gesture is particularly beneficial when it clarifies spoken instruction. This is supported by the *Integrated-Systems Hypothesis*, which suggests that because gesture and speech form an integrated system of meaning during language production (Kendon, 1986; McNeill, 1992), gesture can serve to enhance language comprehension (Kelly et al., 2010). As discussed previously, the relation between gesture and the content of the instructional session (e.g., the accompanying speech, the domain, type of learning) may be one factor that determines when gesture is most advantageous. Beyond the previously discussed relations between gesture, learner characteristics, and task characteristics, there is a wide range of methodological variations across the literature specific to gesture's relation with spoken instruction, including the redundancy of speech and gesture content, the type of gesture used, and how gesture is experienced by a learner. Study 3 asked if these three factors moderate the effect gesture has on learning.

When considering the impact of gesture's characteristics on the present studies, we will: 1) examine the role of redundancy between speech and gesture during learning and to what degree redundancy is needed for analogical reasoning instruction, 2) discuss the benefits of different gesture types and how Study 3 helps us further understand the results of Studies 1 and 2, and finally, 3) review how the added benefit of producing gestures is supported by theoretical and empirical prior work.

Whether gesture is redundant or not with speech may determine gesture's ability to disambiguate. Hostetter's (2011) meta-analysis found that gestures that provide additional information to that provided in speech support more effective comprehension than gestures that provide identical information in speech and gesture. However, neither Dargue and colleagues (2019) or the present meta-analysis found that redundancy of gesture moderated gesture's effects on learning. This finding is supported by previous work that has pitted the two types of gesture against one another: Using a route navigation task, Austin and colleagues (2018) found that when verbal information provides sufficient information, non-redundant gestures do not provide an added benefit compared to redundant gestures. However, any differential advantage between the two forms of gesture may be more nuanced than can be accounted for in Study 3. For example, the advantage of non-redundant gestures may vary by task, such that the degree to which gestures provide additional information to speech may determine gesture's benefits. There is reason to believe that providing non-redundant gestures could be the key to clarifying ambiguous speech or orienting attention to visual content that learners may not attend to without the support of gesture. The findings of Study 2 support this claim: The portion of instruction that is considered *non-redundant*, and in turn, ambiguous, is when the two featurally similar items are being

discussed (e.g., the tiger in the source relation and the featural distractor tiger). Gesture's ability to orient attention away from the featural match when the instruction was ambiguous predicted post-instruction performance, suggesting that the utility of non-redundant gesture to clarify ambiguous speech is particularly useful for comprehension.

In the case of analogical reasoning, the majority of gestures used in Studies 1 and 2 were redundant with the accompanying speech. Future work should explore the benefit of redundant vs non-redundant gestures during analogy instruction. Given that gesture's ability to orient attention when it is most ambiguous predicts post-training performance, perhaps all other gestures incorporated in Studies 1 and 2 are not necessary. And, although redundant gestures should seemingly use fewer cognitive resources because they provide the same information as speech, it is also possible that redundant gestures require additional resources to process both modalities simultaneously. In other words, there may be a cost to redundant gestures that do not provide additional meaningful information to that provided in speech. This may implicate a double edge sword of gesture in Study 1, such that even though providing gestures helped to synchronize visual attention with spoken instruction, it also may have narrowed children's attentional focus and increased their cognitive workload. In turn, children were not able to examine the featural distractor and understand its irrelevance to the analogy structure. Similarly, providing information in two different modalities through non-redundant gestures may facilitate disambiguation, but may also require increased processing capacity. Therefore, it would follow that learners with more advanced cognitive profiles and cognitive capacities are more prepared to learn from gesture instruction, particularly if the gesture is non-redundant with speech. Future research should

further explore the potential implications of redundant and non-redundant gestures on learner's processing capacity.

Beyond redundancy of gesture with speech, Study 3 considered two additional factors that are important methodological variations across the gesture-for-learning literature. Whereas Hostetter (2011) did not investigate gesture type or experience of gesture (i.e., whether the gesture was produced or observed by the learner), Dargue and colleagues (2019) asked whether these factors moderate gesture's effects. In Study 3, we replicated both of Dargue's findings regarding these factors.

Non-significant effects in Studies 1 and 2 are not likely a result of the gesture type used. Study 3 was not the first time researchers have asked if comprehension and learning vary by gesture type used during instruction (e.g., Aussems & Kita, 2019; Macoun & Sweller, 2016; Morett & Chang, 2015). However, these investigations typically limit their comparisons to two or three types of gestures at a time. The meta-analytic approach used in Study 3 allowed the comparison of 5 different gesture types to ask if type moderated gesture's effect on learning. The results obtained suggest that all gesture types explored in Study 3, with the exception of beat gestures, benefit learning. Previous work suggests that when gesture is semantically related to accompanying speech it is more beneficial, which would account for the previously identified benefits of iconic and metaphoric gestures (e.g., Beattie & Shovelton, 1999; Yuan et al., 2019). However, it has also been suggested that gesture benefits comprehension because of its ability to direct attention and facilitate the link between words and visual referents (Richland et al., 2007; Wakefield et al., 2018b). For example, providing both iconic and deictic gestures during instruction helps children synchronize their visual attention with spoken instruction (math equivalence:

Wakefield et al., 2018a; analogical reasoning: Studies 1 & 2) and this synchronization predicts success at posttest (Wakefield et al., 2018b). Because the findings of Study 3 suggest that all gesture types support learning, and because the gestures used in Studies 1 and 2 effectively orient learner's attention, we can assume that the lack of posttest effects of Studies 1 and 2 are not due to the type of gesture used. Rather, the non-significant effects of gesture are likely due to a number of other task and learner variations across the sample, as previously discussed.

The benefits of producing gestures align with theories of embodied cognition and cognitive load. Study 3 found that producing gestures has an added benefit compared to observing gestures. This aligns with theories of *embodied cognition* which propose that active learning through physical movements of one's own body is more effective than passive learning through observing others' movements because active learning capitalizes on integrated sensory-motor processes (e.g., Barsalou, 2008; Paas & Sweller, 2012; Wilson, 2002). In the context of Studies 1 and 2, in which learners *observe* instructor's gestures, gesture can effectively direct attention when observed, but this does necessarily predict learning. Study 3 could help explain why we did not see learning effects in Studies 1 and 2; Although observing gestures encouraged effective visual attention during training, perhaps asking children to produce those gestures would support extension of learning beyond training to posttest. Producing gestures may provide the added benefit of embodying learned content where knowledge is instantiated in not just speech but the learner's own hands. Further, embodiment of learning may facilitate not only directing visual attention, but also synchronizing motor movements with instruction for more successful learning. Future work should explore if embodiment of learning via producing gestures supports analogical reasoning beyond training.

Previous work suggests that gesture helps mitigate limitations of WM when learners *produce* gestures themselves. Specifically, *Cognitive Load Theory* posits that gesture use increases individual's WM capacity, and, in turn, frees up cognitive resources for processing (Cook et al., 2012; Ping & Goldin-Meadow, 2010; Wagner, Nusbaum, & Goldin-Meadow, 2004). In other words, when produced, gestures help reduce the cognitive load placed on the learner (Cherdieu et al. 2017; Cook et al., 2013; McGregor et al., 2009) by off-loading aspects of cognitive processing to the physical environment (e.g., Alibali & Nathan, 2012; Goldin-Meadow et al., 2001). In the case of analogical reasoning, gesture may serve a similar purpose. While not explicitly tested in Studies 1 or 2, even observing gestures may maximize children's limited processing capacity by orienting attention to relevant items in scene analogies. By conveying meaning differently than speech and facilitating connections between spoken instruction and a physical analogy problem, gesture may help children reason analogically about multiple contexts simultaneously. Future work should investigate the relation between children's WM and gesture during analogy instruction – that is, explore whether children's processing capacity is benefited by gesture and whether degree of WM impacts gesture's effects.

In addition, *Cognitive Load Theory* may help us understand the impact of gesture type on learning. For example, gesture types that carry more semantic content and are more semantically related to spoken instruction (i.e., iconic or metaphoric gestures) may be able to capitalize on this function of gesture to mitigate WM limitations. This may be particularly true when gestures are produced rather than observed, such that producing gestures may help to off-load some semantic processing on to the learner's hands. In contrast, although deictic or beat gestures may add additional meaningful information to instruction, using these gesture types may not help to mitigate

WM limitations since they are less tightly linked to the spoken instruction. These assumptions need to be explored in future research to better understand the benefits of producing gesture, and whether those benefits differ depending on the gesture type used.

Future Directions and Next Steps

With this line of work we explored the impact of gesture in a domain not considered before in the gesture-for-learning literature and took a holistic approach to ask under what conditions is gesture most beneficial for learners. Although this work helped us further understand the impact of gesture on learning, there are many additional questions to address, three of which are to investigate: 1) the relation between children's cognitive profile and development of analogical reasoning (e.g., IC *and* WM) and how this relation is impacted by different types of instruction, 2) the effect of producing vs observing gestures during analogical reasoning and whether any benefit of these two types of instruction vary over time, and lastly, 3) the impact of manipulating *when* gestures are used during analogical reasoning instruction, including whether the redundancy of gesture with speech can be optimized for children's learning.

Taking the next step to understand the impact of children's cognitive profile on learning from analogical reasoning instruction. As discussed, gesture may increase learner's WM capacity during tasks of analogical reasoning, as seen in other domains. Study 1 demonstrated that a common measure of IC (Erikson's flanker task) is directly associated with children's performance and visual attention prior to instruction. After instruction, results indicate that gesture facilitated a decoupling of 4-5-year-old children's IC and their analogy performance, suggesting gesture helped minimize the consequences of limited IC. Future work should similarly test the relation between young children's WM and their scene analogy performance by taking

an independent measure of WM. Although it is likely children's WM may correlate with their analogy performance prior to instruction, as was seen with IC, gesture may also help minimize the consequences of limited WM post-instruction.

Additionally, similar to Study 2, future work should investigate the relation between children's task-relevant cognitive profile (IC and WM) and their analogical reasoning ability over development. Rather than treating age as a proxy for cognitive profile, this work should use independent measures of IC and WM to understand the degree to which each of these abilities impact analogy performance over development and if a combination of these abilities best accounts for children's developing analogical reasoning. This work would add empirical evidence to the theoretical and computational findings of previous work positing the role of IC *and* WM in children's analogical reasoning (e.g., Begolli et al., 2018; Doumas et al., 2018; Morrison et al., 2011). Importantly, this work would address if speech+gesture vs speech-alone instruction differentially impact the relation between these abilities and analogical reasoning. The findings of Studies 1 and 2 suggest that a combination of IC and WM would best predict children's analogical reasoning ability prior to instruction, and that using gesture during instruction might alter this relation between cognitive profile and analogy performance – but *how* exactly this relation would be impacted may vary over development.

Manipulating how gesture is experienced and timing of assessment to optimize speech+gesture instruction for analogical reasoning. In order to better understand how contextual variations of the learning environment impact children's analogical reasoning, future research should investigate whether observing gestures is the most advantageous form of instruction. In light of the findings of Study 3 and Dargue and colleagues' meta-analysis it may be that

producing gestures helps children embody the subject matter and effectively process the instruction through the two modalities (speech and gesture). However, asking children to only observe instruction in this domain may not be enough. As discussed, analogical reasoning is notoriously difficult for young children (e.g., Richland et al., 2006; Simms et al., 2018). Therefore, even though observing gestures supports effective visual attention patterns during instruction (Studies 1 and 2), this may not be enough to instantiate the underlying structure of scene analogies in a way that can be extended to problem solving when gesture is no longer providing that visual support. By comparing conditions in which children are asked to observe or produce gestures during instruction we could explore this question.

Furthermore, any differential effects of observing or producing gestures for analogical reasoning may become apparent over time. Previous work finds that learners who imitate movements outperform those who only observed instructional videos, but the effect was only found at a delay assessment not at an immediate posttest measure (Cherdieu et al., 2017), suggesting that integration of motor actions and knowledge require consolidation. Similarly, any added benefit of producing gestures during analogical reasoning instruction may become more apparent after a delay and integration of motor actions and speech can occur.

The benefit of gesture for analogical reasoning may depend on how redundant gesture is with speech. In contrast to Hostetter's meta-analysis, Study 3 and Dargue and colleagues found that both redundant and non-redundant gestures benefit learning and that there is not a significant difference between these forms of gesture. As discussed, a number of methodological variations could account for the remaining heterogeneity in these subgroups, including the degree to which gestures provide additional information to speech. For example, in tasks of analog-

ical reasoning the impact of gesture may vary depending on *how redundant* the gesture is with accompanying speech. Perhaps using gestures in a targeted way, such as *only* when the potentially ambiguous speech is occurring (e.g., when the featurally similar item is discussed), could emphasize to the learner the importance of the information conveyed at that time in speech and gesture. Additional gestures (e.g., those that indicate the chasing relations within scenes) may be unnecessary or even overwhelm children's attentional capacities, as previously discussed. This could be easily tested by manipulating whether the gesture is used throughout instruction (as in Studies 1 and 2) or only during ambiguous speech.

Alternatively, adding gestures that highlight the difference between the two featurally similar items may also benefit learning. For example, future work could manipulate *how* the gesture is used during the ambiguous speech ('*..the tiger is in the same part of the pattern as the lion because they are both chasing…',* Figure 1A). One condition could limit gestures to indicating the relationally important item (e.g., the tiger in the source relation.), as is done in Studies 1 and 2. Another condition could use gestures to indicate that one tiger *is* the correct choice (e.g., using a deictic gesture) and that the featural distractor tiger *is not* the correct choice during the ambiguous speech (e.g., using an iconic X-shaped gesture over the distractor tiger). Thus, the latter condition would seemingly be even *less* ambiguous at this point of the instruction than the former condition. Further, this comparison would contrast non-redundant vs redundant gestures, respectively, and address whether the additional gesture indicating the incorrect choice (e.g., an X-shaped gesture) benefits learning. This proposed future direction would address the utility of redundant gestures for clarifying ambiguous speech during analogical reasoning instruction. Furthermore, this work would add to our understanding of how gesture functions during tasks of an-

alogical reasoning – that is, whether young children need the additional support of entirely redundant gestures or if less is better to limit cognitive workload.

Final Conclusions

Taken together, the findings of these three studies add to the existing body of work investigating gesture's effects for learning. We not only provide further evidence that gesture is, in general, beneficial for learning, but make it clear that there are nuances to these effects. This body of work has important implications for designing teaching methods to support analogical reasoning, but also for using gesture as a teaching tool more broadly. Understanding how gesture functions during analogical reasoning instruction can help educators support the development of this ability that is at the root of a wide range of cognitive skills, including innovation, creativity, and inductive reasoning, which are all critical for future success (for review see Gentner & Smith, 2013). Further, these studies suggest that gesture does not always impact learning in predictable ways. Rather the effects of gesture depend on a variety of contextual and situational factors that comprise the instructional environment, including factors related to the learner themselves, to the content being learned, and to the gesture itself. Future work should continue to investigate the nuances of gesture's effects to help us understand how gesture functions in the classroom.

APPENDIX A

SUMMARY OF STUDIES INCLUDED IN CURRENT META-ANALYSIS

Author(s)	Year	Cou-ntry	Sample	Design	N	Age Group	Exp. of Gesture	Redun-dancy	Gesture Type	Domain	Type of Learning	Delay Learning Measure	Benefit	*d*	Included in other meta-analyses
Aldugom, Fenn, & Wagner Cook	2020	US	study 1 + 2	between	128	adult	observed	Y	deictic	math	concept.	N	N	0.052	
Allen	1995	US		between	112	adult	produced	Y	mixed	language learning	fact	Y	Y	1.901	
Andrä, Mathias, Schwager, Macedonia, & von Kriegstein	2020	DE	study 1	within	54	child	produced	Y	iconic	language learning	fact	Y	N/A	0.266	
Andrä, Mathias, Schwager, Macedonia, & von Kriegstein	2020	DE	study 2	within	43	child	produced	Y	iconic	language learning	fact	Y	N/A	0.288	
Aussems & Kita	2019	UK		between	72	child	observed	Y	iconic	narrative comp	fact	N	Y	0.618	
Aussems, Mumford, & Kita	2021	UK	study 2	between	48	child	observed	Y	iconic	language learning	concept.	N	Y	0.704	
Austin & Sweller	2014	AU	adults	between	94	adult	observed	Y	mixed	spatial reasoning	fact	N	N	0.017	Dargue et al.
Austin & Sweller	2017	AU		between	172	child	observed	Y	mixed	spatial reasoning	fact	N	N	0.202	Dargue et al.
Austin & Sweller	2014	AU	children	between	91	child	observed	Y	mixed	spatial reasoning	fact	N	Y	0.692	Dargue et al.
Austin, Sweller, & Van Bergen	2018	AU	study 2 (complete script)	between	125	adult	observed	Y	mixed	spatial reasoning	fact	N	N	-0.377	Dargue et al.
Austin, Sweller, & Van Bergen	2018	AU	study 2 (incomplet e script)	between	125	adult	observed	N	mixed	spatial reasoning	fact	N	N	0.034	Dargue et al.
Austin, Sweller, & Van Bergen	2018	AU	study 1	between	86	adult	observed	Y	mixed	spatial reasoning	fact	N	N	0.375	Dargue et al.
Baills, Suárez-Gonzáles, Gonzáles-Fuente, & Prieto	2019	ESP	study 1	between	49	adult	observed	Y	metaphor ic	language learning	fact	N	Y	0.626	

Bailles, Suárez-Gonzáles, Gonzáles-Fuente, & Prieto	2019	ESP	study 2	between	56	adult	produced	Y	metaphoric	language learning	fact	N	Y	0.779	
Baylor & Kim	2009	US	with facial expression	between	236	adult	observed	Y	deictic	other	fact	N	N/A	-0.204	
Baylor & Kim	2009	US	without facial expression	between	236	adult	observed	Y	deictic	other	fact	N	N/A	0.266	
Beattie & Shovelton	1999	UK		between	60	adult	observed	Y	iconic	narrative comp	fact	N	Y	1.478	Hostetter Dargue et al.
Beege et al.	2020	DE	study 2	between	121	adult	observed	Y	deictic	other	fact	N	Y	0.673	
Beege et al.	2020	DE	study 1	between	108	adult	observed	Y	deictic	sciences	fact	N	Y	0.728	
Buisine & Martin	2007	FR		within	108	adult	observed	Y	mixed	other	fact	N	N/A	0.392	
Carlotto & Jaques	2016	BR		between	72	adult	observed	Y	deictic	other	concept.	N	N	0.013	
Carlson, Jacobs, Perry, & Church	2014	US		between	56	adult	observed	Y	iconic	sciences	concept.	N	Y	0.852	Dargue et al.
Church, Ayman-Nolle, & Mahootian	2004	US	bilinguals	between	25	child	observed	Y	iconic	spatial reasoning	concept.	N	Y	0.811	Hostetter Dargue et al.
Church, Ayman-Nolle, & Mahootian	2004	US	mono-linguals	between	26	child	observed	Y	iconic	spatial reasoning	concept.	N	Y	1.287	Hostetter Dargue et al.
Church, Garber, & Rogalski	2007	US		within	30	adult	observed	N	iconic	narrative comp	fact	Y	Y	0.570	Hostetter Dargue et al.
Colliot & Jamet	2018	FR		between	42	adult	observed	Y	deictic	sciences	fact	N	Y	0.740	
Congdon et al.	2017	US		between	72	child	observed	N	mixed	math	concept.	Y	N	0.179	Dargue et al.
Craig, Twyford, Irigoyen, & Zipp	2015	US		between	77	adult	observed	Y	deictic	sciences	fact	N	Y	0.712	
Cutica & Bucciarelli	2008	IT	study 2	between	30	adult	observed	Y	mixed	narrative comp	fact	N	N	-0.965	
Cutica & Bucciarelli	2008	IT	study 3 (free recall task)	between	30	adult	observed	Y	mixed	narrative comp	fact	N	Y	1.105	
Cutica & Bucciarelli	2008	IT	study 1	between	38	adult	observed	Y	mixed	narrative comp	fact	N	Y	1.779	

Dahl & Ludvigsen	2014	NO	native language speakers	between	28	adult	observed	Y	mixed	narrative comp	fact	N	N	-0.134	Dargue et al.
Dahl & Ludvigsen	2014	Norway	foreign language speakers	between	46	adult	observed	Y	mixed	narrative comp	fact	N	Y	1.031	Dargue et al.
Dargue & Sweller	2018	AU		between	62	child	observed	Y	iconic	narrative comp	fact	N	N	0.490	Dargue et al.
Davis & Vincent	2019	KOR		between	183	adult	observed	Y	mixed	sciences	fact	N	N	0.371	
Du & Zhang	2019	CHN		between	90	child	produced	Y	deictic	math	concept.	N	Y	0.832	
Eng, Hanna, Leong, & Wang	2013	CA		between	12	adult	observed	Y	metaphoric	language learning	fact	N	N	0.265	
Feyereisen	2006	BE	study 1a	within	52	adult	observed	Y	mixed	narrative comp	fact	N	N/A	-0.107	Hostetter Dargue et al.
Feyereisen	2006	BE	study 2	within	54	adult	observed	Y	iconic	narrative comp	fact	N	N/A	0.338	Hostetter
Feyereisen	2009	BE	older adults	within	20	adult	produced	Y	iconic	narrative comp	fact	N	N/A	1.242	
Feyereisen	2009	BE	younger adults	within	24	adult	produced	Y	iconic	narrative comp	fact	N	N/A	2.395	
Flack & Horst	2017	UK	study 1 + 2	between	48	child	observed	Y	deictic	language learning	fact	N	Y	1.805	
García-Gámez & Macizo	2019	ESP	study 1	within	25	adult	produced	Y	iconic	language learning	fact	N	N/A	0.284	
García-Gámez & Macizo	2019	ESP	study 2	within	32	adult	produced	Y	iconic	language learning	fact	N	N/A	0.794	
Guarino & Wakefield	2020	US		between	323	child	observed	Y	mixed	spatial reasoning	concept.	N	N	0.077	
Guarino, Wakefield, Morrison, & Richland	2021	US		between	57	child	observed	Y	mixed	spatial reasoning	fact	N	N	-0.177	
Halvorson, Bushinski, & Hilverman	2019	US		within	81	adult	observed	Y	iconic	narrative comp	fact	N	N/A	0.029	
Hirata & Kelly	2010	US		between	60	adult	observed	Y	metaphoric	language learning	fact	N	N	-0.559	
Holle et al.	2012	DE	study 1	within	24	adult	observed	Y	beat	narrative comp	fact	N	N	0.034	

Huang, Kim, & Christianson	2019	US		within	23	adult	observed	Y	iconic	language learning	fact	N	Y	0.418	
Ianì & Bucciarelli	2018	IT	study 1a	within	32	adult	observed	Y	iconic	narrative comp	fact	N	N	-0.035	
Ianì & Bucciarelli	2018	IT	study 1b	within	32	adult	observed	Y	iconic	narrative comp	fact	N	Y	0.884	
Ianì et al.	2018	IT	study 2	within	16	adult	observed	Y	iconic	narrative comp	fact	N	Y	0.482	
Kartalkanat & Göksun	2020	TR	children	between	71	child	observed	Y	iconic	narrative comp	fact	N	Y	0.526	
Kartalkanat & Göksun	2020	TR	adults	between	55	adult	observed	Y	iconic	narrative comp	fact	N	Y	0.694	
Kelly & Goldsmith	2004	US		between	39	adult	observed	Y	mixed	sciences	fact	N	N	0.202	Hostetter
Kelly & Lee	2012	US		within	42	adult	observed	Y	iconic	language learning	fact	N	N	-0.145	
Kelly, McDevitt, & Esch	2009	US	study 2	within	24	adult	observed	Y	iconic	language learning	fact	N	N/A	0.352	
Kelly, McDevitt, & Esch	2009	US	study 1	within	27	adult	observed	Y	iconic	language learning	fact	Y	Y	0.637	Dargue et al.
Koumoutsakis, Church, Alibali, Singer, & Ayman-Nolley	2016	US	live instruction	between	33	child	observed	Y	deictic	math	concept.	N	N	0.200	Dargue et al.
Koumoutsakis, Church, Alibali, Singer, & Ayman-Nolley	2016	US	video instruction	between	40	child	observed	Y	deictic	math	concept.	N	Y	0.645	Dargue et al.
Krönke, Mueller, Friederici, & Obrig	2013	DE		within	11	adult	produced	Y	iconic	language learning	fact	Y	N	-0.092	
Liao & Davidson	2016	TW		between	53	child	observed	Y	metaphoric	other	fact	N	Y	0.495	
Llanes-Coromina, Vilà-Giménez, Kushch, Borràs-Comes, & Prieto	2018	ESP	study 2	between	55	child	observed	Y	beat	narrative comp	fact	N	Y	0.546	

Llanes-Coromina, Vilà-Giménez, Kushch, Borràs-Comes, & Prieto	2018	ESP	study 1	within	51	child	observed	Y	beat	narrative comp	fact	N	Y	0.561	
Lüke & Ritterfeld	2014	DE	study 1	within	20	child	observed	Y	iconic	language learning	fact	N	N	-0.520	
Macedonia, Repetto, Ischebeck, & Mueller	2019	DE		within	31	adult	observed	Y	iconic	language learning	fact	Y	N	0.179	
Macoun & Sweller	2016	AU		between	101	child	observed	Y	iconic	narrative comp	fact	N	Y	0.969	Dargue et al.
Mavilidi, Okely, Chandler, Cliff, & Paas	2015	AU		between	111	child	produced	Y	iconic	language learning	fact	Y	N	0.491	
Mayer & DaPra	2012	US	study 2	between	106	adult	observed	Y	mixed	sciences	fact	N	N	0.099	
Mayer & DaPra	2012	US	study 1	between	88	adult	observed	Y	mixed	sciences	fact	N	N	0.187	
Mayer & DaPra	2012	US	study 3	between	115	adult	observed	Y	mixed	sciences	fact	N	N	0.244	
McKern, Dargue, Sweller, Sekine, & Austin	2021	AU		between	120	adult	observed	Y	mixed	narrative comp	fact	N	N/A	-0.336	
Morett & Chang	2015	US		between	57	adult	produced	Y	iconic	language learning	fact	N	N	-0.610	
Mumford & Kita	2014	UK		between	101	child	observed	Y	iconic	language learning	concept.	N	Y	0.685	
Novack, Goldin-Meadow, & Woodward	2015	US	study 1 (3-year-olds)	between	32	child	observed	Y	iconic	language learning	fact	N	Y	2.600	
Ouwehand, van Gog, & Paas	2015	NL		between	34	adult	observed	Y	deictic	sciences	concept.	N	N	-0.361	Dargue et al.
Pi et al.	2019	CHN		between	120	adult	observed	Y	deictic	sciences	fact	N	N	0.439	Dargue et al.
Ping & Goldin-Meadow	2008	US	study 2	between	45	child	observed	Y	iconic	spatial reasoning	concept.	N	Y	0.575	Hostetter Dargue et al.
Ping & Goldin-Meadow	2008	US	study 1	between	52	child	observed	Y	iconic	spatial reasoning	concept.	N	Y	1.013	Hostetter Dargue et al.
Porter	2016	UK		within	24	child	produced	Y	mixed	narrative comp	fact	N	Y	0.904	
Pouw, van Gog, Zwaan, & Paas	2016	NL		between	70	adult	observed	Y	iconic	sciences	concept.	N	N	0.246	

Repetto, Pedroli, & Macedonia	2017	IT		within	20	adult	produced	Y	metaphoric	language learning	fact	N	N/A	0.315	Dargue et al.
Rueckert, Church, Avila, & Trejo	2017	US		between	94	adult	observed	Y	deictic	math	concept.	N	Y	0.574	
Singer & Goldin-Meadow	2005	US		between	160	child	observed	N	mixed	math	concept.	N	Y	0.485	
Stieff, Lira, & Scopelitis	2016	US	study 1	between	70	adult	produced	Y	iconic	sciences	concept.	N	Y	0.773	Dargue et al.
Sweller, Shinooka-Phelan, & Austin	2020	AU		between	60	adult	produced	Y	iconic	language learning	fact	Y	Y	0.947	
Theakston, Coates, & Holler	2014	UK	3-year-olds	between	41	child	observed	Y	iconic	language learning	fact	N	N	0.200	
Theakston, Coates, & Holler	2014	UK	4-year-olds	between	45	child	observed	Y	metaphoric	language learning	fact	N	Y	2.141	
Valenzeno, Alibali, & Klatzky	2003	US		between	25	child	observed	Y	deictic	spatial reasoning	concept.	N	N	0.251	Hostetter Dargue et al.
van Wermeskerken, Fijan, Eielts, & Pouw	2016	NL		between	97	child	observed	Y	mixed	spatial reasoning	fact	N	N/A	1.463	
Vogt & Kauschke	2017	DE	TD children	within	20	child	observed	Y	iconic	language learning	fact	N	N/A	-0.018	Dargue et al.
Wakefield & James	2015	US		between	90	child	produced	N	mixed	spatial reasoning	concept.	N	Y	0.858	Dargue et al.
Wakefield, Novack, Condgon, Franconeri, & Goldin-Meadow	2018	US		between	50	child	observed	N	mixed	math	concept.	N	Y	0.544	
Yeo, Ledesma, Nathan, Alibali, & Church	2017	US		between	82	adult	observed	Y	deictic	math	concept.	N	N	0.493	
Yuan, Gonzáles-Fuente, Baills, & Prieto	2019	CHN		between	64	adult	observed	Y	metaphoric	language learning	fact	N	Y	1.687	
Zhen, Van Hedger, Heald, Goldin-Meadow, & Tian	2019	US		between	108	adult	produced	Y	metaphoric	language learning	fact	Y	Y	1.734	
Zheng, Hirata, & Kelly	2018	US	study 2	within	24	adult	produced	Y	metaphoric	language learning	fact	N	N	0.293	

REFERENCE LIST

References marked with an asterisk indicate studies included in the meta-analysis from Study 3.

Agnetta, B., Hare, B., & Tomasello, M. (2000). Cues to food location that domestic dogs (canis familiaris) of different ages do and do not see. *Animal Cognition, 3*, 107–112. https://doi.org/10.1007/s100710000070a

*Aldugom, M., Fenn, K., & Cook, S. W. (2020). Gesture during math instruction specifically benefits learners with high visuospatial working memory capacity. *Cognitive Research: Principles and Implications*, *5*, 1-12. https://doi.org/10.1186/s41235-020-00215-8

Alexander, P. A., Willson, V. L., White, C. S., & Fuqua, J. D. (1987). Analogical reasoning in young children. *Journal of Educational Psychology, 79*, 401–408. https://doi.org/10.1037/0022-0663.79.4.401

Alibali, M. W., & Nathan, M. J. (2007). Teachers' gestures as a means of scaffolding students' understanding: Evidence from an early algebra lesson. In R. Goldman, R. Pea, B. Barron, & S. J. Derry (Eds.), *Video Research in the Learning Sciences* (pp. 348–365). Mahwah, NJ: Erlbaum.

Alibali, M. W., & Nathan, M. J. (2012). Embodiment in mathematics teaching and learning: Evidence from learners' and teachers' gestures. *Journal of the Learning Sciences*, *21*, 247–286. https://doi.org/10.1080/10508406.2011.611446

Alibali, M. W., Spencer, R. C., Knox, L., & Kita, S. (2011). Spontaneous gestures influence strategy choices in problem solving. *Psychological Science, 22*, 1138–1144. https://doi.org/10.1177/0956797611417722

Alibali, M. W., Young, A. G., Crooks, N. M., Yeo, A., Wolfgram, M. S., Ledesma, I. M., … Knuth, E. J. (2013). Students learn more when their teacher has learned to gesture effectively. *Gesture, 13*, 210–233. https://doi.org/10.1075/gest.13.2.05ali

*Allen, L. Q. (1995). The effects of emblematic gestures on the development and access of Mental representations of French expressions. *The Modern Language Journal, 79*, 521–529. https://doi.org/10.1111/j.1540-4781.1995.tb05454.x

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*, 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

*Andrä, C., Mathias, B., Schwager, A., Macedonia, M., & von Kriegstein, K. (2020). Learning foreign language vocabulary with gestures and pictures enhances vocabulary memory for several months post-learning in eight-year-old school children. *Educational Psychology Review*, *32*, 815-850. https://doi.org/10.1007

*Aussems, S., & Kita, S. (2019). Seeing iconic gestures while encoding events facilitates children's memory of these events. *Child development*, *90*, 1123-1137. https://doi.org/10.1111/cdev.12988

*Aussems, S., Mumford, K. H., & Kita, S. (2021). Prior experience with unlabeled actions promotes 3-year-old children's verb learning. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001071

*Austin, E. E., & Sweller, N. (2014). Presentation and production: The role of gesture in spatial communication. *Journal of Experimental Child Psychology, 122*, 92–103. https://doi.org/10.1016/j.jecp.2013.12.008

*Austin, E. E., & Sweller, N. (2017). Getting to the elephants: Gesture and preschoolers' comprehension of route direction information. *Journal of Experimental Child Psychology*, *163*, 1-14. https://doi.org/10.1016/j.jecp.2017.05.016

*Austin, E. E., Sweller, N., & Van Bergen, P. (2018). Pointing the way forward: Gesture and adults' recall of route direction information. *Journal of Experimental Psychology: Applied, 24*, 490–508. https://doi.org/10.1037/xap0000168

*Bailles, F., Suárez-González, N., González-Fuente, S., & Prieto, P. (2019). Observing and producing pitch gestures facilitates the learning of Mandarin Chinese tones and words. *Studies in Second Language Acquisition*, *41*, 33-58. https://doi.org/10.1017/S0272263118000074

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

*Baylor, A. L., & Kim, S. (2009). Designing nonverbal communication for pedagogical agents: When less is more. *Computers in Human Behavior, 25*, 450–457. https://doi.org/10.1016/j.chb.2008.10.008

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*. https://doi.org/10.18637/ jss.v067.i01

*Beattie, G., & Shovelton, H. (1999). Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. *Semiotica, 123*, 1–30. https://doi.org/10.1515/semi.1999.123.1-2.1

Beaudoin-Ryan, L., & Goldin-Meadow, S. (2014). Teaching moral reasoning through gesture. *Developmental Science*, *17*, 984–990. https://doi.org/10.1111/desc.12180

*Beege, M., Ninaus, M., Schneider, S., Nebel, S., Schlemmel, J., Weidenmüller, J., ... & Rey, G. D. (2020). Investigating the effects of beat and deictic gestures of a lecturer in educational videos. *Computers & Education*, *156*, 103955. https://doi.org/10.1016/j.compedu.2020.103955

Begolli, K. N., Richland, L. E., Jaeggi, S. M., Lyons, M., Klostermann, E. C., & Matlen, B. J. (2018). Executive function in learning mathematics by comparison: incorporating everyday classrooms into the science of learning. *Thinking & Reasoning, 24*, 1–34. https://doi.org/10.1080/13546783.2018.1429306

Biau, E., & Soto-Faraco, S. (2013). Beat gestures modulate auditory integration in speech perception. *Brain and Language, 124*, 143–152. https://doi.org/10.1016/j.bandl.2012.10.008

Blackwell, K.A., Chatham, C.H., Wiseheart, M., & Munakata, Y. (2014). A developmental window into trade-offs in executive function: The case of task switching versus response inhibition in 6-year-olds. *Neuropsychologia, 62*, 356-364. https://doi.org/10.1016/j.neuropsychologia.2014.04.016

Boyer, N., & Ehri, L. C. (2011). Contribution of phonemic segmentation instruction with letters and articulation pictures to word reading and spelling in beginners. *Scientific Studies of Reading*, *15*, 440-470. https://doi.org/10.1080/10888438.2010.520778

Broaders, S. C., Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2007). Making children gesture brings out implicit knowledge and leads to learning. *Journal of Experimental Psychology: General*, *136*, 539–550. https://doi.org/10.1037/0096-3445.136.4.539

Broaders, S. C., & Goldin-Meadow, S. (2010). Truth is at hand: How gesture adds information during investigative interviews. *Psychological Science*, *21*, 623–628. https://doi.org/10.1177/0956797610366082

*Buisine, S., & Martin, J.C. (2007). The effects of speech–gesture cooperation in animated agents' behavior in multimedia presentations. *Interacting with Computers, 19*, 1–10. https://doi.org/10.1016/j.intcom.2007.04.002

Capone, N. C., & McGregor, K. K. (2005). The effect of semantic representation on toddlers' word retrieval. *Journal of Speech, Language, and Hearing Research, 48*, 1468–1480. https://doi.org/10.1044/1092-4388(2005/102)

*Carlotto, T., & Jaques, P. A. (2016). The effects of animated pedagogical agents in an English-as-a-foreign-language learning environment. *International Journal of Human-Computer Studies*, *95*, 15-26. https://doi.org/10.1016/j.ijhcs.2016.06.001

*Carlson, C., Jacobs, S. A., Perry, M., & Breckinridge Church, R. (2014). The effect of gestured instruction on the learning of physical causality problems. *Gesture, 14*, 26–45. https://doi.org/10.1075/gest.14.1.02car

Cherdieu, M., Palombi, O., Gerber, S., Troccaz, J., & Rochet-Capellan, A. (2017). Make gestures to learn: Reproducing gestures improves the learning of anatomical knowledge more than just seeing gestures. *Frontiers in Psychology*, *8*, 1689. https://doi.org/10.3389/fpsyg.2017.01689

Chevalier, N., Blaye, A. (2008). Cognitive flexibility in preschoolers: The role of representation activation and maintenance. *Developmental Science, 11*, 339-353. https://doi.org/ 10.1111/j.1467-7687.2008.00679.x

Chu, M., & Kita, S. (2008). Spontaneous gestures during mental rotation tasks: Insights into the microdevelopment of the motor strategy. *Journal of Experimental Psychology: General, 137*, 706–723. https://doi.org/10.1037/a0013157

Chu, M., & Kita, S. (2011). The nature of gestures' beneficial role in spatial problem solving. *Journal of Experimental Psychology: General, 140*, 102–116. https://doi.org/10.1037/a0021790

*Church, R. B., Ayman-Nolley, S., & Mahootian, S. (2004). The role of gesture in bilingual education: Does gesture enhance learning? *International Journal of Bilingual Education and Bilingualism, 7*, 303–319. https://doi.org/10.1080/13670050408667815

*Church, R. B., Garber, P., & Rogalski, K. (2007). The role of gesture in memory and social communication. *Gesture, 7*, 137–158. https://doi.org/10.1075/gest.7.2.02bre

Church, R. B., Kelly, S. D., & Lynch, K. (2000). Immediate memory for mismatched speech and representational gesture across development. *Journal of Nonverbal Behavior, 24*, 151–174. https://doi.org/10.1023/ A:1006610013873

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.

*Colliot, T., Jamet, É. (2018). Understanding the effects of a teacher video on learning from a multimedia document: an eye-tracking study. *Education Technology Research and Development,* 66**,** 1415–1433. https://doi.org/10.1007/s11423-018-9594-x

Congdon, E. L., Kwon, M. K., & Levine, S. C. (2018). Learning to measure through action and

gesture: Children's prior knowledge matters. *Cognition, 180*, 182–190. https://doi.org/10.1016/j.cognition.2018.07.002

*Congdon, E. L., Novack, M. A., Brooks, N., Hemani-Lopez, N., O'Keefe, L., & Goldin-Meadow, S. (2017). Better together: Simultaneous presentation of speech and gesture in math instruction supports generalization and retention. *Learning and Instruction, 50*. https://doi.org/10.1016/j.learninstruc.2017.03.005

Cook, S. W., Duffy, R. G., & Fenn, K. M. (2013). Consolidation and transfer of learning after observing hand gesture. *Child Development, 84*, 1863–1871. https://doi.org/10.1111/cdev.12097

Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2008). Gesturing makes learning last. Cognition, 106, 1047–1058. https://doi.org/10.1016/j.cognition.2007.04.010

Cook, S. W., Yip, T. K., & Goldin-Meadow, S. (2012). Gestures, but not meaningless movements, lighten working memory load when explaining math. *Language and Cognitive Processes, 27*, 594–610. https://doi.org/10.1080/01690965.2011.567074

*Craig, S. D., Twyford, J., Irigoyen, N., & Zipp, S. A. (2015). A test of spatial contiguity for virtual human's gestures in multimedia learning environments. *Journal of Educational Computing Research*, *53*, 3-14. https://doi.org/10.1177/0735633115585927

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge. https://doi.org/10.4324/9780203807002

*Cutica, I., & Bucciarelli, M. (2008). The deep versus the shallow: Effects of co-speech gestures in learning from discourse. *Cognitive Science, 32*, 921–935. https://doi.org/10.1080/0364021080222203

*Dahl, T. I., & Ludvigsen, S. (2014). How I see what you're saying: The role of gestures in native and foreign language listening comprehension. *Modern Language Journal, 98*, 813–833. https://doi.org/10.1111/modl.12124

Dargue, N., & Sweller, N. (2018a). Not all gestures are created equal: The effects of typical and atypical iconic gestures on narrative comprehension. *Journal of Nonverbal Behavior, 42*, 327–345. https://doi.org/10 .1007/s10919-018-0278-3

*Dargue, N., & Sweller, N. (2018b). Donald Duck's garden: The effects of observing iconic reinforcing and contradictory gestures on narrative comprehension. *Journal of Experimental Child Psychology, 175*, 96– 107. https://doi.org/10.1016/j.jecp.2018.06.004

Dargue, N., Sweller, N. (2020). Learning stories through gesture: Gesture's effects on child and adult narrative comprehension. *Educational Psychological Review, 32***,** 249–276. https://doi.org/10.1007/s10648-019-09505-0

Dargue, N., Sweller, N., & Jones, M. P. (2019). When our hands help us understand: A meta-analysis into the effects of gesture on comprehension. *Psychological Bulletin, 14*, 765–784. https://doi.org/10.1037/bul0000202

Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia, 44*, 2037–2078. https://doi.org/10.1016/j.neuropsychologia.2006.02.006

*Davis, R.O. & Vincent, J. (2019). Sometimes more is better: Agent gestures, procedural knowledge and the foreign language learner. *British Journal of Educational Technology, 50*, 3252-3263. https://doi.org/10.1111/bjet.12732

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7,* 177–188. https://doi.org/10.1016/0197- 2456(86)90046-2

Diamond, A., Kirkham, N., & Amso, D. (2002). Conditions under which young children can hold two rules in mind and inhibit a prepotent response. *Developmental Psychology, 38*, 352–362. https://doi.org/ 10.1037//0012-1649.38.3.352

Dickersin, K. (1997). How important is publication bias? A synthesis of available data. *AIDS Education and Prevention*, *9*, 15-21

Diekelmann, S., & Born, J. (2010). The memory function of sleep. *Nature Reviews Neuroscience*, *11*, 114-126. https://doi.org/10.1038/nrn2762

Doumas, L. A. A., Morrison, R. G., & Richland, L. E. (2018). Individual differences in relational learning and analogical reasoning: A computational model of longitudinal change. *Frontiers in Psychology, 9*, 1–14. https://doi.org/10.3389/fpsyg.2018.01235

Driskell, J. E., & Radtke, P. H. (2003). The effect of gesture on speech production and comprehension. *Human Factors, 45*, 445–454. https://doi.org/10 .1518/hfes.45.3.445.27258

*Du, X., & Zhang, Q. (2019). Tracing worked examples: Effects on learning in geometry. *Educational Psychology*, *39*, 169-187. https://doi.org/10.1080/01443410.2018.1536256

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1*, 170–177. https://doi.org/10.1037/1082- 989X.1.2.170

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 455–463. https://doi.org/10.1111/j.0006-341X.2000 .00455

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj*, *315*, 629-634. https://doi.org/*https://doi.org/10.1136/bmj.315.7109.629*

Egner, T., Hirsh, J. (2005). Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. *Nature Neuroscience, 8*, 1784-1790. https://doi.org/ :10.1038/nn1594

*Eng, K., Hannah, B., Leong, L., & Wang, Y. (2013). Can co-speech hand gestures facilitate learning of non-native tones?. *Proceedings of Meetings on Acoustics of the Acoustical Society of America.* https://doi.org/10.1121/1.4799746

Ferry, A. L., Hespos, S. J., & Gentner, D. (2015). Prelinguistic relational concepts: Investigating analogical processing in infants. *Child Development, 86*, 1386–1405. https://doi.org/10.1111/cdev.12381

*Feyereisen, P. (2006). Further investigation on the mnemonic effect of gestures: Their meaning matters. *European Journal of Cognitive Psychology, 18,* 185–205. https://doi.org/10.1080/09541440540000158

*Feyereisen, P. (2009). Enactment effects and integration processes in younger and older adults' memory for actions. *Memory, 17*, 374–385. https://doi.org/10.1080/09658210902731851

*Flack, Z. M., and Horst, J. S. (2017). Two sides to every story: children learn words better from one storybook page at a time. *Infant and Child Development, 27,* e2047. https://doi.org/10.1002/icd.2047

Flevares, L. M., & Perry, M. (2001). How many do you see? The use of nonspoken representations in first-grade mathematics lessons. *Journal of Educational Psychology, 93*, 330–345. https://doi.org/10.1037//0022-0663.93.2.330

French, R. M., Glady, Y., & Thibaut, J. P. (2017). An evaluation of scanpath-comparison and machine-learning classification algorithms used to study the dynamics of analogy making. *Behavior Research Methods, 49*, 1291–1302. https://doi.org/10.3758/s13428-016-0788-z

French, R. M., & Thibaut, J. P. (2014). Using eye-tracking to predict children's success or failure on analogy tasks. *Livre/Conférence Proceedings of the Thirty-Sixth Annual Meeting of the Cognitive Science Society*, 2222–2227.

*García-Gámez, A. B., & Macizo, P. (2019). Learning nouns and verbs in a foreign language: The role of gestures. *Applied Psycholinguistics*, *40*, 473-507. https://doi.org/10.1017/S0142716418000656

Gentner, D. (1983). Structure mapping: A theoretical framework for analogy. *Cognitive Science,*

7, 155–170. https://doi.org/10.1016/S0364-0213(83)80009-3

Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child Development, 59*, 47–59. https://doi.org/10.2307/1130388

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science, 34*, 752–775. https://doi.org/10.1111/j.1551-6709.2010.01114.x

Gentner, D., Levine, S. C., Ping, R., Isaia, A., Dhillon, S., Bradley, C., & Honke, G. (2016). Rapid learning in a children's museum via analogical comparison. *Cognitive Science, 40*, 224–240. https://doi.org/10.1111/cogs.12248

Gentner, D., & Smith, L. A. (2013). Analogical learning and reasoning. In D. Reisberg (Ed.), *The Oxford handbook of Cognitive Psychology* (pp. 668-681). New York, NY: Oxford University Press.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12*, 306–355. https://doi.org/10.1016/0010-0285(80)90013-4

Glady, Y., French, R. M., & Thibaut, J. P. (2017). Children's failure in analogical reasoning tasks: A problem of focus of attention and information integration? *Frontiers in Psychology, 8,* 1–13. https://doi.org/10.3389/fpsyg.2017.00707

Glady, Y., Thibaut, J.P., & French, R. (2010). Visual strategies in analogical reasoning development: A new method for classifying scanpaths. *Proceedings of Thirty-Fifth Annual Meeting of the Cognitive Science Society*, 2398–2403. https://doi.org/10.13140/2.1.4107.1365

Glenberg A. M. & Robertson, D. A. (1999). Indexical understanding of instructions. *Discourse Processes*, *28,* 1–26. https://doi.org/10.1080/01638539909545067

Goldin-Meadow, S. (2015). From action to abstraction: Gesture as a mechanism of change. *Developmental Review, 38*, 167–184. https://doi.org/10.1016/j.dr.2015.07.007

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

Goldin-Meadow, S., Levine, S. C., Zinchenko, E., Yip, T. K., Hemani, N., & Factor, L. (2012). Doing gesture promotes learning a mental transformation task better than seeing gesture. *Developmental Science, 15*, 876–884. https://doi.org/10.1111/j.1467-7687.2012.01185

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

Goldin-Meadow, S. & Sandhofer, C. M. (1999). Gestures convey substantive information about a child's thoughts to ordinary listeners. *Developmental Science*, *2*, 67–74. https://doi.org/10.1111/1467-7687.00056

Gordon, P. C., & Moser, S. (2007). Insight into analogies: Evidence from eye movements. *Visual Cognition, 15,* 20–35. https://doi.org/10.1080/13506280600871891

Goswami, U. (1989). Relational complexity and the development of analogical reasoning. *Cognitive Development, 4*, 251–268. https://doi.org/10.1016/0885-2014(89)90008-7

Goswami, U. (2001). Analogical reasoning in children. *Proceedings of the International Conference on Conceptual Structures*, 437–470. https://doi.org/10.4324/9781315804729

Goswami, U., & Brown, A. L. (1989). Melting chocolate and melting snowmen: Analogical reasoning and causal relations. *Cognition, 35*, 69–95. https://doi.org/10.1016/0010-0277(90)90037-K

Grassmann, S., & Tomasello, M. (2010). Young children follow pointing over words in interpreting acts of reference. *Developmental Science, 13*, 252-263. https://doi.org/ 10.1111/j.1467-7687.2009.00871.x

*Guarino, K. F. & Wakefield, E. M. (2020). Teaching analogical reasoning with co-speech gesture shows children where to look, but only boosts learning for some. *Frontiers in Psychology, 11*, 1-14. https://doi.org/10.3389/fpsyg.2020.575628

Guarino, K. F., Wakefield, E, M., Morrison, R. G., & Richland, L. E. (2019). Looking patterns during analogical reasoning: Generalizable or task-specific? *Forty-First Annual Meeting of the Cognitive Science Society*, 387-392.

*Guarino, K. F., Wakefield, E, M., Morrison, R. G., & Richland, L. E. (in press). Exploring how visual attention, inhibitory control, and co-speech gesture instruction contribute to children's analogical reasoning ability.

Halford, G. S. (1993). *Children's understanding: The development of mental models.* Hillsdale, NJ: Lawrence Erlbaum

*Halvorson, K. M., Bushinski, A., & Hilverman, C. (2019). The role of motor context in the beneficial effects of hand gesture on memory. *Attention, Perception, & Psychophysics*, *81*, 2354-2364. https://doi.org/10.3758/s13414-019-01734-3

Hardison, D. M. (2003). Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics*, *24*, 495-522. https://doi.org/10.1017/S0142716403000250

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107– 128. https://doi.org/10.3102/10769986006002107

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*, 1539–1558. https://doi.org/10.1002/sim.1186

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*, 557–560. https://doi.org/10.1136/bmj.327.7414.557

*Hirata, Y. and Kelly, S. D. 2010. The effects of lips and hands on auditory learning of second language speech sounds. *Journal of Speech, Language and Hearing Research*, 53: 298–310. https://doi.org/10.1044/1092-4388(2009/08-0243

*Holle, H., Obermeier, C., Schmidt-Kassow, M., Friederici, A. D., Ward, J., & Gunter, T. C. (2012). Gesture facilitates the syntactic analysis of speech. *Frontiers in Psychology, 3,* 74. https://doi.org/10.3389/fpsyg .2012.00074

Holle, H., Obleser, J., Rueschemeyer, S. A., & Gunter, T. C. (2010). Integration of iconic gestures and speech in left superior temporal areas boosts speech comprehension under adverse listening conditions. *NeuroImage, 49*, 875–884. http://doi.org/10.1016/j.neuroimage.2009.08 .058

*Huang, X., Kim, N., & Christianson, K. (2019). Gesture and vocabulary learning in a second language. *Language Learning*, *69*, 177-197. https://doi.org/10.1111/lang.12326

Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological Bulletin, 137*, 297–315. https://doi.org/10.1037/a0022128

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica, 137*, 151–171. https://doi.org/10.1016/j.actpsy.2010.11.003

*Ianì, F., & Bucciarelli, M. (2018). Relevance of the listener's motor system in recalling phrases enacted by the speaker. *Memory*, *26*, 1084-1092. https://doi.org/10.1080/09658211.2018.1433214

*Ianì, F., Burin, D., Salatino, A., Pia, L., Ricci, R., & Bucciarelli, M. (2018). The beneficial effect of a speaker's gestures on the listener's memory for action phrases: The pivotal role of the listener's premotor cortex. *Brain and language*, *180*, 8-13. https://doi.org/10.1016/j.bandl.2018.03.001

Ioannidis, J. P. (1998). Effect of the statistical significance of results on the time to

completion and publication of randomized efficacy trials. *Jama*, *279*, 281-286. https://doi.org/10.1001/jama.279.4.281

Ji, D., & Wilson, M. A. (2007). Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nature Neuroscience, 10,* 100–107. https://doi.org/10.1038/nn1825

*Kartalkanat, H., & Göksun, T. (2020). The effects of observing different gestures during storytelling on the recall of path and event information in 5-year-olds and adults. *Journal of Experimental Child psychology*, *189*, 104725. https://doi.org/10.1016/j.jecp.2019.104725

Kelly, S. D. (2001). Broadening the units of analysis in communication: Speech and nonverbal behaviours in pragmatic comprehension. *Journal of Child Language*, *28*, 325–349. https://doi.org/10.1017/S0305000901004664

*Kelly, S. D., & Goldsmith, L. (2004). Gesture and right hemisphere involvement in evaluating lecture material. *Gesture, 4*, 25–42. https://doi.org/10 .1075/gest.4.1.03kel

Kelly, S. D., Hirata, Y., Manansala, M., & Huang, J. (2014). Exploring the role of hand gestures in learning novel phoneme contrasts and vocabulary in a second language. *Frontiers in Psychology, 5*, 1–11. http://doi.org/10.3389/fpsyg.2014.00673

*Kelly, S. D., & Lee, A. L. (2012). When actions speak too much louder than words: Hand gestures disrupt word learning when phonetic demands are high. *Language and Cognitive Processes*, *27*, 793-807. https://doi.org/10.1080/01690965.2011.581125

*Kelly, S. D., McDevitt, T., & Esch, M. (2009). Brief training with co speech gesture lends a hand to word learning in a foreign language. *Language and Cognitive Processes, 24*, 313–334. https://doi.org/10 .1080/01690960802365567

Kelly, S., Ozyürek, A., & Maris, E. (2010). Two sides of the same coin: Speech and gesture mutually interact to enhance comprehension. *Psychological Science, 21*, 260–267. https://doi.org/10.1177/0956797609357327

Kendon, A. (1986). Current issues in the study of gesture. *The biological foundations of gestures: Motor and semiotic aspects*, *1*, 23-47.

Kendon, A. (1994). Do gestures communicate?: A review. *Research on Language and Social Interaction, 27*, 175–200. https://doi.org/10.1207/s15327973rlsi2703_2

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

*Koumoutsakis, T., Church, R. B., Alibali, M. W., Singer, M., & Ayman-Nolley, S. (2016).

Gesture in instruction: Evidence from live and video lessons. *Journal of Nonverbal Behavior*, *40*, 301–315. https://doi.org/10.1007/s10919-016-0234-z

Krauss, R. M., Dushay, R. A., Chen, Y., & Rauscher, F. (1995). The communicative value of conversational hand gesture. *Journal of Experimental Social Psychology, 31*, 533–552. https://doi.org/10.1006/jesp .1995.1024

⋆Krönke, K. M., Mueller, K., Friederici, A. D., & Obrig, H. (2013). Learning by doing? The effect of gestures on implicit retrieval of newly acquired words. *Cortex*, *49*, 2553-2568. https://doi.org/10.1016/j.cortex.2012.11.016

Kushch, O., Igualada, A., & Prieto, P. (2018). Prominence in speech and gesture favour second language novel word learning. *Language, Cognition and Neuroscience, 33*, 992–1004. https://doi.org/10.1080/23273798.2018 .1435894

Lajevardi, N., Narang, N. S., Marcus, N., & Ayres, P. (2017). Can mimicking gestures facilitate learning from instructional animations and static graphics? *Computers & Education, 110*, 64–76. https://doi.org/10.1016/j.compedu.2017.03.010

Landmann, N., Kuhn, M., Piosczyk, H., Feige, B., Baglioni, C., Spiegelhalder, K., et al. (2014). The reorganisation of memory during sleep. *Sleep Medicine Review, 18*, 531–541. https://doi.org/10.1016/j.smrv.2014.03.00

Levine, S. C., Goldin-Meadow, S., Carlson, M. T., & Hemani-Lopez, N. (2018). Mental transformation skill in young children: The role of concrete and abstract motor training. *Cognitive Science, 42,* 1207–1228. https://doi.org/10.1111/cogs.12603

*Liao, M. Y., & Davidson, J. W. (2016). The effects of gesture and movement training on the intonation of children's singing in vocal warm-up sessions. *International Journal of Music Education*, *34*, 4-18. https://doi.org/10.1177/0255761415614798

Lipsey, M. W., & Wilson, D. (2001). *Practical Meta-Analysis (Applied Social Research Methods)*. Thousand Oaks, CA: SAGE Publications, Inc.

*Llanes-Coromina, J., Vilà-Giménez, I., Kushch, O., Borràs-Comes, J., & Prieto, P. (2018). Beat gestures help preschoolers recall and comprehend discourse information. *Journal of Experimental Child Psychology*, *172*, 168-188. https://doi.org/10.1016/j.jecp.2018.02.004

*Lüke, C., & Ritterfeld, U. (2014). The influence of iconic and arbitrary gestures on novel word learning in children with and without SLI. *Gesture*, *14*, 204-225. https://doi.org/10.1075/gest.14.2.04luk

*Macedonia, M., Repetto, C., Ischebeck, A., & Mueller, K. (2019). Depth of encoding through observed gestures in foreign language word learning. *Frontiers in psychology*, *10*, 33. https://doi.org/10.3389/fpsyg.2019.00033

*Macoun, A., & Sweller, N. (2016). Listening and watching: The effects of observing gesture on preschoolers' narrative comprehension. *Cognitive Development, 40*, 68–81. https://doi.org/10.1016/j.cogdev.2016.08.005

Margoliash, D., & Fenn, K. M. (2008). Sleep and memory consolidation in audition. In A. I. Basbaum, A. Kaneko, G. M. Shepherd, & G. Westheimer (Eds.), *The senses: A comprehensive reference* (pp. 895–912). San Diego, CA: Academic Press.

Markman, A. B., & Gentner, D. (1993). Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language, 32*, 517–535. https://doi.org/10.1006/jmla.1993.1027

Markman, A. B., & Wood, K. (2009). *Tools for innovation: The science behind the practical methods that drive new ideas*. New York, NY: Oxford University Press.

*Mayer, R. E., & DaPra, C. S. (2012). An embodiment effect in computer-based learning with animated pedagogical agents. *Journal of Experimental Psychology: Applied*, *18*, 239-252. https://doi.org/10.1037/a0028616

*Mavilidi, M., Okely, A. D., Chandler, P., Dylan, P., Cliff, D. P., & Paas, F. (2015). Effects of integrated physical exercises and gestures on preschool children's foreign language vocabulary learning. *Educational Psychology Review, 27*. 413-426. https://doi.org/10.1007/s10648-015-9337-z.

McGaugh, J. L. (2000). Memory--a century of consolidation. *Science*, *287*, 248-251. https://doi.org/10.1126/science.287.5451.248

McGregor, K. K., Rohlfing, K. J., Bean, A., & Marschner, E. (2009). Gesture as a support for word learning: The case of under. *Journal of Child Language*, *36*, 807–828. https://doi.org/10.1017/S0305000908009173

*McKern, N., Dargue, N., Sweller, N., Sekine, K., & Austin, E. (2021). Lending a hand to storytelling: Gesture's effects on narrative comprehension moderated by task difficulty and cognitive ability. *Quarterly Journal of Experimental Psychology*, 17470218211024913. https://doi.org/10.1177/17470218211024913

McNeill, D. (1992). *Hand and mind.* Chicago, IL: University of Chicago Press.

McNeil, N. M., Alibali, M. W., & Evans, J. L. (2000). The role of gesture in children's comprehension of spoken language: Now They need It, now they don't. *Journal of Nonverbal Behavior, 24*, 131–150. https://doi.org/10.1023/A

McNeill, D., Cassell, J., & McCullough, K.E. (1994). Communicative effects of speech-

mismatched gestures. *Research on Language and Social Interaction, 27*, 223–237. https://doi.org/10.1207/s15327973 rlsi2703_4

Merlin, T., Weston, A., & Tooher, R. (2009). Extending an evidence hierarchy to include topics other than treatment: revising the Australian levels of evidence'. *BMC medical research methodology*, *9*, 1-8. https://doi.org/10.1186/1471-2288-9-34

*Morett, L. M. & Chang, L. (2015) Emphasizing sound and meaning: pitch gestures enhance Mandarin lexical tone acquisition. *Language, Cognition and Neuroscience, 30*, 347-353, https://doi.org/10.1080/23273798.2014.923105

Morrison, R. G., Doumas, L. A. A., & Richland, L. E. (2011). A computational account of children's analogical reasoning: Balancing inhibitory control in working memory and relational representation. *Developmental Science, 14*, 516–529. https://doi.org/10.1111/j.1467-7687.2010.00999.x

Morrison, R. G., Krawczyk, D. C., Holyoak, K. J., Hummel, J. E., Chow, T. W., Miller, B. L., & Knowlton, B. J. (2004). A neurocomputational model of analogical reasoning and its breakdown in frontotemporal lobar degeneration. *Journal of Cognitive Neuroscience, 12*, 260–271. https://doi.org/10.1162/089892904322984553

*Mumford, K.H., & Kita, S. (2014). Children use gesture to interpret novel verb meanings. *Child Development, 85*, 1181–1189. https://doi.org/10.1111/cdev.12188

Murphy, A., Zheng, Y., Shivaram, A., Vollman, E., & Richland, L.E. (2021). Bias and sensitivity to task constraints in spontaneous relational attention, *Journal of Experimental Child Psychology, 202*, 1-23. https://doi.org/10.1016/j.jecp.2020.104981

Namy, L. L., & Gentner, D. (2002). Making a silk purse out of two sow's ears: Young children's use of comparison in category learning. *Journal of Experimental Child Psychology, 131*, 5–15. https://doi.org/10.1037//0096-3445.131.1.5

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*, 903-910. https://doi.org/10.1177/0956797613518351

*Novack, M. A., Goldin-Meadow, S., & Woodward, A. (2015). Learning from gesture: How early does it happen? *Cognition, 142*, 138–147. https://doi.org/10.1016/j.cognition.2015.05.018

*Ouwehand, K., van Gog, T., & Paas, F. (2015). Effects of gestures on older adults' learning from video-based models. *Applied Cognitive Psychology, 29*, 115–128. https://doi.org/10.1002/acp.3097

Paas, F., & Sweller, J. (2012). An evolutionary upgrade of cognitive load theory: Using the

human motor system and collaboration to support the learning of complex cognitive tasks. *Educational Psychology Review, 24*, 27-45. https://doi.org/10.1007/s10648-011-9179-2

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology*, *61*, 991-996. https://doi.org/10.1016/j.jclinepi.2007.11.010

*Pi, Z., Zhang, Y., Zhu, F., Xu, K., Yang, J., & Hu, W. (2019). Instructors' pointing gestures improve learning regardless of their use of directed gaze in video lectures. *Computers & Education*, *128*, 345-352. https://doi.org/10.1016/j.compedu.2018.10.006

*Ping, R., & Goldin-Meadow, S. (2008). Hands in the air: Using ungrounded iconic gestures to teach children conservation of quantity. *Developmental Psychology, 44*, 1277–1287. https://doi.org/10.1037/0012-1649.44.5.1277

Ping, R. M., & Goldin-Meadow, S. (2010). Gesturing saves cognitive resources when talking about nonpresent objects. *Cognitive Science, 34*, 602-619. https://doi.org/10.1111/j.1551-6709.2010.01102.x

Ping, R., Ratliff, K., Hickey, E., & Levine, S. C. (2011). Using manual rotation and gesture to improve mental rotation in preschoolers. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 459–464.

*Porter, A. (2016). A helping hand with language learning: teaching French vocabulary with gesture. *The Language Learning Journal*, *44*, 236-256. https://doi.org/10.1080/09571736.2012.750681

Post, L. S., Gog, T. Van, Paas, F., & Zwaan, R. A. (2013). Effects of simultaneously observing and making gestures while studying grammar animations on cognitive load and learning. *Computers in Human Behavior, 29*, 1450–1455. https://doi.org/10.1016/j.chb.2013.01.005

Pouw, W. T. J. L., de Nooijer, J. A., van Gog, T., Zwaan, R. A., & Paas, F. (2014). Toward a more embedded/extended perspective on the cognitive function of gestures. *Frontiers in Psychology, 5*, 359. https://doi.org/10.3389/fpsyg.2014.00359

*Pouw, W. T. J. L., van Gog, T., Zwaan, R. A., & Paas, F. (2016). Augmenting instructional animations with a body analogy to help children learn about physical systems. *Frontiers in Psychology*, *7*, 860. https://doi.org/10.3389/fpsyg.2016.00860

Rattermann, M. J. & Gentner, D. (1998). More evidence for a relational shift in the development of analogy: Children's performance on a causal-mapping task. *Cognitive Development, 13*, 453–478. https://doi.org/10.1016/S0885-2014(98)90003-X

*Repetto, C., Pedroli, E., & Macedonia, M. (2017). Enrichment effects of gestures and pictures on abstract words in a second language. *Frontiers in Psychology, 8*, 2136. https://doi.org/10.3389/fpsyg.2017.02136

Richland, L. E. (2015). Linking gestures: Cross-cultural variation during instructional analogies. *Cognition and Instruction, 33*, 295–321. https://doi.org/10.1080/07370008.2015.1091459

Richland, L. E., & Burchinal, M. R. (2013). Early executive function predicts reasoning development. *Psychological Science, 24*, 87–92. https://doi.org/10.1177/0956797612450883

Richland, L. E., Chan, T. K., Morrison, R. G., & Au, T. K. F. (2010). Young children's analogical reasoning across cultures: Similarities and differences. *Journal of Experimental Child Psychology, 105*, 146–153. https://doi.org/10.1016/j.jecp.2009.08.003

Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology, 94*, 249–273. https://doi.org/10.1016/j.jecp.2006.02.002

Richland, L. E., & Simms, N. (2015). Analogy, higher order thinking, and education. *Wiley Interdisciplinary Reviews: Cognitive Science, 6,* 177–192. https://doi.org/10.1002/wcs.1336

Richland, L. E., Zur, O., & Holyoak, K. J. (2007). Cognitive supports for analogies in the mathematics classroom. *Science, 316*, 1128–1129. https://doi.org/10.1126/science.1142103

Rohlfing, K. J., Longo, M. R., & Bertenthal, B. I. (2012). Dynamic pointing triggers shifts of visual attention in young infants. *Developmental Science, 15*, 426–435. https://doi.org/10.1111/j.1467-7687.2012.01139.x

Rosenthal, R. (1986). Meta-analytic procedures for social science research. Sage Publications: Beverly Hills. *Educational Researcher*, *15*, 18-20. https://doi.org/10.3102/0013189X015008018

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin, 118*, 183–192. https://doi.org/10.1037/0033-2909.118.2.183

Rowe, M. L., Ozçaliskan, S., & Goldin-Meadow, S. (2008). Learning words by hand: Gesture's role in predicting vocabulary development. *First Language, 28,* 182–199. https://doi.org/10.1177/0142723707088310

Rowe, M. L., Silverman, R. D., & Mullan, B. E. (2013). The role of pictures and gestures as nonverbal aids in preschoolers' word learning in a novel language. *Contemporary Educational Psychology, 38*, 109–117. https://doi.org/10.1016/j.cedpsych.2012.12.001

Rücker, G., Carpenter, J. R., & Schwarzer, G. (2011). Detecting and adjusting for small-study effects in meta-analysis. *Biometrical Journal Biometrische Zeitschrift, 53*, 351–368. https://doi.org/10.1002/bimj .201000151

*Rueckert, L., Church, R. B., Avila, A., & Trejo, T. (2017). Gesture enhances learning of a complex statistical concept. *Cognitive Research Principles & Implications, 2*, 2–7. https://doi.org/10.1186/s41235-016-0036-1

Rueda, M. R., Fan, J., McCandliss, B. D., Halparin, J. D., Gruber, D. B., Lercari, L. P., & Posner, M. I. (2004). Development of attentional networks in childhood. *Neuropsychologia, 42*, 1029–1040. https://doi.org/10.1016/j.neuropsychologia. 2003.12.012

Schmidt, F. L., Oh, I.S., & Hayes, T. L. (2009). Fixed vs. random models in meta-analysis: Model properties and comparison of differences in results. *British Journal of Mathematical and Statistical Psychology, 62,* 97–128. https://doi.org/10.1348/000711007X255327

Simms, N. K., Frausel, R. R., & Richland, L. E. (2018). Working memory predicts children's analogical reasoning. *Journal of Experimental Child Psychology, 166*, 160–177. https://doi.org/10.1016/j.jecp.2017.08.005

*Singer, M. A., & Goldin-Meadow, S. (2005). Children learn when their teachers' gesture and speech differ. *Psychological Science, 16*, 85–89. https://doi.org/10.1111/j.0956-7976.2005.00786.x

Singh, J. (2013). Critical appraisal skills programme. *Journal of Pharmacology and Pharmacotherapeutics*, *4*, 76.

So, W. C., Shum, P. L. C., & Wong, M. K. Y. (2015). Gesture is more effective than spatial language in encoding spatial information. *Quarterly Journal of Experimental Psychology*, *68*, 2384–2401. https://doi.org/10.1080/17470218.2015.1015431

Starr, A., Vendetti, M. S., & Bunge, S. A. (2018). Eye movements provide insight into individual differences in children's analogical reasoning strategies. *Acta Psychologica, 186*, 18–26. https://doi.org/10.1016/j.actpsy.2018.04.002

Sternberg, R. J. (Ed.). (1988). *The nature of creativity: Contemporary psychological perspectives*. New York, NY: Cambridge University Press.

*Stieff, M., Lira, M. E., & Scopelitis, S. A. (2016). Gesture supports spatial thinking in STEM. *Cognition and Instruction, 34*, 80–99. https://doi.org/10.1080/07370008.2016.1145122

Straube, B., Green, A., Weis, S., Chatterjee, A., & Kircher, T. (2009). Memory effects of

speech and gesture binding: cortical and hippocampal activation in relation to subsequent memory performance. *Journal of cognitive neuroscience*, *21*, 821-836. https://doi.org/10.1162/jocn.2009.21053

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning, 55*, 661–699. https://doi.org/10.1111/j.0023-8333.2005.00320.x

*Sweller, N., Shinooka-Phelan, A., & Austin, E. (2020). The effects of observing and producing gestures on Japanese word learning. *Acta Psychologica*, *207*, 103079. https://doi.org/10.1016/j.actpsy.2020.103079

Taylor, R., Reeves, B., Ewings, P., Binns, S., Keast, J., & Mears, R. (2000). A systematic review of the effectiveness of critical appraisal skills training for clinicians. *Medical Education, 34*, 120–5. https://doi.org/10.1046/j.1365-2923.2000.00574.x

*Theakston, A. L., Coates, A., & Holler, J. (2014). Handling agents and patients: Representational co-speech gestures help children comprehend complex syntactic constructions. *Developmental Psychology, 50*, 1973–1984. https://doi.org/10.1037/a0036694

Thibaut, J. P., & French, R. M. (2016). Analogical reasoning, control and executive functions: A developmental investigation with eye-tracking. *Cognitive Development, 38*, 10–26. https://doi.org/10.1016/j.cogdev.2015.12.002

Thibaut, J. P., French, R., & Vezneva, M. (2010). The development of analogy making in children: Cognitive load and executive functions. *Journal of Experimental Child Psychology, 106*, 1–19. https://doi.org/10.1016/j.jecp.2010.01.001

Thompson, L. A., & Massaro, D. W. (1986). Evaluation and integration of speech and pointing gestures during referential understanding. *Journal of Experimental Child Psychology, 42*, 144–168. https://doi.org/10.1016/0022-0965(86)90020-2

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need?: A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics, 35,* 215–247. https://doi.org/10.3102/1076998609346961

*Valenzeno, L., Alibali, M. W., & Klatzky, R. (2003). Teachers' gestures facilitate students' learning: A lesson in symmetry. *Contemporary Educational Psychology, 28*, 187–204. https://doi.org/10.1016/S0361-476X(02)00007-3

*van Wermeskerken, M., Fijan, N., Eielts, C., & Pouw, W. T. J. L. (2016). Observation of depictive versus tracing gestures selectively aids verbal versus visual–spatial learning in primary school children. *Applied Cognitive Psychology, 30*, 806–814. https://doi.org/10.1002/acp.3256

Viechtbauer W (2010). “Conducting meta-analyses in R with the metafor package.” *Journal of Statistical Software*, *36*, 1–48. https://doi.org/10.18637/jss.v036.i03

Viskontas, I. V, Morrison, R. G., Holyoak, K. J., Hummel, J. E., & Knowlton, B. J. (2004). Relational integration, inhibition, and analogical reasoning in older adults. *Psychology and Aging, 19*, 581–591. https://doi.org/10.1037/0882-7974.19.4.581

*Vogt, S. S., & Kauschke, C. (2017). With some help from others' hands: Iconic gesture helps semantic learning in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *60*, 3213-3225. https://doi.org/10.1044/2017_JSLHR-L-17-0004

von Hippel, P. T. (2015). The heterogeneity statistic I^2 can be biased in small meta-analyses. *BMC Medical Research Methodology, 15*, 1-8.https://doi.org/10.1186/s12874-015-0024-z

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

Wakefield, E. M., Hall, C., James, K. H., & Goldin-Meadow, S. (2018a). Gesture for generalization: Gesture facilitates flexible learning of words for actions on objects. *Developmental Science, 21*. https://doi.org/10.1111/desc.12656

*Wakefield, E. M., & James, K. H. (2015). Effects of learning with gesture on children’s understanding of a new language concept. *Developmental Psychology, 51*, 1105–1114. https://doi.org/10.1037/a0039471

*Wakefield, E., Novack, M. A., Congdon, E. L., Franconeri, S., & Goldin-Meadow, S. (2018b). Gesture helps learners learn, but not merely by guiding their visual attention. *Developmental Science, 21*. https://doi.org/10.1111/desc.12664

Wakefield, E., Novack, M. A., & Goldin-Meadow, S. (2017). Unpacking the ontogeny of gesture understanding: How movement becomes meaningful across development. *Child Development, 89*, 245-260. https://doi.org/10.1111/cdev.12817

Walker, M. P. (2005). A refined model of sleep and the time course of memory formation. *Behavioral and Brain Sciences*, *28*, 51-64. https://doi.org/10.1017/S0140525X05000026

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

Wilson, M. A., & McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science*, *265*, 676-679. https://doi.org/10.1126/science.8036517

*Yeo, A., Ledesma, I., Nathan, M.J. Alibali, M.W., & Church, R.B. (2017). Teachers' gestures and students' learning: sometimes "hands off" is better. *Cognitive Research*, 2, 1-11. https://doi.org/10.1186/s41235-017-0077-0

*Yuan, C., González-Fuente, S., Baills, F., & Prieto, P. (2019). Observing pitch gestures favors the learning of Spanish intonation by Mandarin speakers. *Studies in Second Language Acquisition*, *41*, 5-32. https://doi.org/10.1017/S0272263117000316

*Zhen, A., Van Hedger, S., Heald, S., Goldin-Meadow, S., & Tian, X. (2019). Manual directional gestures facilitate cross-modal perceptual learning. *Cognition, 187,* 178–187. https:// doi.org/10.1016/j.cognition.2019.03.004

*Zheng, A., Hirata, Y., & Kelly, S. D. (2018). Exploring the effects of imitating hand gestures and head nods on L1 and L2 Mandarin tone production. *Journal of Speech, Language, and Hearing Research*, *61*, 2179-2195. https://doi.org/10.1044/2018_JSLHR-S-17-0481

VITA

Katharine F. Guarino graduated *cum laude* with her B.A. in Psychology from Mount Holyoke College. During her undergraduate years, she worked as research assistant and teaching assistant for Dr. Katherine Binder, and lead two independent research projects under the advisorship of Dr. Francine Duetsch and Dr. Mara Breen. Upon graduation in May 2013, she began a three-year position as a lab manager in a Developmental Cognitive Neuroscience lab under the advisorship of Dr. Alison Preston and Dr. Margaret Schlichting researching the developmental trajectory of inference and memory formation. In the fall of 2016, she began her graduate training in Developmental Psychology at Loyola University Chicago under the advisorship of Dr. Elizabeth Wakefield researching the utility of co-speech gesture for teaching analogical reasoning. As a culmination of her graduate work, for her dissertation she investigated *when* gesture is most beneficial for learning across a wide range of contextual and situational variations of the learning environment.