COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 44 (2020) e12891

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12891

Analogical Comparison Promotes Theory-of-Mind Development

Christian Hoyos,[a] William S. Horton,[a] Nina K. Simms,[b] Dedre Gentner[a]

[a]*Department of Psychology, Northwestern University*
[b]*Spatial Intelligence and Learning Center, Northwestern University*

Received 4 April 2018; received in revised form 16 June 2020; accepted 13 July 2020

Abstract

Theory-of-mind (ToM) is an integral part of social cognition, but how it develops remains a critical question. There is evidence that children can gain insight into ToM through experience, including language training and explanatory interactions. But this still leaves open the question of *how* children gain these insights—what processes drive this learning? We propose that analogical comparison is a key mechanism in the development of ToM. In Experiment 1, children were shown true- and false-belief scenarios and prompted to engage in multiple comparisons (e.g., belief vs. world). In Experiments 2a, 2b, and 3, children saw a series of true- and false-belief events, varying in order and in their alignability. Across these experiments, we found that providing support for comparing true- and false-belief scenarios led to increased performance on false-belief tests. These findings show that analogical comparison can support ToM learning.

Keywords: Theory-of-mind; False beliefs; Analogy; Comparison; Relational processing

1. Introduction

Humans are uniquely talented at reasoning about the thoughts of other members of our species—a skill that is critical to our ability to cooperate intelligently and to form complex social systems. The knowledge and skills that allow us to infer mental states such as goals, intentions, and beliefs is referred to collectively as "theory-of-mind" (ToM). Because of its central role in human cognitive and social life, there is a large body of research on the development of ToM (Flavell, Green, Flavell, Watson, & Campione, 1986; Gopnik & Astington, 1988; Gopnik & Wellman, 1994, 2012; Perner, 1991; Wellman, 1990, 2014). This research has led to the development of tasks that tap into specific

Correspondence should be sent to Dedre Gentner, Department of Psychology, 2029 Sheridan Rd., Evanston, IL 60208. E-mail: gentner@northwestern.edu

aspects of ToM, as laid out by Wellman and colleagues (Peterson, Wellman, & Slaughter, 2012; Wellman & Liu, 2004). There is considerable research on the use of these tasks to assess children's understanding of different mental states (e.g., desires, beliefs, emotions).

One important class of tasks tests children's ability to reason about others' *false* beliefs. For example, in the unexpected-location task (Baron-Cohen, Leslie, & Frith, 1985; Wimmer & Perner, 1983), children watch as a character (e.g., Sally) places an object in one location; then in Sally's absence, the object is moved to a new location. Once Sally returns, the child is asked where Sally will look for the object. To answer this question correctly, children must understand that because Sally did not *see* the object change locations, she *believes* it to still be where she left it, and as a result, that is where she will *search* for it. Another commonly used task is the unexpected-contents task (Perner, Leekham, & Wimmer, 1987), in which children see a container that appears to hold one thing, but which actually holds something totally different. Children are then asked what a naïve person, who has not yet looked inside, would think is inside the box.

When someone possesses a false belief, their mental contents are not in accord with reality, and this will be reflected in their behavior. If a child can grasp this difference between mental states and reality, it suggests that they have some understanding that mental states are subjective and representational (Dennett, 1978; Perner, 1991). Thus, false-belief understanding is widely considered a critical milestone in ToM development. Children begin passing traditional false-belief tasks between the ages of 3 and 5 years, as attested by a large body of research (Peterson et al., 2012; Wellman, Cross, & Watson, 2001; Wellman & Liu, 2004).

There is also external support for the validity of tasks assessing understanding of false beliefs. Consistent with the idea that an understanding of others' minds contributes to social competency (Astington, 2003), high performance on false-belief tasks is associated with greater popularity in early childhood (Slaughter, Imuta, Peterson, & Henry, 2015). Furthermore, training in ToM improves children's ability to lie—indubitably a useful social skill (Ding, Wellman, Wang, Fu, & Lee, 2015). Overall, the evidence shows that children's performance on false-belief tasks is a good indication of their understanding of mental states as distinct from reality.

Although there is considerable support for the emergence of ToM between 3 and 5 years, there is also research suggesting that false-belief understanding may arise much earlier than previously thought, during infancy (see reviews by Barone, Corradi, & Gomila, 2019; Scott & Baillargeon, 2017). These studies use implicit, non-verbal measures such as violation of expectation and anticipatory looking in simplified false-belief tasks to assess infants' understanding of another's beliefs. However, interpretation of these findings lacks consensus (e.g., Apperly & Butterfill, 2009; Friedman & Leslie, 2005; Kovács, Téglás, & Endress, 2010; Setoh, Scott, & Baillargeon, 2016). What exactly infants and toddlers know about others' minds is an important question, and we will return to this issue in Section 6. Nonetheless, it is well-attested that preschool children improve dramatically in their ability to reason about false-belief scenarios. Understanding the skills and experiences that support false-belief improvement during the preschool period is vitally important and is the focus of the present work.

Theoretical reviews by Wellman and colleagues tracing performance on false-belief tasks across age have shown that false-belief understanding is part of a stable developmental trajectory of increasingly sophisticated reasoning about mental states (Peterson et al., 2012; Wellman & Liu, 2004). For instance, children tend to pass tasks that assess reasoning about others' desires before tasks assessing reasoning about others' beliefs. Gopnik and Wellman (1994, 2012; Wellman, 2014) have proposed an influential framework—the *theory–theory*—that characterizes the development of ToM understanding. Here, *theories* refer to overarching systems of interconnected concepts that can be used to explain and predict events.[1] Under this approach, ToM development is characterized in terms of theory change. When children are confronted with evidence that cannot be accommodated by their existing theory of how the world works, children shift to a different hypothesis that better accounts for their accumulated experience.

An important question is how children generate these hypotheses. Or, to put it another way, what happens between 3 and 5 years of age that allows children to represent the difference between true and false beliefs? Some research suggests that gains in general cognitive ability and executive function are important in the development of ToM (Benson, Sabbagh, Carlson, & Zelazo, 2013; Devine & Hughes, 2014; Perner & Lang, 1999). However, other studies (discussed below) show that ToM can be improved through training experiences, including the acquisition of supportive language. Thus, gains in ToM are at least partly driven by learning. But what kind of learning?

In the present work, we test the idea that analogical comparison and abstraction aid children's acquisition of ToM. In the following sections, we first describe training studies that show that ToM insight can be promoted by specific experiences, and then describe in more detail how analogical processes might support learning the system of relations that underlies ToM. We then present our experiments testing this claim.

1.1. Training ToM

Over and above possible contributions of gains in executive function, there is evidence from training studies that experience plays a key role in children's ToM development. A recent meta-analysis surveyed 45 training studies (N = 1,529) that exposed children to some kind of experience or intervention intended to improve their ToM (Hofmann et al., 2016). The results of the meta-analysis showed a moderate-to-large effect size, suggesting that ToM can indeed be effectively improved by experience. Many of these training studies have focused on one or both of two general approaches: providing language training (e.g., Ding et al., 2015; Hale & Tager-Flusberg, 2003; Lohmann & Tomasello, 2003), and providing or eliciting explanations and answers to questions (e.g., Amsterlaw & Wellman, 2006; Clements, Rustin, & McCallum, 2000; Lecce, Bianco, Demicheli, & Cavallini, 2014; Slaughter & Gopnik, 1996).

Several aspects of language learning have been found to be important to ToM acquisition. These include acquiring sentential complement syntax (de Villiers & Pyers, 2002; Hale & Tager-Flusberg, 2003), acquiring mental-state verbs (Moore, Pure, & Furrow,

1990; Pyers & Senghas, 2009), and engaging in discourse about mental states (Lohmann & Tomasello, 2003). As de Villiers (2005) theorized, learning to use sentential complements (e.g., "She says the doll is in the chest") involves learning the same syntactic construction that is needed for belief statements ("She thinks the doll is in the chest"). Importantly, this construction allows the truth of the embedded clause to be different from the truth of the main clause. Thus, learning sentential complement syntax may provide children with the means to represent that mental contents do not always match reality. Another aspect of language that may be important is the acquisition of mental-state verbs that use these syntactic frames, such as *think*, *know*, or *believe* (e.g., Pyers & Senghas, 2009).

Consistent with claims about the importance of language for ToM acquisition, Lohmann and Tomasello (2003) developed a training study in which they found gains in false-belief understanding after training on either discourse about mental states or sentential complement syntax, with the greatest gains occurring when children received both kinds of training. The idea that multiple elements of language contribute to false-belief understanding is further supported by a meta-analysis conducted by Milligan, Astington, and Dack (2007). On this evidence, language provides an important set of tools through which children can consider others' perspectives.

Other studies have found gains in ToM through providing children with an explanatory framework for understanding mental states—either by providing explanations invoking mental states or by eliciting explanations and answers from children together with scaffolding to help them improve their understanding. For instance, Amsterlaw and Wellman (2006) used a self-explanation approach and found that prompting children to explain a protagonist's behavior within a false-belief scenario improved their ToM across 24 sessions. Other studies have found that directly providing mentalistic explanations of a protagonist's behavior within a false-belief scenario can promote gains in ToM (Clements et al., 2000; Lecce et al., 2014; Slaughter & Gopnik, 1996). For instance, Clements et al. (2000) gave children an unexpected-location task in which children were corrected if they responded incorrectly to the key question as to where Sally would look for the toy: "Sally will look in the chest because that's where she last put it and she doesn't know that Anne moved it." Children receiving these explanations across two sessions over 2 weeks made gains from pretest to posttest. These studies show that directly providing children with explanations of behavior that invoke mental states can promote ToM development over time.

Two conclusions stand out from these training studies. First, they show that explanatory and linguistic experience can improve children's understanding of mental states. Second, the successful training studies have typically utilized a combination of many different training tasks over many sessions. Various combinations of corrective feedback, instructional explanations, self-explanation, and language training have been used to improve children's false-belief understanding. For instance, one such protocol involved six sessions over 2 weeks; each session included exposure to three different false-belief tasks and explicit training with mental-state vocabulary through the use of storybooks rich in mental-state content (Ding et al., 2015).

These studies provide critical evidence that ToM is at least partly learned, and not simply acquired through maturation. But because of the variety of training tasks used, we cannot tell which aspects of ToM interventions are critical in promoting learning (Ding et al., 2015; Hofmann et al., 2016). In particular, we cannot isolate which learning processes were involved in children's gains. That is, these studies by their nature show *that* experience can improve ToM, but not *how*. Our goal in the present research is to highlight one cognitive mechanism that may support learning in this domain. Specifically, we aim to test the role of analogical comparison in promoting ToM insights.

1.2. Why analogical comparison is relevant to the development of ToM

Comparison—the process of discovering similarities and/or differences between two things—has a long history in psychology. Traditional approaches modeled similarity either as closeness of concepts represented as points in a multidimensional space (e.g., Shepard, 1962) or as the intersection between sets of independent features (e.g., Tversky, 1977). Although both these approaches can successfully model many kinds of tasks, they fall short in situations in which it is necessary to take account of common relations (Goldstone & Son, 2005; Markman, 1989; Medin, Goldstone, & Gentner, 1993). In such cases, it is necessary to model the alignment of relational structure. In this approach, often termed "analogical comparison," comparison is not just matching features but matching relational patterns. Although models of analogical processing differ in detail, there is general agreement on the idea that analogical processing involves finding common relational structure (Doumas & Hummel, 2013; Falkenhainer, Forbus, & Gentner, 1989; Gentner, 1983, 2010; Gentner & Markman, 1997; Holyoak & Thagard, 1997; Hummel & Holyoak, 1997; for reviews, see Gentner & Forbus, 2011; Goldstone, Day, & Son, 2010; Kokinov & French, 2006).

This is important in considering ToM learning because the key insights needed for ToM are relational in nature (Bach, 2014; Baldwin & Saylor, 2005; Lohmann & Tomasello, 2003). For example, if a child hears that "X believes that Y is true," they need to understand that this describes a relation between X and Y—roughly, *believes*(*X*, *Y*), rather than "Y is true." To take a more complex example, consider the belief-desire schema described by Bartsch and Wellman (1995)—*persons who desire the content of a representation P and believe that action A will bring about the realization of P are disposed to perform action A*. This schema is part of the ToM knowledge that a child must come to understand (Bartsch & Wellman, 1995). As Bach (2014) puts it: "it is often overlooked that the belief-desire law is a relational category. The members of relational categories are united on the basis of shared relational structure, e.g. causal or functional properties (Gentner, 2005; Gentner & Kurtz, 2005; Markman & Stillwell, 2001), and they are multiply realized by objects" (p. 356). The ability to perceive common relational structure across different instances is critical for capturing ToM learning.

There is abundant evidence that analogical comparison acts to highlight common relational structure, which may then be retained in memory and applied to new instances as a more general schema (Christie & Gentner, 2010; Doumas & Hummel, 2013; Gentner,

2010; Gick & Holyoak, 1983; Goldstone et al., 2010; Loewenstein, Thompson, & Gentner, 1999; Markman & Gentner, 1993a). In particular, there is considerable evidence that young children can capitalize on analogical comparison to acquire new abstractions (Casasola, 2005; Chen & Mo, 2004; Childers & Paik, 2009; Childers et al., 2016; Christie & Gentner, 2010; Ferry, Hespos, & Gentner, 2015; Gentner & Hoyos, 2017; Gentner & Medina, 1998; Haryu, Imai, & Okada, 2011; Hoyos & Gentner, 2017; Kotovsky & Gentner, 1996; Loewenstein & Gentner, 2001; Namy & Gentner, 2002). Thus, analogical comparison provides a means through which specific experiences can be converted into more general schemas. With respect to ToM, we hypothesize that analogical processes can reveal common relational patterns—for example, across various false-belief situations—providing a means through which children can generalize across experiences to acquire an abstract understanding of how beliefs relate to behavior.

Our goal in the present work is to test whether analogical comparison can aid children's acquisition of ToM, in particular the critical insight that thoughts may differ from reality. For specificity, we use the *structure-mapping* theory of analogy to draw predictions (Falkenhainer et al., 1989; Gentner, 1983, 2003, 2010; Gentner & Forbus, 2011). Although theories of analogical processing largely agree on basic tenets, their implementations are variable, and each model addresses different phenomena more readily than others. We adopt structure-mapping theory in part because—in addition to accounting for phenomena that other models also explain (e.g., highlighting common structure; Doumas & Hummel, 2013)—it readily explains two patterns of analogical comparison that may be key for ToM development: gradual abstraction and alignable differences.

In structure-mapping theory, the key idea is that comparison involves aligning two representations according to common relational structure (*structural alignment*); on the basis of this alignment, further inferences may then be projected from one analog to the other (Falkenhainer et al., 1989; Gentner & Markman, 1997; Markman & Gentner, 1993a; Medin et al., 1993). The commonalities and differences between two situations are found by determining the maximal (or near-maximal) structurally consistent alignment between the representations of the two situations. This common relational structure can form the basis for a category abstraction—a relational schema that may be applied to other instances, allowing the learner to infer missing information in the new situation. As new instances are aligned with the schema, it is further abstracted (Ferry et al., 2015; Kuehne, Forbus, Gentner, & Quinn, 2000). Thus, although the initial schema may be context-specific, applying it to further instances results in gradual abstraction (Bach, 2014; Gentner & Medina, 1998).

Structure-mapping processes can also serve to highlight *alignable differences*—differences that play the same role in the common structure—both in adults (Gentner & Gunn, 2001; Markman & Gentner, 1993b; Sagi, Gentner, & Lovett, 2012) and in children (Gentner et al., 2016; Hoyos & Gentner, 2017). This is important in the current context because understanding of ToM requires perceiving the difference between situations in which mental states correspond with reality (true belief) and those in which mental states differ from reality (false belief). Noticing these differences necessarily involves a comparison, either between mental content and the state of affairs in the world or between the

thoughts of different people. Indeed, some studies have found that children are more successful on false-belief tasks when the scenarios incorporate a second character with a different perspective or belief from the main character (Lewis, Hacquard, & Lidz, 2012; Pham, Bonawitz, & Gopnik, 2012). These scenarios provide children the opportunity to make an informative contrast that can highlight the critical alignable difference in the characters' knowledge or beliefs and bolster subsequent reasoning.

Of course, we are not claiming that analogical comparison is the only way children gain insight into mental states. On the contrary, other kinds of experience, such as explanation and language-learning, are likely contributors to this learning. Nor are we the first to propose that structure mapping is a key mechanism in ToM development. For instance, Bach (2014) reviewed theories of ToM learning and suggested that structure mapping offers a plausible mechanism for how children could gain such insight from experience: "A central appeal of Gentner and colleagues' *structure-mapping theory* is its explanation for how learners develop relational category knowledge. The core idea is that comparison, which involves the structural-alignment of two representations, induces an epistemic focus on relations while demoting the epistemic importance of object-realizers. Such highlighted relations can then be abstracted and serve as the basis for relational concepts that describe relational categories" (p. 356). In a related proposal, Baldwin and Saylor (2005) suggested that structure-mapping processes coupled with linguistic interaction could drive ToM development. Their reasoning is that since the mind is invisible, linguistic interactions serve as indicators of mental states: "The overarching hypothesis predicts that language will frequently serve to trigger comparison, and hence structure mapping, across distinct behavioral scenarios that could support, among other things, inferences about internal, mentalistic causes that underlie others' behavior" (Baldwin & Saylor, 2005, p. 138). For example, when children hear the same mental-state verbs used in similar syntactic constructions across situations, this may invite alignment across those situations that will promote the abstraction of belief relations. For instance, a child who hears *Johnny thinks that the box has crayons inside* and *Abby thinks that it has blocks inside* may compare the two situations (and the two locutions) and arrive at a common insight. Moreover, the mutual bootstrapping hypothesis (Gentner, 2010) suggests that the connection between comparison and word learning can work in both directions: Hearing similar words across different situations invites comparison between the situations, and observing similarities between situations can help children learn the meanings of those words (Gentner, 2003, 2010, 2016; Gentner & Medina, 1998). This could help explain why caregivers' mental-state language relates to children's ToM (e.g., Ruffman, Perner, & Parkin, 1999).

In the present work, we seek to provide a more direct test of the role of analogical comparison in ToM development. Our studies were guided by four principles derived from structure-mapping theory and research:

1. The process of structural alignment renders common structure more salient, thus promoting schema abstraction (Christie & Gentner, 2010; Doumas & Hummel, 2013; Gentner & Medina, 1998; Gentner & Namy, 1999, 2006; Gick & Holyoak, 1983).

2. Comparison between highly alignable events can lead to highlighting of alignable differences (contrast; Gentner & Markman, 1994; Markman & Gentner, 1993b, 1996; Sagi et al., 2012).
3. High overall similarity between situations both invites spontaneous comparison and facilitates structural alignment (Gentner & Kurtz, 2006; Gentner & Toupin, 1986; Hoyos & Gentner, 2017).
4. Although the initial schema formed by comparing two examples will often be concrete and context-specific, once a schema has been formed, further examples can be compared with it, resulting in gradual abstraction of the schema (*progressive alignment*); thus, the less the learner knows about a domain, the more dependent they are on overall similarity and other supports to perceive and align relational structure (Kotovsky & Gentner, 1996; Kuehne, Forbus, et al., 2000; Kuehne, Gentner, & Forbus, 2000).

1.3. Testing the analogical comparison hypothesis

The evidence reviewed thus far has shown that experience is an important driver of ToM development. However, it remains unclear what cognitive mechanisms are deployed to benefit from these experiences. Our studies test one specific hypothesis: that analogical comparison can foster insight into ToM, as assessed by improved performance on false-belief tasks. We also investigate whether relational language interacts with comparison processes in ToM learning, as suggested by Baldwin and Saylor (2005; see also Gentner, 2003, 2010; Gentner & Christie, 2010).

In Experiment 1, we intentionally take a heavy-handed approach, engaging children in many guided comparisons to provide an initial proof-of-concept that analogical comparisons can support false-belief reasoning. Specifically, we test whether asking children to carry out explicit comparisons between one character's thoughts and another's, and between thoughts and reality, will help them understand the idea of false belief (i.e., that two individuals can hold differing mental states and that thoughts may differ from reality). As described above, an important aspect of our hypothesis is that analogical comparison will reveal commonalities across events and promote abstraction of a common schema (Principle 1 above). In the remaining studies, we explore how children's spontaneous comparisons might be elicited and supported to facilitate false-belief reasoning. In these studies, we also take a more targeted approach to test specific hypotheses that arise from structure-mapping theory, namely that high overall similarity should invite and support comparison (Experiments 2a and 3), that successful comparisons yield schemas that potentiate further alignments (Experiment 2b), that comparisons highlight important (alignable) differences (Experiments 2a, 2b, and 3), and that younger or less knowledgeable learners will require greater support for fruitful comparisons (Experiments 2a and 3).

2. Experiment 1

The goal of Experiment 1 was to test whether children could make gains in false-belief reasoning in a single session when given comparison-based training. Using a pretest/

posttest procedure, we created three between-subjects training conditions. In the key experimental condition (*Compare-Thoughts*), we showed children a scenario in which one character has a true belief about the contents of a box and another character has a false belief. Within this scenario, we had children compare each character's mental state with reality, and also compare the characters' mental states. Then, to solidify children's grasp of these distinctions, we showed them a second true-belief/false-belief scenario. After comparing mental states the same ways within the second scenario, children were shown the two scenarios simultaneously and were asked to find the correspondences between them. In sum, children made three kinds of comparisons during training: (a) between a character's belief and the state of the world, (b) between a character holding a true belief and a character holding a false belief, and (c) between different situations involving characters with true and false beliefs (see Fig. 1 for a sample comparison pair). This condition tested our prediction that intensive analogical comparison of beliefs would help children gain insight into ToM.

In the two control conditions, children were not shown any mental-state content. In the *Compare-Items* condition, children saw event scenarios that did not depict or discuss beliefs and answered comparison questions, but these questions concerned objects, not mental

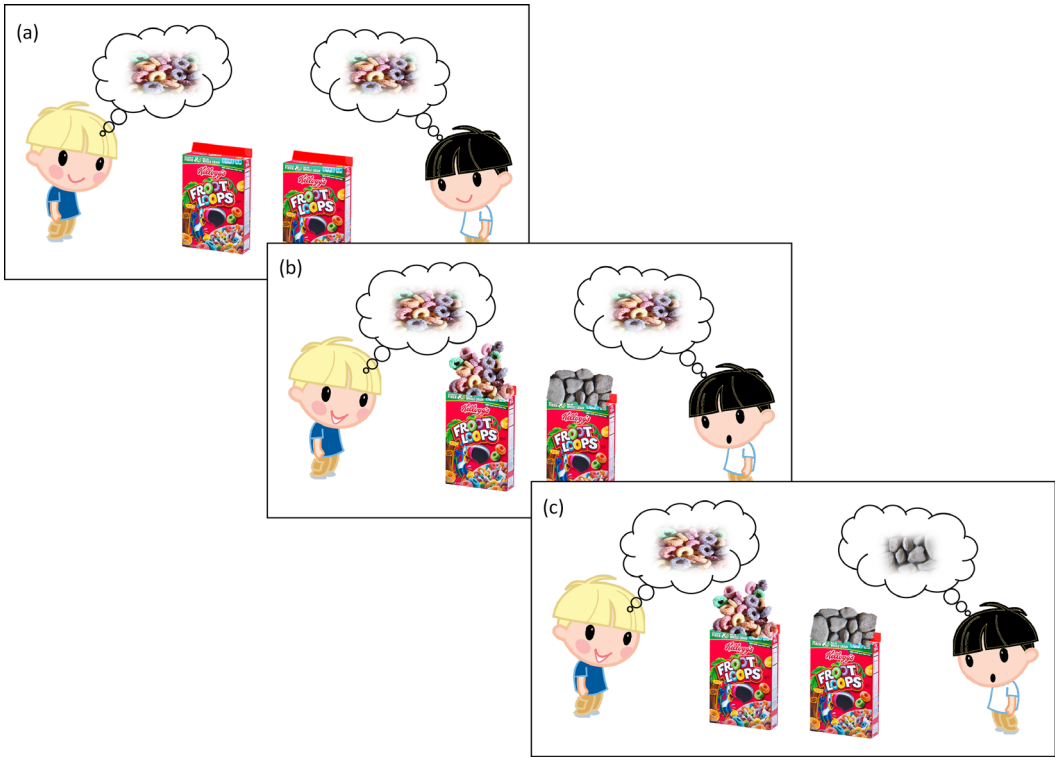

Fig. 1. The first training scenario used in the Compare-Thoughts condition of Experiment 1 (a) before the contents were revealed, (b) immediately after the contents were revealed, and (c) at the final outcome.

states. This condition was included in response to findings in adults showing improved perspective-taking following an unrelated comparison task (Todd, Hanko, Galinksy, & Mussweiler, 2011). Thus, this condition tested whether practice making comparisons more generally (as opposed to specifically comparing thoughts, as in the key experimental condition) would improve false-belief performance, and whether that would be sufficient to explain any improvements in the Compare-Thoughts condition. In the *Baseline* condition, children had no intervening task between the pretest and posttest. We predicted that because the comparisons that children made in the Compare-Thoughts condition should highlight important relational structure directly related to true and false beliefs, children in this condition should outperform children in the control conditions on the posttest.

In addition, in Experiment 1 (and in each of the experiments we report here), we also tested for potential effects of gender on false-belief learning. Although most studies measuring false-belief task performance in 3- to 5-year-old children have not bothered to examine gender separately, there are a handful of reports of better performance by girls (e.g., Cutting & Dunn, 1999; Walker, 2005). In one study directly examining potential gender differences in the false-belief development of children age 2–6 (Charman, Ruffman, & Clements, 2002), there was a "significant but weak advantage" for girls compared to boys, and notably, in the meta-analysis of ToM training studies conducted by Hofmann et al. (2016), gender was a marginal ($p = .06$) moderator of training effect sizes. Thus, although the effects of gender on false-belief understanding do not appear to be overwhelming, we wanted to include gender in our analyses to probe whether our training procedures would have a differential effect across girls and boys.

Finally, our pretest and posttest tasks were designed to assess both near and far transfer. For near transfer, we used an unexpected-contents task—the same kind of task as that used in the training phase. To assess far transfer, we used two other standard false-belief tasks widely used in the developmental literature—an unexpected-location task and a verbal false-belief task. In each of these tasks, different versions were run in the pretest and posttest, and the specific content differed from that present in the training phase scenarios. An important question is the degree to which children will transfer beyond the unexpected-contents training scenario to pass the other false-belief tasks.

2.1. Methods

2.1.1. Participants

Eighty-two 4.5- to 5-year olds from the greater Evanston/Chicago area participated in this study (39 females, mean age = 56 months, range = 54–61 months). Six additional children were tested but excluded from analysis for failing to complete the experiment ($n = 6$), and/or not understanding English ($n = 3$). Another 18 children were excluded for ceiling performance in the pretest. The racial and economic composition of the sample reflected the local population, with the majority of children coming from European American, middle- and upper-middle-class families. Children received a small gift for their participation.

Children were randomly assigned to the Compare-Thoughts (n = 28, 13 females, mean age = 57 months, range = 54–61 months), Compare-Items (n = 26, 12 females, mean age = 56 months, range = 54–59 months), or Baseline (n = 28, 14 females, mean age = 57 months, range = 54–59 months) conditions. There were no differences across conditions in child age, $F(2, 79) = 0.96$, $p = .39$, $\eta^2 = .02$, or gender, $\chi^2(2) = 1.02$, $p = .95$, $V = .04$. Importantly, children also showed no differences across groups in false-belief performance on the pretest, $F(2, 79) = 1.05$, $p = .36$, $\eta^2 = .03$.

2.1.2. Design and procedure

The pretest and posttest were each composed of three different false-belief tasks. The same three types of tasks were used in the pretest and posttest, and the order of the tasks was counterbalanced across participants. The dependent measure was how many false-belief tasks children passed at posttest. We compared children's posttest performance, controlling for pretest performance, across the three groups.

2.1.2.1. False-belief pretest/posttest: Children participated individually, either at Northwestern University or at the child's preschool. The false-belief tests were displayed on a laptop. Although most of the work on false-belief understanding in children has involved puppets or dolls, presentation of false-belief scenarios via laptop or tablet computer has become increasingly common (e.g., Liu, Sabbagh, Gehring, & Wellman, 2009; Pellicano, 2007; Wu, Haque, & Schulz, 2018). A meta-analysis by Wellman Cross and Watson (2001) found no impact on differences on ToM development due to task- or medium-related factors, including whether the scenario protagonist and target objects are real objects or images or presented via video. Thus, for ease of administration, both our training procedures and pretest/posttest tasks were conducted via laptop.

Using PowerPoint, simplified images of characters and events were displayed in semi-animated fashion on the laptop computer. Children first completed the diverse-desires task (Repacholi & Gopnik, 1997; Wellman & Liu, 2004; Wellman & Woolley, 1990)—an easy task for 4-year olds which was meant to serve as a warm-up task (see Supplementary Material for full description of prompts used in tasks and training). Then children completed the pretest, composed of three different false-belief tasks. These included the unexpected-location task (Baron-Cohen et al., 1985; Wimmer & Perner, 1983), the unexpected-contents task (Perner et al., 1987; Wellman & Liu, 2004), and a verbal false-belief task (Siegal & Beattie, 1991; Wellman & Bartsch, 1989; Wellman & Liu, 2004). In all tasks, children had to answer both a memory question and a target question. The target questions were the key questions that probed for children's false-belief understanding. The purpose of the memory questions was to ensure that children were paying attention and following the information presented. No feedback was given.

At the posttest, each child completed the same three false-belief tasks, with different specific scenarios. These tasks were used to measure possible changes in false-belief reasoning following the training procedures, with the unexpected-contents task probing near transfer, and the unexpected-location and verbal false-belief tasks probing far transfer (although note that "near transfer" and "far transfer" only apply to the Compare-Thoughts

group, who received training with unexpected-contents scenarios). The specific version of each task type assigned to either the pretest or the posttest was counterbalanced across children and conditions. In this manner, each child completed six false-belief tasks total —three at the pretest and three at the posttest—with each task involving both a target question and a memory question.

2.1.2.2. Thought-bubbles training: Because our primary training task involved animations that represented characters' mental states through visual thought bubbles, we wanted to ensure that the children in our study understood the role of thought bubbles in visually representing another person's thoughts. As shown by Wellman, Hollander, and Schult (1996), even minimal experience with visual representations of characters' thoughts can allow young children to appreciate thought-bubble depictions as "showing what someone is thinking." Thus, following the pretest tasks, all children were given brief training on thought bubbles, adapted from Wellman et al. (1996).

In our study, children saw a screen depicting a character with a thought bubble. The experimenter explained to the child that looking inside the thought bubble showed what someone was thinking. The child was asked to report what was inside the thought bubble, and then in a second image, to distinguish between what the character was doing and what the character was thinking. All children received thought-bubbles training, regardless of condition, even though only children in the experimental condition (Compare-Thoughts) saw thought bubbles during training. This was done to avoid the possibility that any gains seen in the Compare-Thoughts condition could be attributed to the thought-bubble training itself.[2] No thought bubbles were used in the pretest and posttest.

2.1.2.3. Training conditions: After the thought-bubbles training, children were randomly assigned to one of three training conditions: Compare-Thoughts, Compare-Items, or Baseline. In the Compare-Thoughts condition, children saw two containers and two characters involved in an unexpected-contents situation. In the classic version of the unexpected-contents task (e.g., Perner et al., 1987), children are shown a box that appears to contain one thing but contains something different. After the child is shown the box's true contents, they are introduced to a character who has never seen inside the box, and the child is asked what the character thinks is inside the box. Young children often incorrectly answer that the character will already know what the box contains. For our training sequence, thought bubbles displayed what the character thought was inside the box. This allowed us to ask children to compare mental states as well as states of the world.

In the Compare-Thoughts condition, children initially saw two cereal boxes, which opened to reveal that one contained cereal and the other contained rocks. Then the boxes were closed and two characters were introduced. Thought bubbles showed that each character thought his box contained cereal (see Fig. 1a). The child first asked to directly compare the characters' mental states, as depicted in their thought bubbles: "Are Jay and Luke thinking the same or different?" Then children were asked to contrast the actual

contents of the boxes: "Do the boxes contain the same or different things?" Next, the contents of the boxes were revealed to the characters. For each character, the experimenter asked, "Was he thinking the same or different than what was inside the box?" This question was intended to prompt the child to compare mental states with reality—revealing that the character either had a true belief or a false belief. Importantly, each question could be answered on the basis of the visual information in the training scenario—they did not require false-belief understanding by the child. Rather, we hypothesized that engaging in this series of comparisons would lead children to notice the difference between true and false beliefs. Children nearly always answered these questions correctly. Their answers were confirmed by the experimenter (e.g., "Yes! He's thinking the same!"). See Supplementary Material for full description of procedure.

After this, children were presented with a second unexpected-contents scenario, parallel to the first scenario but with new boxes, contents, and characters. The same sequence of questions was repeated for this scenario. After this second scenario was completed, the two scenes—each with its own boxes and its own characters—were shown simultaneously, and children were asked to identify what was the same between the two stories: "Remember these two stories? Can you tell me what's the same between these two stories?" The goal was to promote structural alignment between the situations and thereby foster noticing the common relational structure. Thus, this training procedure had children make three distinct types of comparisons: comparisons between characters' mental states, comparisons between mental states and reality, and comparison across two situations that had these relations embedded within them.

The Compare-Items condition was designed to test whether performance in the experimental condition could be due simply to engaging in comparison itself, regardless of content. In this condition, children were asked to make comparisons between objects, not thoughts; the scenarios did not involve hidden contents or any reference to or depiction of mental states. One scenario in this condition, for example, showed children two characters, each of whom had brought various items to a picnic. Children were asked to make comparisons between the items. There were three pairs of items, and for each pair, children were asked, "Did Jay and Luke bring the same things to the picnic?" Children completed two scenarios with this line of questioning. This training procedure had a similar number of comparison questions to the Compare-Thoughts condition.

The Baseline condition had no intervening task between the pretest and posttest; children went directly from the thought-bubbles training procedure to the posttest.

2.2. Results

2.2.1. Analysis plan

For each false-belief task in the pretest and posttest, children were given a score of 1 if they answered *both* the target and memory questions correctly; children with perfect pretest scores were excluded. Pretest and posttest means are displayed in Table 1.

Table 1
Means (and *SD*s) of theory-of-mind (ToM) scores across training conditions

	Condition	Pretest[a]	Posttest[a]	Difference Score
Experiment 1	Compare-Thoughts	0.71 (0.66)	1.46 (0.96)	0.75 (1.00)
	Compare-Items	1.00 (0.80)	1.23 (1.03)	0.23 (0.65)
	Baseline	0.93 (0.81)	1.18 (1.16)	0.25 (0.70)
Experiment 2a	High Alignability (HA)	1.38 (0.81)	2.12 (0.97)	0.75 (0.84)
	Low Alignability (LA)	1.12 (0.78)	1.41 (0.95)	0.29 (0.93)
Experiment 2b	Reversed	1.05 (0.73)	1.24 (0.88)	0.18 (0.80)
Experiment 3	High Alignability (HA)	0.85 (0.77)	1.20 (0.88)	0.35 (0.92)
	Low Alignability (LA)	0.46 (0.60)	0.64 (0.81)	0.18 (0.82)

[a]Pretest/posttest ToM scores are out of a possible three points; however, children with perfect pretest scores were excluded from analysis.

2.2.2. Training effects

To assess whether children in each condition improved between the pretest and posttests, we calculated gain scores by subtracting the number of tasks the child passed in the pretest (out of 3) from the number of tasks they passed in the posttest (out of 3) and compared these to zero. Because we excluded children with perfect pretest scores, these difference scores could theoretically range from −2 to 3; however, the actual range was −1 to 3. The mean gain in the Compare-Thoughts condition was significantly greater than zero ($M = 0.75$, $SD = 1.00$), $t(27) = 3.95$, $p = .001$, $d = 0.75$. Mean gain in the Compare-Items condition was marginally greater than zero ($M = 0.23$, $SD = 0.65$), $t(25) = 1.81$, $p = .083$, $d = 0.35$, as was the mean gain in the Baseline condition ($M = 0.25$, $SD = 0.70$), $t(27) = 1.89$, $p = .070$, $d = 0.36$.

To examine differences and interactions across conditions, an ANCOVA with pretest score as a covariate, posttest score (out of 3) as the dependent variable, and condition and gender as between-subjects factors was conducted. Pretest scores significantly predicted posttest scores, $F(1, 75) = 60.02$, $p < .001$, $\eta^2 = .45$. The analysis revealed a significant main effect of condition, $F(2, 75) = 4.16$, $p = .019$, $\eta^2 = .10$. Planned contrasts indicated that children in the Compare-Thoughts condition ($M^3 = 1.64$, $SE = 0.14$) demonstrated better false-belief understanding on the posttest than children in either the Compare-Items condition ($M = 1.15$, $SE = 0.14$), $p = .018$, $d = 0.66$, or the Baseline condition ($M = 1.13$, $SE = 0.14$), $p = .012$, $d = 0.69$. The Compare-Items and Baseline conditions did not significantly differ from each other, $p = .925$, $d = 0.03$. See Fig. 2 for pretest and posttest performance by condition.

Interestingly, there was also a significant main effect of gender, $F(1, 75) = 14.53$, $p < .001$, $\eta^2 = .16$. Females ($M = 1.62$, $SE = 0.12$) demonstrated better false-belief understanding on the posttest than males ($M = 1.00$, $SE = 0.11$). There was also a marginal interaction between condition and gender, $F(2, 75) = 2.87$, $p = .063$, $\eta^2 = .07$. Bonferroni post-hoc tests showed that females showed the predicted pattern: Those in the Compare-Thoughts condition ($M = 2.16$, $SE = 0.20$) had marginally better posttest

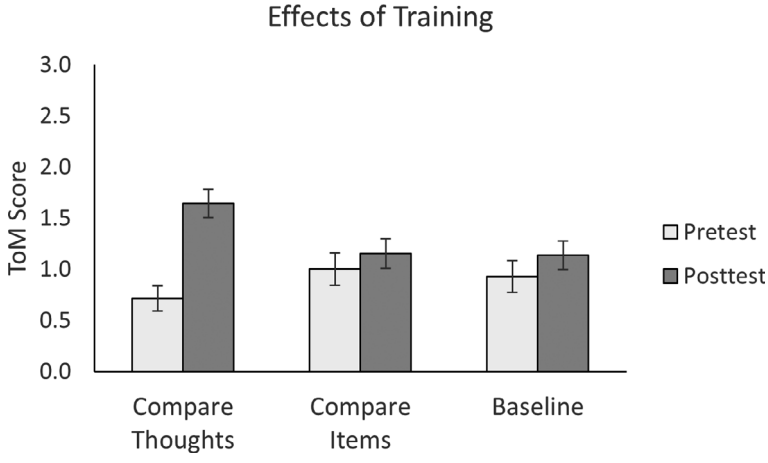

Fig. 2. Pretest and posttest performance across conditions in Experiment 1. Pretest performance did not differ across conditions, but children in the Compare-Thoughts condition scored significantly higher on the posttest than children in the control conditions after controlling for pretest scores. Plotted posttest means are estimated marginal means; error bars depict ±1 *SEM*.

performance than those in the Compare-Items condition (M = 1.51, SE = 0.21), p = .102, d = 0.87, and significantly better posttest performance than those in the Baseline condition (M = 1.19, SE = 0.20), p = .003, d = 1.32. For males, performance in the Compare-Thoughts condition (M = 1.12, SE = 0.19) did not differ from performance in the Compare-Items condition (M = 0.79, SE = 0.19), p = .679, d = 0.45, or the Baseline condition (M = 1.08, SE = 0.19), p = 1.00, d = 0.06.

2.2.3. Near and far transfer

An important question is whether at posttest, the Compare-Thoughts group transferred beyond the unexpected-contents task (that closely resembled training) to improve on the other two false-belief tasks (the unexpected-location and verbal false-belief tasks). If they are only able to show near transfer (i.e., to the unexpected-contents task), this will suggest that the schema they formed during training was context-specific. The breakdown of performance by task is shown in Fig. 3. To examine this, we analyzed performance on the near-transfer task and the two far-transfer tasks separately. Our primary interest is in whether and how far the Compare-Thoughts group was able to transfer. However, we also analyzed performance in the Compare-Items and Baseline groups. Since these two groups did not receive training with the unexpected-contents task, the distinction between near and far transfer does not apply. Nevertheless, we include them for calibration.

2.2.3.1. Near transfer: To examine improvement on the near-transfer task, we conducted McNemar's exact tests. There was a significant pre-to-post increase in children passing the unexpected-contents task in the Compare-Thoughts group, p = .001, V = .38. This was not the case for the Compare-Items group, p = 1.00, V = .76, nor for the Baseline group, p = .453, V = .52.

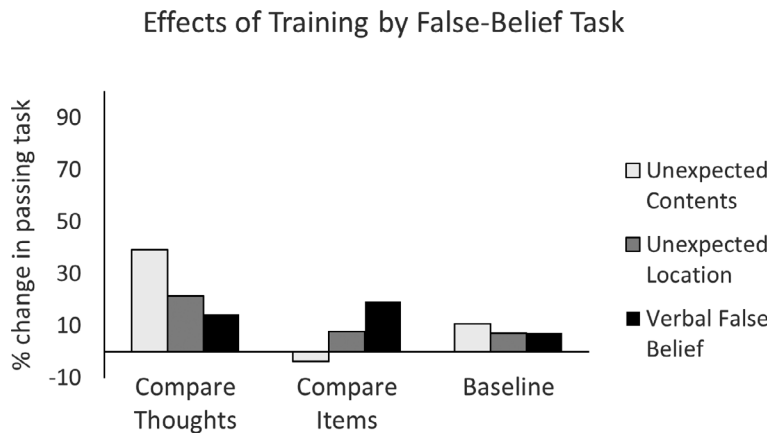

Fig. 3. Change in percentage of children passing each task from pretest to posttest in Experiment 1. This shows that the gains made by children in the Compare-Thoughts condition are primarily in the unexpected-contents task (the near-transfer task).

To examine condition differences, we conducted a multinomial logistic regression with pretest as a covariate and condition as a fixed factor. Including sex and the condition by sex interaction did not improve the model and so these factors were excluded. The final model was significant, $\chi^2(3) = 36.08$, $p < .001$. The Compare-Thoughts group was significantly more likely to improve on the near-transfer task, relative to the Baseline group, $\beta = 2.02$, Wald's $\chi^2 = 6.81$, $p = .009$; the Compare-Items group was not, $\beta = -0.68$, Wald's $\chi^2 = 0.77$, $p = .379$. Thus, for the near-transfer task, of the two training groups, only the Compare-Thoughts group outperformed the Baseline group on the posttest; and only the Compare-Thoughts group showed improvement from pretest to posttest.

2.2.3.2. Far transfer: To examine improvement from pretest to posttest on these tasks, we computed new pretest and posttest scores, excluding performance on the unexpected-contents task. Gain scores were compared to zero. Mean gains in the Compare-Thoughts group were significantly greater than zero ($M = 0.36$, $SD = 0.83$), $t(27) = 2.29$, $p = .030$, $d = 0.43$, as were those in the Compare-Items group ($M = 0.27$, $SD = 0.45$), $t(25) = 3.04$, $p = .006$, $d = 0.60$. Mean gains in the Baseline group were not significantly different from zero ($M = 0.14$, $SD = 0.59$), $t(27) = 1.28$, $p = .212$, $d = 0.24$.

To compare across conditions, we conducted an ANCOVA parallel to the one examining overall performance using these new scores. We found no effect of condition, $F(2, 75) = 0.97$, $p = .384$, $\eta^2 = .03$. This suggests that the two training groups demonstrated similar overall performance in the transfer tasks, and that the greater overall performance in the Compare-Thoughts condition stemmed largely from passing the unexpected-contents task. There was an effect of sex, $F(1, 75) = 15.64$, $p < .001$, $\eta^2 = .17$, as well as a marginal sex by condition interaction, $F(2, 75) = 2.64$, $p = .078$, $\eta^2 = .07$, that mirrored the patterns in the overall analysis.[4] Females did better on the

far-transfer tasks than males, and this may have driven the gender effects seen in the overall analysis. However, we are unable to draw strong conclusions about gender because we did not find effects of sex in any of the further three studies.[5]

2.3. Discussion

The results of Experiment 1 provide support for the hypothesis that comparison processes support gaining insight into mental states. Children in the Compare-Thoughts condition demonstrated significantly better false-belief understanding on the posttest than the groups who did not compare thoughts (controlling for pretest performance). The advantage of the Compare-Thoughts condition on the near-transfer task cannot be attributed to comparison processing per se, because the Compare-Items group performed no differently than the Baseline group. It also cannot be attributed to the presence of thought bubbles because all three groups experienced thought bubbles. These results provide evidence that analogical comparison can highlight commonalities and distinctions that are critical for supporting ToM insights.

There were no condition differences on the far-transfer tasks, suggesting that the advantage of the Compare-Thoughts condition was limited to the near-transfer task—that is, to the unexpected-contents task, whose relational structure closely matches that of the training scenario. However, both the Compare-Thoughts and Compare-Items groups made significant gains on the far-transfer tasks from pretest to posttest. Inspection of Fig. 3 suggests that the Compare-Thoughts group made gains on near-transfer and on one of the far-transfer tasks—the unexpected-location task. This pattern in the Compare-Thoughts group could suggest that there was some degree of transfer from unexpected-contents to unexpected-location. The pattern in the Compare-Items group is more puzzling; this group appears to have made gains on only one task—verbal false belief. It's possible that being prompted to compare conferred some benefit (e.g., Todd et al., 2011), or that verbal interaction or some other aspect of the training procedures led to these gains. However, it is unclear why such benefits would have appeared selectively for only this task; the Compare-Items group made no gains in the near-transfer task. Regardless, being asked to compare, per se, and engaging in ancillary features of the training are not sufficient by themselves to account for the complete pattern seen here.

Also, interestingly, the effect appeared to be more pronounced for females. This is consistent with earlier work showing a female advantage in ToM (Charman et al., 2002) and with evidence for the positive effects of gender on gains in false-belief performance through training (Hofmann et al., 2016). However, we do not find gender effects in any of our three subsequent studies; furthermore, as reviewed in Section 6, evidence for gender effects in ToM is quite mixed.

The results of Experiment 1 provide support for the idea that analogical comparison can facilitate false-belief understanding. However, the concentrated comparison experience provided here serves more as a proof-of-concept than as evidence that comparison processes drive ToM learning in the child's typical experience. In the Compare-Thoughts condition, children were asked to compare mental states to states of the world, true-belief

mental states to false-belief mental states, and whole situations involving true and false mental states to each other. Clearly, this level of intensive, pedagogically directed comparison is not likely to happen in real life. Thus, in Experiment 2, we investigated whether children could gain insight from their own spontaneous comparisons. Instead of asking children direct comparison questions, we showed children a sequence of short vignettes depicting true or false beliefs and asked questions about individual vignettes. The idea is that, provided that the vignettes are sufficiently similar, children will spontaneously compare them; and this may help children gain insight into the distinction between true and false beliefs.

Another key factor in Experiment 2 is the order in which children received the vignettes. In Experiment 2a, we chose an order that has been investigated in prior analogical research and found to be effective in promoting comparison—the *repetition-break* pattern (Loewenstein & Heath, 2009). In the repetition-break pattern, two or more highly alignable events are stated in sequence, fostering alignment and schema abstraction (Principle 1). This is followed by a final event that is highly alignable, but differs in some key way. When this final event is aligned with the schema, then this alignable difference should "pop out" (Principle 2). This narrative structure is widely used in fables and folk stories as a means of creating interest and highlighting a key point (Loewenstein, Raghunathan, & Heath, 2011). Notably, this pattern is often used in children's stories (e.g., Three Billy Goats Gruff, Three Little Pigs). For example, in the Three Little Pigs, three pigs each build a house. The first little pig builds a house of straw, but a wolf blows it down and devours him. The second little pig builds a house of sticks; but the wolf also blows this house down and devours the second pig. The third little pig builds a house of bricks, and again the wolf tries to blow the house down. But this time the wolf fails; the house is too strong. Instead of devouring the third pig, the wolf is soundly defeated.[6] This narrative pattern may be especially effective for children because the first two situations are similar enough to be aligned even by a young child, leading the child to form a common schema. When the third situation is encountered, it is close enough to align with the schema—with the result that the difference pops out as an alignable difference.

In Experiment 2a, our goal was to allow children to note the contrast between true and false beliefs. To do this, we constructed a repetition-break sequence that began with two analogous true-belief stories and ended with a false-belief story whose relational structure largely matched that of the first two stories, but which ended differently (in that the belief turned out to be false). The idea is that children will spontaneously align the first two stories, rendering the common true-belief structure salient. The subsequent false-belief story will initially align with this schema, leading to the candidate inference that the character's belief is true. When the belief turns out to be false, we expect this alignable difference to stand out as highly salient, thus calling attention to the difference between true- and false-belief events. Our goal was to highlight the idea that thoughts do not always match reality, which may make children curious as to when this occurs. Importantly, the same examples experienced in a different sequence—for example, in reverse—should not facilitate insight into belief schemas as effectively as the repetition-break order. This prediction was tested in Experiment 2b.

Experiment 2 also tested the claim that high overall similarity should strengthen comparison effects. According to Principle 3, this happens for two reasons: First, high overall similarity between two sequential events increases the likelihood that children will spontaneously compare them; and, second, high overall similarity facilitates structural alignment, especially in early learning (Gentner, Anggoro, & Klibanoff, 2011; Gentner, Loewenstein, & Hung, 2007; Gentner & Medina, 1998; Gentner & Toupin, 1986; Goldstone & Son, 2005; Hoyos & Gentner, 2017). Thus, we predicted that children presented with vignettes that are highly similar would be more likely to align the first two stories and extract their common relational structure, and therefore to notice the difference between true- and false-belief events.

Another motivation for Experiment 2 is that in Experiment 1, children in the Compare-Thoughts training received more exposure to mental states than did those in the other conditions. In Experiment 2, we equated exposure to mental-state depictions. Instead, what varied was the predicted likelihood that children would compare and align this information across instances. If we observe better performance on false-belief tasks when structural alignment across training instances is facilitated, this will provide evidence that comparison can support false-belief understanding.

In Experiment 2, we also asked whether the effectiveness of training could be related to children's command of mental-state language. As discussed earlier, positive effects of mental-state language have been found for performance on false-belief tasks (de Villiers & Pyers, 2002; Lohmann & Tomasello, 2003; Pyers & Senghas, 2009; Rabkina, Nakos, & Forbus, 2019). Therefore, we included an additional elicitation task in which we measured children's spontaneous production of mental-state verbs. If spontaneous use of mental-state verbs is related to children's propensity to encode situations in terms of mental states, then children who produced mental-state verbs in the elicitation task would benefit more from the training than those who did not.

A final question concerns transfer. In Experiment 1, the Compare-Thoughts group showed advantages only on the unexpected-contents task, showing transfer only to the task that resembled training. In Experiment 2, we asked whether the new training protocol would support transfer to the other false-belief tasks in the posttest (the unexpected-location and verbal false-belief tasks).

In sum, Experiment 2 tested two key predictions that arise from structure mapping: First, high overall similarity should invite and support comparison across examples, leading to better understanding of the underlying structures involved (Experiment 2a); second, experiencing examples in a repetition-break sequence should highlight the common structure in true-belief scenarios, and consequently make the key difference in the false-belief scenario more apparent (Experiment 2b). In addition, we aimed to clarify how features of our methods (explicit prompts to compare, differing exposure to thought bubbles) affect children's false-belief understanding. Finally, we extended our investigation to include mental-state language—another factor that has been found to be important in false-belief understanding.

3. Experiment 2a

In Experiment 2a, we again used a pretest-training–posttest design, within a single session. As in Experiment 1, we used the unexpected-contents task in our training. Our goal in the key condition was to lead children to notice the difference between false beliefs and true beliefs. To do this, we used a repetition-break sequence (Loewenstein & Heath, 2009) that began with two parallel (and readily alignable) true-belief scenarios, followed by a highly alignable false-belief scenario. Children saw a series of three stories. In each scenario, a character looked at a box—for example, a crayon box—and a thought bubble appeared with the character's belief about its contents (e.g., crayons). Then the contents of the box were revealed. In the first two stories, the character's belief was correct (true-belief events). In the third story, the character's belief was shown to be incorrect (false-belief event). The prediction is that structural alignment between the first two situations will render their common true-belief structure salient. When this common schema is applied to the final false-belief scenario, the critical difference between true and false beliefs will stand out.

To test the prediction that structural alignment is critical, we varied the similarity among the scenarios (see Fig. 4). In the *High-Alignability* (*HA*) condition, the three stories were similar in characters and objects. The *Low-Alignability* (*LA*) condition used the same sequence of two true-belief events followed by a false-belief event, but the characters and objects differed across the stories, making it harder for children to align across them. If comparison processing is critical to this learning, children in the HA condition will show more gains than those in the LA condition.

In addition to equating exposure to mental states, this simpler method was intended to reduce demands on attention. For each scenario, children attended to a single character and container, and there were fewer questions than in Experiment 1. Because the procedure was expected to be less demanding than that of Experiment 1, we extended the age range to the whole 4–5 period.

3.1. Methods

3.1.1. Participants

Eighty-one 4- to 5-year olds from the greater Evanston/Chicago area participated in this study (39 females, mean age = 54 months, range = 48–60 months). Eight additional children were tested but excluded from analyses for bringing a distracting toy into the testing area (n = 2), refusing to complete the study (n = 1), not answering all questions or not answering non-target questions appropriately during the study (n = 4), or experimenter error (n = 1). Another 50 children were excluded for ceiling performance in the pretest. The demographic make-up was similar to that of Experiment 1.

Children were assigned to the HA (n = 40, 20 females, mean age = 54 months, range = 48–60 months) or the LA (n = 41, 19 females, mean age = 53 months, range = 48–60 months) conditions. There were no differences across conditions in child

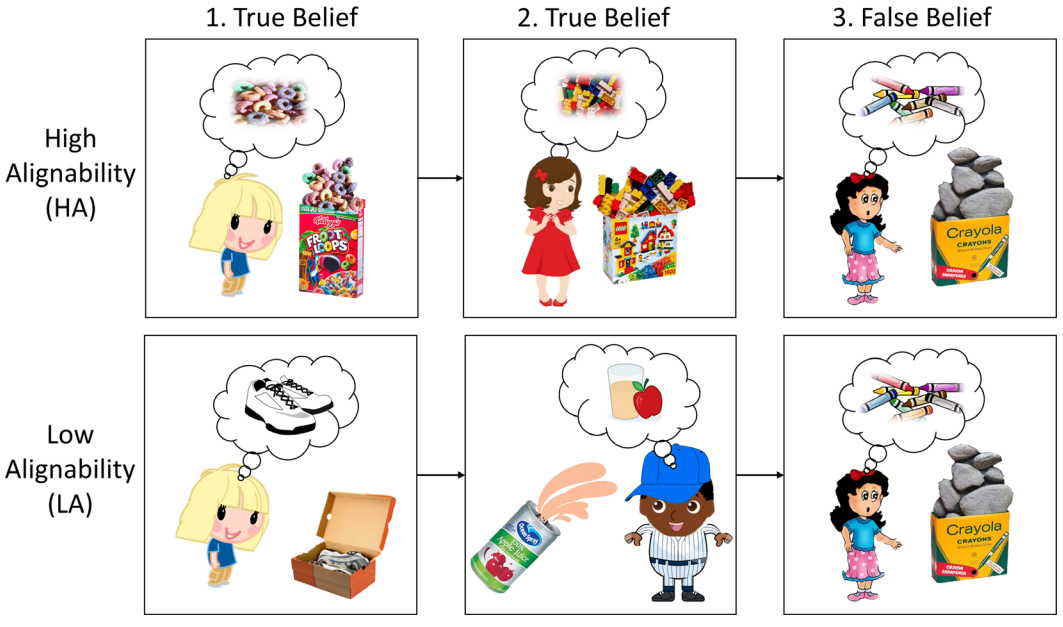

Fig. 4. Sample stills of stories shown to children during training in Experiment 2a. The first two stories were true-belief events followed by a final false-belief event. In the High Alignability condition, the stories used similar language and characters to invite and facilitate comparison across the stories. In the LA condition, the stories had differing language and characters to make comparison less likely.

age, $F(1, 79) = 2.30$, $p = .134$, $\eta^2 = .03$, or gender, $\chi^2(1) = 0.11$, $p = .742$, $V = 0.04$. Importantly, children also showed no differences across groups in false-belief performance on the pretest, $F(1, 79) = 2.06$, $p = .155$, $\eta^2 = .02$.

3.1.2. Design and procedure

After completing the diverse desires warm-up task, children completed a mental-state language elicitation task, adapted from a task used by Pyers and Senghas (2009). They were shown a cartoon story about two brothers engaged in deception and asked to describe what was occurring in each panel of the comic. At each panel, children were prompted with "What's happening here?" or "And then what happened?" but were not asked any specific questions about the stories. Their utterances were transcribed and coded as to whether children used mental-state verbs to describe the scenes (e.g., *want*, *think*, *know*).

Following the elicitation task, all children completed the false-belief pretest (consisting of three tasks: unexpected-contents; unexpected-location; verbal false-belief) and the thought-bubbles training procedure as in Experiment 1. Then children were randomly assigned to either the HA or the LA condition. In both conditions, children saw three stories presented sequentially: two true-belief stories followed by a false-belief story. Specifically, the first two events showed "expected contents" situations while the third showed the classic unexpected-contents situation. See Supplementary Material for full procedure.

In each of the three stories, children saw a box with obvious contents (such as cereal). Children were asked to predict what was inside the box, and the experimenter confirmed ("Yes, probably cereal"). This ensured children understood what the contents of the box should be. Then a character (Kim) who also had not yet seen inside the box appeared, with a thought bubble depicting her belief as to its contents (cereal). That is, children did not need to infer the character's thoughts because they were depicted in the scenario. The experimenter said, "Kim thinks that there's cereal in there. . .if you look there, you can see what she is thinking. So let's open the box and see what's inside" (see Fig. 4). Then the experimenter revealed the contents of the box, and the child was asked "Was she right?" The idea was to make sure children had noticed that the character's belief matched (or did not match) the true contents of the box. The correct answer was confirmed by the experimenter (e.g., "That's right—she thought there was cereal inside the box, and there was!"). Children (in this and the following studies) were nearly perfect in their answers to these questions. In the first two stories, the characters' predictions were right. In the third story (a false-belief story), the character's prediction was wrong. The idea was that if the child had successfully aligned the first two true-belief scenarios, then the contrast with the false-belief outcome in the third story should be highly salient (see Fig. 4 and Supplementary Material for full procedure).

We manipulated the alignability of the stories in four ways: (1) the characters and objects were highly similar in the HA condition and much less similar in the LA condition; (2) the same mental verb "think" was used in each story in the HA condition; in the LA condition, "think" was used in stories 1 and 3 and "believe" was used in story 2; (3) the spatial orientation of the box and the character were identical across the stories in the HA condition, but varied across stories in the LA condition; and (4) the location of the stories on the screen progressed in an easy-to-compare pattern (top left, center, bottom right) in the HA condition. This progression was disrupted in the LA condition (top left, middle right, and bottom center) to make alignment less likely.

Following the training scenarios, each child then completed three false-belief tasks for the posttest. Again, an unexpected-contents task examined near transfer, while the unexpected-location and verbal false-belief tasks allowed us to look for evidence for far transfer.

Unlike in Experiment 1, there was no explicit directive to compare across the stories. High similarity across the events should increase the likelihood that children spontaneously align across them. Thus, we predicted that the HA group would be more likely to align the first two stories and extract their common relational structure, therefore facilitating their ability to notice the difference between true and false beliefs.

3.2. Results

As in Experiment 1, children were given a score of 1 if they answered *both* the target and memory questions correctly; children with perfect pretest scores were excluded. Pretest and posttest means are displayed in Table 1. For the Elicitation task, we coded whether children produced one or more mental-state verbs (*want*, *think*, or *know*) in their

retelling of the cartoon. We used a categorical variable to indicate whether the child did or did not produce a mental-state verb.

3.2.1. Training effects

To assess whether significant improvement following training occurred in each condition, we compared gains from pretest to posttest to zero. Mean gains in the HA condition ($M = 0.75$, $SD = 0.84$) were significantly higher than zero, $t(39) = 5.65$, $p < .001$, $d = 0.89$, as were the gains in the LA condition ($M = 0.29$, $SD = 0.93$), $t(40) = 2.02$, $p = .050$, $d = 0.32$.

To examine effects and interactions of condition, an ANCOVA with pretest score as a covariate; posttest score (out of 3) as the dependent variable; and condition, gender, and mental-state language as between-subjects factors was conducted. Pretest scores significantly predicted posttest scores, $F(1, 72) = 26.84$, $p < .001$, $\eta^2 = .27$ (see Fig. 5 and Table 1 for pretest and posttest means).

The analysis revealed a main effect of condition, $F(1, 72) = 7.43$, $p = .008$, $\eta^2 = .09$, whereby children in the HA condition passed significantly more of the false-belief tasks on the posttest ($M^7 = 2.05$, $SE = 0.13$) than children in the LA condition ($M = 1.54$, $SE = 0.13$). In contrast to Experiment 1, no effects of gender emerged; females ($M = 1.88$, $SE = 0.13$) and males ($M = 1.71$, $SE = 0.13$) performed equally well on the posttest, $F(1, 72) = 0.78$, $p = .381$, $\eta^2 = .01$, and the interaction between gender and condition was also not significant, $F(1, 72) = 1.32$, $p = .255$, $\eta^2 = .02$.

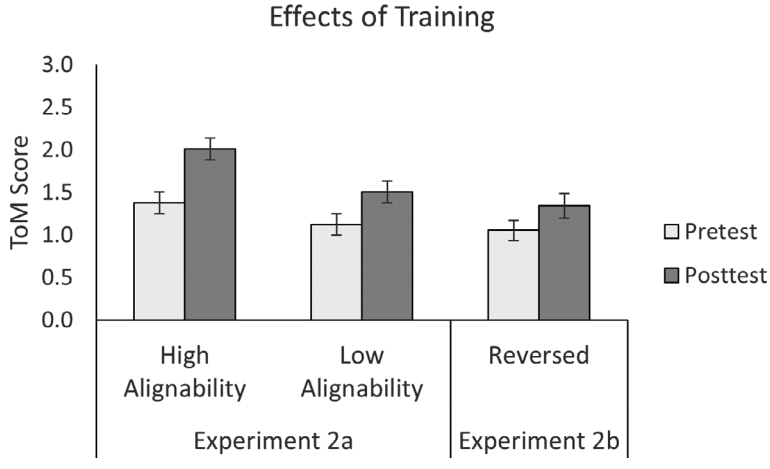

Fig. 5. Pretest and posttest performance across conditions in Experiments 2a and 2b. Pretest performance did not differ across conditions, but children in the High-Alignability condition scored significantly higher on the posttest than children in the Low-Alignability and Reversed conditions (who did not differ) after controlling for pretest scores. Both High-Alignability and Low-Alignability conditions (but not the Reversed condition) showed significant gains from pretest to posttest. Plotted posttest means are estimated marginal means; error bars depict ±1 *SEM*.

3.2.2. Mental-state language

Examination of mental-state language showed that 37 children (46%) produced at least one mental-state verb in the elicitation task, whereas the other 44 children (54%) did not. Most of these uses were of the verb *want*. The ANCOVA did not find a significant difference in false-belief understanding between children who produced mental-state language and those that did not, $F(1, 72) = 2.61$, $p = .11$, $\eta^2 = .04$, nor was there an interaction between condition and mental-state language, $F(1, 72) = 1.19$, $p = .278$, $\eta^2 = .02$.

Nonetheless, we ran planned comparisons on false-belief performance between children who had and had not produced mental-state language within each condition. We found a marginal effect in the LA condition: Children who produced mental-state language ($M = 1.79$, $SE = 0.19$) in the LA condition demonstrated marginally better false-belief understanding after training than children who did not ($M = 1.29$, $SE = 0.18$), $p = .059$, $d = 0.60$. However, there was no difference between children who had ($M = 2.10$, $SE = 0.20$) and had not ($M = 2.00$, $SE = 0.18$) produced mental-state language in the HA condition, $p = .712$, $d = 0.12$.

3.2.3. Near and far transfer

To examine the breadth of children's transfer, we analyzed the near- and far-transfer tasks separately. See Fig. 6 for a breakdown of false-belief performance across tasks.

3.2.3.1. Near transfer: To examine improvement on the near-transfer task (the unexpected-contents task), we conducted McNemar's exact tests. There was a significant increase in children passing the unexpected-contents task after training in both the HA group, $p = .006$, $V = .37$, and the LA group, $p = .021$, $V = .26$. To directly compare performance across conditions, we also conducted a multinomial logistic regression with

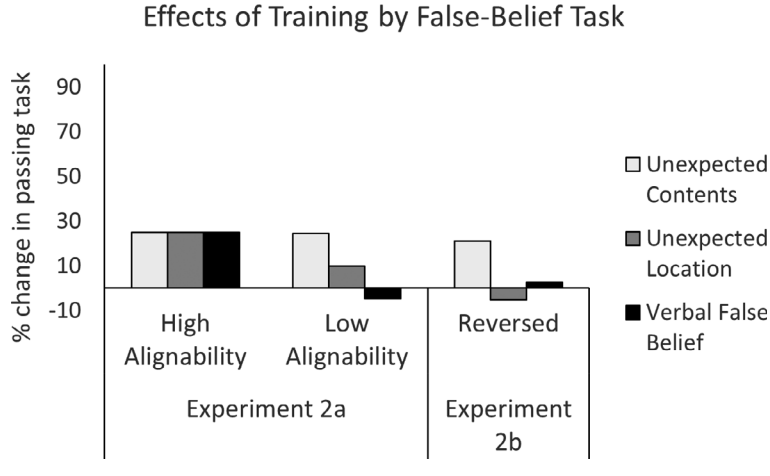

Fig. 6. Change in percentage of children passing each task from pretest to posttest for Experiments 2a and 2b.

pretest as a covariate and condition as a fixed factor. Mental-state language, sex, and their interactions with each other and with condition were also considered as factors; however, only mental-state language improved the model with pretest and condition, so the final model included only these three variables, $\chi^2(3) = 20.21$, $p < .001$. Condition did not predict improvements on the unexpected-contents task, $\beta = 0.76$, Wald's $\chi^2 = 1.37$, $p = .242$; however, mental-state language did, $\beta = -2.19$, Wald's $\chi^2 = 8.29$, $p = .004$. Children who did not produce any mental-state language were less likely than those who did to pass the unexpected-contents task. Thus, children in the HA and LA groups succeeded in near transfer equally well, but children who used mental-state language were more likely to succeed than those who did not.

3.2.3.2. Far transfer: To assess whether children in each condition improved on the far-transfer measures (the unexpected-location and verbal false-belief tasks), gain scores from pretest to posttest on these tasks were compared to zero. Mean gains in the HA were significantly greater than zero ($M = 0.50$, $SD = 0.72$), $t(39) = 4.42$, $p < .001$, $d = 0.70$; in contrast, the gains in the LA group were not ($M = 0.05$, $SD = 0.77$), $t(40) = 0.40$, $p = .688$, $d = 0.06$.

To examine condition differences, we ran an ANCOVA parallel to the one examining overall performance, considering only the far-transfer measures on the pretest and posttest. We found that children in the HA condition ($M = 1.23$, $SE = 0.11$) significantly outperformed those in the LA condition ($M = 0.74$, $SE = 0.11$), $F(1, 72) = 9.52$, $p = .003$, $\eta^2 = .12$. There were no effects or interactions with children's production of mental-state language on the far-transfer tasks. In sum, children in the HA condition were better able to transfer beyond the unexpected-contents task to pass the other two false-belief tasks than children in the LA condition.

3.3. Discussion

As predicted, children who received training with highly alignable examples demonstrated better false-belief understanding than children who received training with less alignable examples. It appears that children can benefit from a series of alignable events to gain insight into ToM. Importantly, there was no explicit invitation to children to compare the events (in contrast to Experiment 1). Thus, children's performance after training in this task resulted from their own spontaneous engagement in comparison processes.

It is also noteworthy that children in the HA condition transferred beyond the unexpected-contents task at posttest. That is, they showed better performance than those in the LA condition on the unexpected-location and verbal false-belief tasks. It might seem surprising that seeing *less* variable exemplars should lead to broader transfer, but this is predicted if children who experience more variable examples are less likely or less able to compare them. Only the HA group made significant gains on these far-transfer tasks, despite engaging in similar training procedures to the LA group, including seeing the same types of false-belief stories and answering the same questions.

We also found some indication that mental-state language—specifically, children's spontaneous production of mental-state verbs—can interact with analogical processing. Children who produced mental-state language fared better than those who did not in the LA condition, whereas there was no difference for children in the HA condition. These results are in line with prior work. Early in learning, before children have learned domain relations, they tend to focus on object matches (Gentner, 1988; Richland, Morrison, & Holyoak, 2006). Thus, they require high overall similarity to successfully align two situations (Gentner & Toupin, 1986; Loewenstein & Gentner, 2001). But with increasing relational knowledge, learners become able to align relationally similar situations even when the situations lack concrete similarity (Gentner, 2003, 2010; Gentner et al., 2007; Gentner & Rattermann, 1991; Kotovsky & Gentner, 1996). Against this background, we suggest that children who had more relational knowledge (as reflected in their spontaneous use of mental-state language) may have been more likely to encode the events in terms of mental relations, which in turn would have invited alignment even when surface-level commonalities were lacking. Children who did not produce such verbs may have been less likely to encode the events in terms of the relevant relations, and therefore more reliant on overall similarity to successfully align the events.

4. Experiment 2b

The results of Experiment 2a provide encouragement for the idea that children's own spontaneous structure-mapping processes can drive insight into ToM. In the preceding study, high similarity encouraged and supported children's alignment. However, Experiment 2a also provided what structure mapping would predict to be an ideal repetition-break sequence. Our processing account predicts that the repetition-break sequence is effective because (a) the first two events (whose relational structure is identical) are spontaneously aligned to create a true-belief schema in which the character's expectations concerning the contents of the container are fulfilled and (b) because that schema largely aligns with the final event, the alignable difference between them (that the contents are *not* what was expected) stands out sharply, highlighting the difference between true and false beliefs. If this account is correct, then experiencing the events in this order was important for children to gain false-belief insight.

To test this prediction, in Experiment 2b, we altered the order of the events in the HA condition to eliminate the repetition-break sequence and compared this to the LA and HA conditions in Experiment 2a. Specifically, we reversed the order of events such that children started training with the false-belief event, followed by the two true-belief events. In this order, the first event (false belief) and the second event (true belief) should not spontaneously align, so the alignable difference between the events should not pop out.

We predicted that this order would fail to make the contrast between true- and false-belief salient, and thus would not support success on subsequent false-belief tasks. Thus, if our processing account is correct, children should show lower performance in this reversed condition than in the comparable repetition-break condition. This outcome would

lend support to the claim that children's online analogical generalization and inference processes can support ToM insight.

4.1. Methods

4.1.1. Participants

Thirty-eight 4-year olds from the greater Evanston/Chicago area participated in this study (16 females, mean age = 53 months, range = 48–59 months). Six additional children were tested but were excluded from analysis for not finishing the study or not answering non-target questions appropriately ($n = 3$) and/or experimenter error ($n = 4$). Another 27 children participated in the study but were excluded from analysis for ceiling performance on the pretest. The demographic make-up was similar to that of Experiments 1 and 2a.

All children were assigned to the Reversed condition in Experiment 2b. When we compared this Reversed condition to the LA and HA conditions in Experiment 2a, there were no differences across conditions in child age, $F(2, 116) = 1.96$, $p = .146$, $\eta^2 = .03$; in gender, $\chi^2(2) = 0.49$, $p = .783$, $V = 0.06$; nor in false-belief performance on the pretest, $F(2, 116) = 1.89$, $p = .155$, $\eta^2 = .03$.

4.1.2. Design and procedure

The overall order of tasks was identical to that of Experiment 2a. Participants first completed the diverse desires warm-up task, followed by the mental-state language elicitation task, the false-belief pretest, and the thought-bubbles training. Then, in the training task, children were given the HA events from Experiment 2a, but in reversed order. Children first saw the false-belief event, followed by the two true-belief events. The procedure and questioning within each event were identical to those of Experiment 2a. Following this training, children completed the false-belief posttest.

4.2. Results

4.2.1. Training effects

To assess improvement after training, gains from pretest to posttest in the Reversed condition were compared to zero. Children in this condition, unlike the HA and LA conditions in Experiment 2a, did not show significant gains ($M = 0.18$, $SD = 0.80$), $t(37) = 1.42$, $p = .164$, $d = 0.23$.

To assess effects and interactions with condition, an ANCOVA comparing the performance in the Reversed condition to the HA and LA training conditions in Experiment 2a, with pretest score as a covariate; posttest score (out of 3) as the dependent variable; and condition, gender, and mental-state language as between-subjects factors was conducted. Pretest scores significantly predicted posttest scores, $F(1, 106) = 39.56$, $p < .001$, $\eta^2 = .27$ (see Fig. 5 and Table 1 for pretest and posttest means).

The analysis revealed a significant effect of condition, $F(2, 106) = 6.64$, $p = .002$, $\eta^2 = .11$. Planned comparisons showed that performance in the HA condition of Experiment 2a was significantly greater than in the Reversed condition of Experiment 2b ($M^3 = 1.34$, $SE = 0.14$), $p = .001$, $d = 0.60$. However, performance in the LA condition of Experiment 2a was not significantly different from the Reversed condition, $p = 0.409$, $d = 0.19$. These results are shown in Fig. 5.

As in the analysis in Experiment 2a, this analysis including the Reversed condition did not yield any effects of gender. Females ($M = 1.70$, $SE = 0.12$) performed no better than males ($M = 1.54$, $SE = 0.10$) on the posttest, $F(1, 106) = 1.12$, $p = .293$, $\eta^2 = .01$, and there was no significant interaction between gender and condition, $F(2, 106) = 0.69$, $p = .503$, $\eta^2 = .01$. No other effects or interactions were significant (Fig. 7).

Because the Reversed condition did not seem to benefit learning, we did not examine near- and far-transfer effects for this group.

4.2.2. Mental-state language

We also examined performance within the Reversed condition by mental-state language production. Within this condition, 15 children (39%) had produced at least one mental-state verb (*want*, *think*, or *know*) in the elicitation task, and 23 had not (61%). However, children who produced mental-state language ($M = 1.34$, $SE = 0.24$) performed no better than children who did not ($M = 1.34$, $SE = 0.17$) after Reversed training, $p = .980$, $d = 0.01$.

4.3. Discussion

As predicted, children in the Reversed condition demonstrated poorer false-belief understanding than the HA group in Experiment 2a, who had seen the same highly alignable

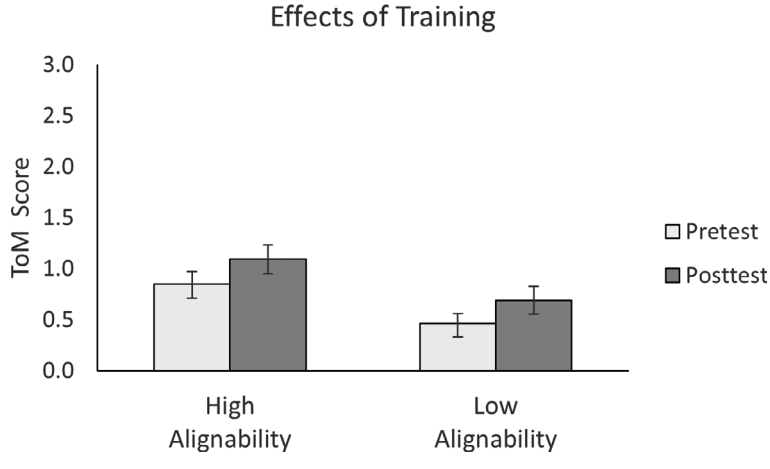

Fig. 7. Pretest and posttest performance across conditions in Experiment 3. Pretest performance differed significantly across conditions. However, children in the High-Alignability condition scored significantly higher on the posttest than children in the Low-Alignability condition after controlling for pretest scores. Only the High-Alignability condition showed significant gains from pretest to posttest. Plotted posttest means are estimated marginal means; error bars depict ±1 *SEM*.

examples, but in an ideal repetition-break order. Despite seeing these highly alignable events, the Reversed condition was also no better after training than the LA group, who had seen less alignable events. In fact, although both the HA and LA groups given the repetition-break sequence made significant gains from pretest to posttest, the Reversed condition did not. Even those who spontaneously produced mental-state language failed to show improvement. This shows that the order of events—and thus the kinds of online comparisons children engaged in—was a strong determinant of what they learned. The repetition-break order supported insights about false beliefs, but the reversed order did not.

How did the repetition-break sequence support learning across the events? Using the repetition-break sequence allowed children to extract a schema from the first two true-belief examples—namely, that the character's (and the child's) belief would be confirmed. The final example began in the same way as the others, inviting the child to apply the schema. Aligning the schema with the final situation led to the inference that the event would unfold in the same way—that the box's contents would match the character's (and the child's) expectations. Thus, the outcome, disconfirming the character's belief, stood out as a salient difference. This contrast made the difference between the true- and false-belief events pop out. In the Reversed condition, the false-belief event was shown first, without first making the true-belief structure salient. Even though the same three events were shown, children made no gains in this condition. This suggests that first creating a schema by comparing the two true-belief events was critical for highlighting the alignable difference between the true- and false-belief events.

5. Experiment 3

In the previous studies, we focused on 4-year olds, an age at which false-belief understanding is rapidly developing. In Experiment 3, we examined whether 3½-year olds could benefit from the repetition-break training we provided in Experiment 2a. We used the same procedures and conditions as in Experiment 2a. As before, our prediction was that children in the HA condition should see greater gains in false-belief reasoning than those in the LA condition (based on Principle 3). This would lend further strength to our claim that these learning processes are important for ToM development, as the benefits of comparison would extend across multiple age groups.

5.1. Methods

5.1.1. Participants

Seventy-nine 3½-year olds from the greater Evanston/Chicago area participated in this study (34 females, mean age = 44 months, range = 41–47 months). Six additional children were tested but were excluded from analysis for at least one of the following: refusing to answer questions (n = 4), not speaking English (n = 1), and experimenter error

(n = 2). Another 17 children participated but were excluded from analysis for ceiling performance in the pretest.

Children were assigned to the HA (n = 40, 17 females, mean age = 44 months, range = 41–47 months) or the LA conditions (n = 39, 17 females, mean age = 44 months, range = 41–47 months). There were no differences across conditions for age, $F(1, 77) = 0.07$, $p = .786$, $\eta^2 = .00$, or gender, $\chi^2(1) = 0.01$, $p = .922$, $V = 0.01$. Though children were randomly assigned to condition, we did find a significant difference across conditions in false-belief pretest scores, $F(1, 77) = 6.24$, $p = .015$, $\eta^2 = .08$. Children in the HA condition started out with a higher pretest mean than the LA condition (see Table 1).

5.1.2. Design and procedure

The procedure was identical to Experiment 2a. In a single session, children first completed the diverse desires warm-up task, followed by the mental-state language elicitation task, the false-belief pretest, and the thought-bubbles training. Children were then randomly assigned to the HA or LA conditions for training. Finally, they completed the false-belief posttest last.

5.2. Results

As in the previous experiments, children were given a score of 1 if they answered both the target and memory questions correctly; children with perfect pretest scores were excluded (see Table 1 for pretest and posttest means). For the elicitation task, we again measured whether children produced any mental-state verbs (*want*, *think*, or *know*).

5.2.1. Training effects

To assess improvement following training, gains from pretest to posttest were compared to zero. Mean gains in the HA condition ($M = 0.35$, $SD = 0.92$) were significantly greater than zero, $t(39) = 1.42$, $p = .021$, $d = 0.38$; however, gains in the LA condition ($M = 0.18$, $SD = 0.18$) were not, $t(38) = 1.36$, $p = .181$, $d = 0.22$.

To compare performance across conditions, an ANCOVA with pretest score as a covariate; posttest score (out of 3) as the dependent variable; and condition, gender, and mental-state language as between-subjects factors was conducted. Pretest scores significantly predicted posttest scores, $F(1, 70) = 11.32$, $p = .001$, $\eta^2 = .14$.

The analysis revealed a significant effect of condition, $F(1, 70) = 4.10$, $p = .047$, $\eta^2 = .06$. Children in the HA condition ($M^7 = 1.09$, $SE = 0.14$) showed significantly better performance after training than the LA condition ($M = 0.69$, $SE = 0.14$), after controlling for pretest performance.

Unlike Experiment 1, but like Experiment 2, gender was not related to children's false-belief understanding following training. Females ($M = 0.76$, $SE = 0.14$) and males ($M = 1.02$, $SE = 0.13$) performed equally well on the posttest, $F(1, 70) = 1.73$, $p = .193$, $\eta^2 = .02$, and gender did not interact with condition, $F(1, 70) = 0.43$, $p = .515$, $\eta^2 = .01$. No other effects or interactions were significant.

5.2.2. Mental-state language

Among this sample of 3½-year olds, 34 children (43%) produced at least one mental-state verb (*want*, *think*, *or know*), and 45 children (57%) did not. We compared posttest performance for children who did and did not produce mental-state language in each condition. As in Experiment 2a, performance in the HA group did not differ between children who did ($M = 1.20$, $SE = 0.20$) or did not ($M = 0.98$, $SE = 0.19$) produce mental-state language, $p = .416$, $d = 0.25$. However, in contrast to Experiment 2a, performance in the LA group also did not differ between children who did ($M = 0.66$, $SE = 0.20$) or did not ($M = 0.72$, $SE = 0.17$) produce mental-state language, $p = .842$, $d = 0.06$.

5.2.3. Near and far transfer

To examine the breadth of children's transfer, we analyzed the near- and far-transfer tasks separately.

5.2.3.1. Near transfer: To examine improvement on the near-transfer task, we conducted McNemar's exact tests. There was a significant increase in children passing the unexpected-contents task after training in the HA group, $p = .008$, $V = .42$, but not the LA group, $p = .109$, $V = .47$. To directly compare performance across conditions, we also conducted a multinomial logistic regression with pretest as a covariate and condition as a fixed factor. Mental-state language, sex, and their interactions with each other and with condition were also considered as factors; however, none of these improved the fit of the model and so were excluded. The final model with pretest and condition was significant, $\chi^2(2) = 18.73$, $p < .001$. However, condition was not related to the likelihood of passing the unexpected-contents task, $\beta = -0.782$, Wald's $\chi^2 = 2.32$, $p = .128$. Thus, the HA group was more likely to improve on the near-transfer task than the LA group, but comparing groups directly shows that the HA group was not more likely overall to pass the near-transfer task than the LA group.

5.2.3.2. Far transfer: To assess whether children in each condition improved on the far-transfer measures (the unexpected-location and verbal false-belief tasks), gain scores from pretest to posttest on these tasks were compared to zero. Mean gains in the HA group were not significantly greater than zero ($M = 0.05$, $SD = 0.71$), $t(39) = 0.44$, $p = .660$, $d = 0.07$; nor were the gains in the LA group ($M = 0.05$, $SD = 0.77$), $t(38) = 0.27$, $p = .786$, $d = 0.04$.

To examine condition differences in children's far transfer, we conducted an ANCOVA parallel to the one in the overall analysis, considering only the far-transfer tasks on the pretest and posttests. An effect of condition was marginally significant, $F(1, 70) = 3.78$, $p = .056$, $\eta^2 = .05$. For 3½-year olds, performance on the far-transfer posttest tasks was better following HA ($M = 0.58$, $SE = 0.10$) than LA training ($M = 0.29$, $SE = 0.10$).

5.3. Discussion

This study tested whether the ToM performance of 3½-year olds—who are not typically able to reliably reason about false belief—could benefit from analogical comparison. We found that the answer is yes: 3½-year olds can gain insight from experiencing the sequence of comparisons embodied in the repetition-break sequence. And like the 4-year olds in Experiment 2a, 3½-year olds demonstrated better false-belief understanding following training with highly alignable examples. However, this ability is more fragile for the younger children: Unlike the 4-year olds in Experiment 2a, who made gains in both conditions, the 3½-year olds showed significant gains only in the HA condition. This pattern is notable because the HA group started with higher pretest performance, and thus had less growth potential than the LA group. The 3½-year olds also failed to transfer beyond the unexpected-contents task, even in the HA condition, suggesting that their analogical generalization was fairly concrete and context-specific.

6. General discussion

In four experiments, we found evidence that analogical comparison can help children gain an understanding of the critical false-belief insight that mental states can differ from reality. In all four studies, we used a pretest-training–posttest sequence to assess the effects of analogical comparison. Experiment 1 was intended as a proof-of-concept to test whether children would benefit from intensive explicit comparisons between true- and false-belief scenarios. We showed 4-year-old children training scenarios with two juxtaposed cartoon characters—one with a true belief and one with a false belief—and prompted them to make comparisons between mental states and reality, between true and false beliefs, and between scenarios involving these distinctions. Controlling for pretest scores, children in this intensive comparison condition demonstrated significantly better false-belief understanding at posttest than children who were given other kinds of comparisons or no training. This finding is evidence that analogical comparison can support ToM insight, at least when children are given a concentrated set of explicit comparisons. This sets the stage for asking our central question—can children's own spontaneous comparison processes promote gains in ToM?

In the remaining studies (Experiments 2a, 2b, and 3), children were exposed to sequences of discrete scenarios, an approach that more closely resembles what could happen in real life. These studies allow us to test four principles derived from structure mapping: (1) Structural alignment renders common structure more salient, thus promoting schema abstraction; (2) comparison between highly alignable events promotes highlighting of alignable differences; (3) high overall similarity between situations both invites spontaneous comparison and facilitates structural alignment; and (4) the less the learner knows about the relational structure of a domain, the more dependent they are on overall similarity and other supports to perceive and align relational structure.

Principles (1) and (2) are combined in the repetition-break sequence (Loewenstein & Heath, 2009). In this sequence, two highly alignable events are juxtaposed, inviting comparison and schema abstraction. This is immediately followed by a final event that partly aligns with that schema—inviting alignment and inference-projection—but deviates from it in a critical way. This renders the key difference salient—a kind of pop-out effect. Our hypothesis was that this sharp contrast would support learning. Specifically, we hypothesized that the repetition-break sequence could be harnessed to show children the difference between true and false beliefs. A further prediction, stemming from Principles (3) and (4), was that these early learners would show more gains when given high-overall-similarity sequences (as in the HA conditions) than when given less surface-similar events (as in the LA conditions). This prediction was confirmed in Experiments 2a and 3.

Experiment 2a tested whether the repetition-break sequence (Principles 1 and 2) would be effective in conveying the true-belief–false-belief distinction. This study also tested the prediction that high overall similarity promotes spontaneous comparison and structural alignment (Principle 3). Thus, we varied the surface similarity among the three events in the repetition-break sequence (TB, TB, and FB). In the High-Alignability condition, the three events were overall similar—they shared surface similarities as well as relational similarity. In the Low-Alignability condition, the events shared relational structure but not surface similarity—that is, they had somewhat dissimilar objects, characters, and spatial configurations (see Fig. 4). The results bore out our predictions. Both groups showed significant gains from pretest to posttest (consistent with Principles 1 and 2), though the High-Alignability group performed better than the Low-Alignability group (consistent with Principle 3). Further, whereas the Low-Alignability group showed gains only on the trained task (the unexpected-contents task), the High-Alignability group also showed far-transfer to new false-belief tasks (unexpected location and verbal false-belief). These findings underline that false belief may not be an all-or-nothing insight; children may learn it in a highly context-specific way (as in our Low-Alignability group) and gradually generalize the schema as they encounter more examples. This incremental learning course is naturally predicted by the structure-mapping account[8] (Kuehne, Forbus, et al., 2000).

In Experiment 3, we ran the same design with a younger group (3½-year olds): Children were given the repetition-break sequence with either a High-Alignability set or a Low-Alignability set. Again, we found that children gained significantly in ToM performance after repetition-break training (consistent with Principles 1–3). But unlike the 4-year olds, the 3½-year olds showed gains only in the High-Alignability condition. That is, consistent with Principle 4, younger children required greater overall similarity to derive the analogical insight.

A further prediction of the structure-mapping account is that, to benefit from the repetition-break sequence, the learner must be able to align the first two items and form a common schema. We tested this prediction in Experiment 2b by reversing the repetition-break sequence used in the High-Alignability condition from Experiment 2a—resulting in a FB, TB, and TB sequence. We predicted that without an early pair of highly alignable events, spontaneous schema formation would not occur, and no inferences would be projected to the third event (Principle 1). As predicted, 4-year olds failed to make significant gains in

false-belief understanding when they received this reversed order. This finding underscores the importance of structural similarity in the alignment process; even though high surface similarity was maintained, we saw no evidence of schema-formation. Critically, this finding also highlights the role of alignable differences in supporting children's false-belief insight.

Alignable differences can be powerful sources for learning. Alignable differences have been shown to aid category learning in adults (Higgins & Ross, 2011) and word learning in children (Shao & Gentner, 2016; Waxman & Klibanoff, 2000). Alignable differences are faster to notice than non-alignable differences (Sagi et al., 2012), are more likely to be mentioned when people describe comparisons (Gentner & Gunn, 2001; Gentner & Markman, 1994), and are better remembered (Markman & Gentner, 1997). These patterns are predicted by structure mapping and are readily modeled by its computational model, SME (the Structure-mapping Engine; Falkenhainer et al., 1989; Forbus, Ferguson, Lovett, & Gentner, 2017). SME has also been used to simulate the rapid "pop out" of alignable differences in a visual comparison task (Sagi et al., 2012), using the same alignment process that it uses for analogical processing in general. To our knowledge, SME is the only simulation of analogical comparison that has been able to capture the phenomena of alignable differences. Alignable differences arise naturally from SME's mapping process, which involves an initially symmetric structural alignment between two representations (Falkenhainer et al., 1989; Forbus et al., 2017; Wolff & Gentner, 2011). Though many other tenets of structure mapping are widely shared among analogy researchers (Doumas, Hummel, & Sandhofer, 2008; Gentner & Holyoak, 1997; Hummel & Holyoak, 1997; see Kokinov & French, 2006), initial symmetric alignment is a unique assumption among analogical process models.

New inferences and alignable differences follow from this initial symmetric alignment; thus, it is possible to learn something that is not yet understood in either of the separate items (e.g., Christie & Gentner, 2010; Kotovsky & Gentner, 1986; Kurtz, Miao, & Gentner, 2001). This also means that the learner does not have to know the point of the alignment in advance. In contrast, many other models of analogical processing use a driver-recipient structure in which a designated relational structure in the base is projected to the target (e.g., Doumas et al., 2008; Hummel & Holyoak, 1997). In terms of false belief, this means that a learner would need to be able to identify the belief structure (e.g., a false-belief schema) in one scenario to map it onto a new scenario. SME's alignment-first process predicts that a learner may be able to align two scenarios even before the belief structures of either are fully identified, and this may result in highlighting the common relational structure. This initial schema will typically be quite concrete, but it can provide the seed for further generalization. Thus, the structure-mapping account provides a domain-general mechanism through which children can generate new hypotheses (Bach, 2014; Christie & Gentner, 2010; Gentner & Hoyos, 2017; Xu, 2016).

6.1. Language

Prior research and theory support the idea that acquiring language for domain relations contributes to children's ability to detect and process these relations (Bach, 2014; Baldwin & Saylor, 2005; Gentner, 2003, 2010; Gentner & Christie, 2010). In particular, there

is evidence that children's ToM learning is supported by acquiring mental-state verbs and sentence-complement syntax (de Villiers & Pyers, 2002; Hale & Tager-Flusberg, 2003; Lohmann & Tomasello, 2003; Moore et al., 1990; Pyers & Senghas, 2009). Therefore, in Experiments 2 and 3, we included an elicitation task in which we measured children's spontaneous production of mental-state verbs. We found that 4-year-old children who produced mental-state verbs in the elicitation task outperformed those who did not in the Low-Alignability condition, but not in the High-Alignability condition (Experiment 2a). This is consistent with the idea that children who entered the study with greater insight into mental life—as evidenced by their spontaneous use of mental verbs—were able to encode the events in terms of beliefs as well as physical events (Bach, 2014; Baldwin & Saylor, 2005; San Juan & Astington, 2012). Thus, they could spontaneously match the two true-belief events, even when the events lacked surface similarity. In contrast, children with less initial knowledge required the support of high surface similarity, and therefore succeeded only in the High-Alignability condition. Consistent with this reasoning, we did not see effects of mental-state language with 3½-year olds (Experiment 3), nor when the repetition-break sequence was reversed (Experiment 2b).

6.2. Relation to other accounts of ToM development

Our approach is partly consistent with Gopnik and Wellman's influential theory–theory of the development of ToM (Gopnik & Wellman, 1994, 2012; Wellman, 2014). The key assumptions of this account—that ToM understanding requires representing systems of interconnected concepts, and that this system is acquired through experience—have received considerable empirical support (e.g., Amsterlaw & Wellman, 2006; Clements et al., 2000; Lecce et al., 2014; Slaughter & Gopnik, 1996). On this approach, children fail false-belief tasks because their existing model of mental-state reasoning does not account for how false beliefs relate to behavior. As children encounter phenomena that cannot be explained by their current model, they develop auxiliary hypotheses, which may become more prominent with experience.

This account is highly appealing, and roughly compatible with our own view. But it lacks an account of how these auxiliary hypotheses are generated. As Bach (2014, p. 355) puts it "supporters of STT [the scientific theory-theory] must identify a general learning mechanism for hypothesis discovery that can supplement the processes of probabilistic causal modeling and/or auxiliary hypothesis assimilation. Without such a unified account of domain-general learning mechanisms, defenders of STT remain vulnerable to nativist critiques." The structure-mapping account could intersect with this approach by providing a mechanism through which children can generate new hypotheses (Bach, 2014; Christie & Gentner, 2010; Gentner & Hoyos, 2017; Rabkina, McFate, & Forbus, 2018; Xu, 2016), which can be evaluated against evidence, potentially leading to improved models, or even to conceptual change.

Another difference between these two approaches is that whereas the structure-mapping view naturally predicts incremental generalization from concrete schemas to more abstract schemas, the theory–theory view does not appear to have any mechanism for gradual

abstraction of hypotheses. As discussed earlier, structure mapping predicts that the initial schema formed through comparison is often fairly concrete, especially in children's learning. This initially concrete schema becomes more abstract if the learner aligns that schema with further instances (e.g., Kotovsky & Gentner, 1996; Kuehne, Forbus, et al., 2000). This process has been modeled by SAGE, which uses SME to build up generalizations (Forbus et al., 2017; Kuehne, Forbus, et al., 2000; McLure, Friedman, & Forbus, 2015). Evidence for initially concrete representations in ToM come from studies that have found that seemingly superficial changes in false-belief tasks can have significant effects on performance (e.g., whether the entity changing locations in an Unexpected-Location task is an object or an animate being; Rai & Mitchell, 2004; Symons, McLaughlin, Moore, & Morine, 1997). Consistent with this pattern, we found that children in the Low-Alignability condition showed gains only on the near-transfer task, whereas those in the High-Alignment condition showed gains on both near transfer and far transfer.

The theory–theory and our analogical learning account both foreground learning of the knowledge structures in their explanations of ToM development. Some other accounts, in contrast, suggest that abstract knowledge about beliefs and desire is present very early and is potentially innate (Leslie, Friedman, & German, 2004; Scott & Baillargeon, 2017). These accounts point to recent demonstrations of success in infants and toddlers on simplified false-belief tasks (for reviews, see Barone et al., 2019; Scott & Baillargeon, 2017) and suggest that failures on traditional false-belief measures in preschoolers are due to other task demands (e.g., Friedman & Leslie, 2005; Kovács et al., 2010; Setoh et al., 2016). Specifically, improvements in false-belief reasoning are thought to stem largely from increased inhibitory control or processing capacity, rather than changes in the underlying conceptual knowledge, which they suggest is continuous across development. However, interpretations of the infant and toddler findings currently lack consensus (e.g., Heyes, 2014; Low, Apperly, Butterfill, & Rakoczy, 2016; Scott & Baillargeon, 2017). A number of researchers have questioned the interpretation of these results (Heyes, 2014; Ruffman & Perner, 2005), including whether these spontaneous-response infant tasks form a coherent construct (Apperly, 2012; Poulin-Dubois & Yott, 2018; Powell, Hobbs, Bardis, Carey, & Saxe, 2018) comparable to that found for explicit-response tasks (Peterson et al., 2012; Wellman et al., 2001; Wellman & Liu, 2004). For instance, the "two-systems" account (Apperly & Butterfill, 2009) suggests that infants may have a separate, limited system for tracking beliefs and behaviors, but that a more flexible reasoning system is used by older children and adults. The two-system account, along with theory–theory and other accounts (e.g., Benson et al., 2013), maintains a role for conceptual knowledge changes in children's developing ToM, and thus are broadly compatible with our findings. However, it is unclear how continuity accounts would explain the current findings. It seems unlikely that analogical training would lead to a change in inhibitory control or processing capacity during a single session. In addition, it is not clear how such an account could predict the differences found here between HA and LA training, or between near and far transfer.

Whatever the starting point of infants, the fact remains that preschool age children improve dramatically in their ability to reason coherently about others' minds, and in their ability to marshal their understanding to guide their expectations and behaviors in

scenarios involving others. To do so, children must represent the relationships among reality, others' mental states, and others' behaviors, and use these structures to interpret and predict events in complex, realistic scenarios (Apperly, 2012). Analogical comparison provides a natural and powerful framework for understanding how children do so.

6.3. Relation to other training studies

Of course, analogical comparison is not the only mechanism that contributes to acquiring ToM. Prior training studies have demonstrated gains in children's ToM using a variety of methods (see meta-analysis by Hofmann et al., 2016). However, we note that these training studies have typically included multiple sessions over 2–4 weeks. In contrast, in the present studies, children given analogical comparison experience showed significant gains in a single session.

Explanation tasks (e.g., asking children to explain someone's searching behavior) have been used in some successful training studies (Amsterlaw & Wellman, 2006; Knoll & Charman, 2000). Explanation has been widely used to support learning in adults and children (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Legare & Lombrozo, 2014; Lombrozo, 2010, 2012; Siegler, 2002), and it seems likely that this process contributes to ToM learning. Interestingly, there is evidence that explanation tasks can recruit comparison (Chin-Parker & Bradner, 2017; Edwards, Williams, Gentner, & Lombrozo, 2019; Hoyos & Gentner, 2017; Thibodeau, Crow, & Flusberg, 2016), and there is evidence that comparison experience can give rise to explanation (Wang & Baillargeon, 2008). We speculate that analogical comparison may be effective in combination with explanation and other processes in promoting understanding of ToM. For example, in Slaughter and Gopnik's (1996) study, when children gave incorrect answers in an appearance-reality task (e.g., "I always thought it was soap"), they received corrective feedback with contrastive statements such as "No you didn't, you thought they were golf balls." This may have led children to compare the past and current mental states. Likewise, the instructional explanations provided to children in Appleton and Reddy (1996) and Clements et al. (2000) often included statements that explicitly contrasted two mental states.

6.4. Limitations and future directions

One open question concerns the variable effects of gender. In Experiment 1, we found that females demonstrated better ToM understanding than males after training, especially in the critical Compare-Thoughts condition. However, we did not replicate either the main effect of gender or any interactions with training conditions in Experiments 2a, 2b, or 3. Thus, we cannot draw conclusions about gender effects from our studies. More broadly, the evidence for gender effects in ToM development is mixed. When effects of gender have been examined, they have generally been weak and/or elusive, especially when compared to other potential moderators, such as age (e.g., Charman et al., 2002) or length of training (Hofmann et al., 2016).

A possible limitation of the current findings concerns our use of thought bubbles in the training sequence. We might ask whether children's learning was restricted to contexts

involving thought bubbles. This does not appear to be the case, as the pretest and posttest tasks did not contain thought bubbles. That children showed gains on these ToM tasks suggests that they had learned something general about mental states and their relation to the state of the world.

A larger concern here is whether and how the comparison mechanism we have described would occur in real life, where people's thoughts are not clearly displayed. Clearly, for structure mapping to serve as a viable candidate for driving ToM, it must be the case that children can spontaneously make the kinds of comparisons described here on their own. One possibility is that parents' use of mental-state language can invite comparison across events (Baldwin & Saylor, 2005; San Juan & Astington, 2012). Developmental studies have shown that common labels invite comparison (Christie & Gentner, 2010; Gentner & Namy, 1999; Namy & Gentner, 2002). It seems likely that mental-state language could serve as an impetus for what Gentner and Medina (1998) termed "symbolic juxtaposition" of mental states. Indeed, there is evidence that children whose parents produce more mental-state language go on to show earlier acquisition of ToM (Devine & Hughes, 2014).

There is also evidence suggesting that spontaneous comparison might support children's real-life ToM learning. Researchers working with corpora of children's speech have found that young children do sometimes produce contrastive statements in which two mental states are directly compared (Bartsch & Wellman, 1995; Wellman & Estes, 1987). For example, Bartsch and Wellman (1995) found that children as young as two could produce contrastive statements such as "I like it – but you don't like it." Shatz, Wellman, and Silber (1983) found that 20% of the mental-state utterances produced by the 3-year olds in their sample were in a contrastive form. Sabbagh and Callanan (1998) found that explicit contrastive statements increased in frequency between 3 and 5 years of age. This evidence suggests that children do engage in spontaneous mental-state comparisons. Based on this evidence, an intriguing question for future work is whether children who produce more contrastive statements early in development go on to show earlier false-belief understanding than children who do not produce these kinds of contrastive statements (see Silvey, Gentner, Richland, & Goldin-Meadow, 2017, for evidence that children's early use of specific comparisons predicts their later performance on analogy tasks).

A key open question arising from the present findings is whether children retain these gains over time. Future research should investigate children's retention of the insights they derive from comparing events. Research should also explore factors that may interact with analogical comparison to promote long-term retention and transfer. For example, would the use of relational language during the training session improve retention, as has been found in other studies of analogical learning (Gentner et al., 2011; Loewenstein & Gentner, 2005)? Another factor that might promote long-term retention is self-explanation (Chi et al., 1989; Legare & Lombrozo, 2014; Lombrozo, 2010, 2012; Siegler, 2002)—for example, asking children to compare what happened in a pair of contrasting true- and false-belief scenarios. Combining analogical comparison with explanation of the commonalities and differences could lead to better mental-state representations, which would grant

children greater cognitive control over their attention and behavior during ToM reasoning (e.g., Munakata, Snyder, & Chatham, 2014). Indeed, although our emphasis here has been on children's understanding of false belief, the same analogical comparison processes might support children's understanding of diverse desires, and perhaps other insights as well. This idea is consistent with research that emphasizes the role of structure-mapping mechanisms in social cognitive development (Christie, 2017; Gerson, 2014).

Another approach that suggests a link between structure mapping and ToM development comes from the comparative literature. Analogy researchers have pointed out that humans' capacity for relational thinking far exceeds that of non-human primates (Gentner, 2010; Gentner & Christie, 2008; Penn, Holyoak, & Povinelli, 2008). In the domain of social relations, there is evidence that humans' social cognitive abilities far outpace those of non-human primates (Tomasello, 2014). Tomasello suggests that one key difference is that humans—even children—readily represent and reverse collaborative roles, whereas apes such as chimpanzees do not (Carpenter, Tomasello, & Striano, 2005). Tomasello (2014) proposes that this capacity for relational thinking—in particular, the ability to represent role-based concepts—arose from a need for social coordination among humans. Thus, over phylogeny, our relational processing abilities may have developed hand-in-hand with our social cognitive skills. If so, then it is perhaps not surprising that structure mapping is instrumental in the acquisition of mental-state understanding. Exploring how analogical comparisons contribute to a larger trajectory of social development will be a fruitful area for future work.

7. Conclusion

Understanding how humans gain insight into mental states is of central interest in cognitive science. Prior studies have shown that children's understanding of others' mental states can be improved through training, but not *how*. Our findings provide evidence for the role of analogical comparison as a learning process in ToM development. In our studies, structural alignment among events helped children understand the critical distinction between true and false beliefs. We suggest that this domain-general learning mechanism is a key process in allowing children to acquire the insights that guide mental-state reasoning.

Acknowledgments

We thank the families and preschools who participated in the research; Ashley Isaia, Emma Bulzoni, and Katie Cha for help with data collection; Sue Hespos, Erin Anderson, and members of the Cognition and Language Lab for comments and discussion; and Ken Forbus, Irina Rabkina, and Constantine Nakos for discussions of computational mechanisms. This work was supported by ONR Grant N00014-16-1-2613 to Dedre Gentner and an NSF Graduate Research Fellowship DGE-0824162 to Christian Hoyos.

Notes

1. Such structures are often referred to as *schemas*; we will primarily use this more general term.
2. Indeed, other research has shown that thought-bubble training is beneficial for helping deaf children (Wellman & Peterson, 2013) and children with autism (Paynter & Peterson, 2013; Wellman et al., 2002) better understand the nature of thoughts and theory of mind.
3. Reported means from the ANCOVA are estimated marginal means.
4. Overall, females (M = 0.89, SE = 0.09) were more successful on the far transfer tasks than males (M = 0.38, SE = 0.09). In addition, Bonferroni-corrected comparisons revealed that females showed predicted condition differences, with significantly better performance in the Compare-Thoughts group (M = 1.19, SE = 0.16) than Baseline (M = 0.62, SE = 0.16), p = .043, d = 1.32, though Compare-Items (M = 0.88, SE = 0.17) did not differ significantly from Compare-Thoughts, p = .574, d = .53. For males, there were no significant differences between Compare-Thoughts (M = 0.29, SE = 0.15) and Compare-Items (M = 0.43, SE = 0.16), p = 1.00, d = .24, or Baseline (M = 0.43, SE = 0.16), p = 1.00, d = .24.
5. Gender analyses across studies may have also been underpowered. However, the main effect of gender in Experiment 2a was not (power = .95), and this effect was not significant.
6. The repetition-break pattern is also used in jokes and in advertisements, as discussed by Loewenstein and Heath (2009) and Rozin et al. (2006). An example from Mastercard's "priceless" campaign is as follows: "18-speed bike: $1,235. Shipping bike to Italy: $281. Map of Tuscany: 4,000 lira. Seven days without e-mail: priceless."
7. Reported means are estimated marginal means.
8. However, we note that in our studies, the near transfer task was also the first one following training; future work will need to tease apart the effects of similarity and temporal proximity to the training examples.

References

Amsterlaw, J., & Wellman, H. M. (2006). Theories of mind in transition: A microgenetic study of the development of false belief understanding. *Journal of Cognition and Development*, *7*(2), 139–172. https://doi.org/10.1207/s15327647jcd0702_1

Apperly, I. A. (2012). What is "theory of mind"? Concepts, cognitive processes and individual differences. *The Quarterly Journal of Experimental Psychology*, *65*(5), 825–839. https://doi.org/10.1080/17470218.2012.676055

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, *116*(4), 953. https://doi.org/10.1037/a0016923

Appleton, M., & Reddy, V. (1996). Teaching three-year-olds to pass false belief tests: A conversational approach. *Social Development*, *5*(3), 275–291. https://doi.org/10.1111/j.1467-9507.1996.tb00086.x

Astington, J. W. (2003). Sometimes necessary, never sufficient: False-belief understanding and social competence. In B. Repacholi & V. Slaughter (Eds.), *Individual differences in theory of mind: Implications for typical and atypical development* (pp. 13–38). New York: Psychology Press.

Bach, T. (2014). A unified account of general learning mechanisms and theory of mind development. *Mind and Language*, *29*(3), 351–381. https://doi.org/10.1111/mila.12055

Baldwin, D. A., & Saylor, M. (2005). Language promotes structural alignment in the acquisition of mentalistic concepts. In J. Astington & J. Baird (Eds.), *Why language matters for theory of mind* (pp. 123–143). New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195159912.003.0007

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*, 37–46. https://doi.org/10.1016/0010-0277(85)90022-8

Barone, P., Corradi, G., & Gomila, A. (2019). Infants' performance in spontaneous-response false belief tasks: A review and meta-analysis. *Infant Behavior and Development*, *57*. https://doi.org/10.1016/j.infbeh.2019.101350

Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind*. New York: Oxford University Press. https://doi.org/10.1002/1520-6807(199601)33:1<87:AID-PITS2310330105>3.0.CO;2-C

Benson, J. E., Sabbagh, M. A., Carlson, S. M., & Zelazo, P. D. (2013). Individual differences in executive functioning predict preschoolers' improvement from theory-of-mind training. *Developmental Psychology*, *49*(9), 1615–1627. https://doi.org/10.1037/a0031056

Carpenter, M., Tomasello, M., & Striano, T. (2005). Role reversal imitation and language in typically-developing infants and children with autism. *Infancy*, *8*, 253–278. https://doi.org/10.1207/s15327078in0803_4

Casasola, M. (2005). Can language do the driving? The effect of linguistic input on infants' categorization of support spatial relations. *Developmental Psychology*, *41*(1), 183. https://doi.org/10.1037/0012-1649.41.1.183

Charman, T., Ruffman, T., & Clements, W. (2002). Is there a gender difference in false belief development? *Social Development*, *11*(1), 1–10. https://doi.org/10.1111/1467-9507.00183

Chen, Z., & Mo, L. (2004). Schema induction in problem solving: A multidimensional analysis. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *30*, 583–600. https://doi.org/10.1037/0278-7393.30.3.583

Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*(2), 145–182. https://doi.org/10.1207/s15516709cog1302_1

Childers, J. B., & Paik, J. H. (2009). Korean-and English-speaking children use cross-situational information to learn novel predicate terms. *Journal of Child Language*, *36*(01), 201–224. https://doi.org/10.1017/S0305000908008891

Childers, J. B., Parrish, R., Olson, C. V., Burch, C., Fung, G., & McIntyre, K. P. (2016). Early verb learning: How do children learn how to compare events? *Journal of Cognition and Development*, *17*(1), 41–66. https://doi.org/10.1080/15248372.2015.1042580

Chin-Parker, S., & Bradner, A. (2017). A contrastive account of explanation generation. *Psychonomic Bulletin & Review*, *24*(5), 1387–1397. https://doi.org/10.3758/s13423-017-1349-x

Christie, S. (2017). Structure-mapping for social learning. *Topics in Cognitive Science*, *9*(3), 758–775. https://doi.org/10.1111/tops.12264

Christie, S., & Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. *Journal of Cognition and Development*, *11*(3), 356–373. https://doi.org/10.1080/15248371003700015

Clements, W. A., Rustin, C., & McCallum, S. (2000). Promoting the transition from implicit to explicit understanding: A training study of false belief. *Developmental Science*, *3*, 88–92. https://doi.org/10.1111/1467-7687.00102

Cutting, A. L., & Dunn, J. (1999). Theory of mind, emotion understanding, language, and family background: Individual differences and interrelations. *Child Development*, *70*, 853–865. https://doi.org/10.1111/1467-8624.00061

de Villiers, J. G. (2005). Can language acquisition give children a point of view? In J. Astington & J. Baird (Eds.), *Why language matters for theory of mind* (pp. 186–219). New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195159912.003.0010

de Villiers, J. G., & Pyers, J. E. (2002). Complements to cognition: A longitudinal study of the relationship between complex syntax and false-belief understanding. *Cognitive Development*, *17*, 1037–1060. https://doi.org/10.1016/S0885-2014(02)00073-4

Dennett, D. (1978). Beliefs about beliefs. *Behavioral and Brain Sciences*, *1*(4), 568–570. https://doi.org/10.1016/0010-0277(83)90004-5

Devine, R., & Hughes, C. (2014). Relations between false belief understanding and executive function in early childhood: A meta-analysis. *Child Development*, *85*(5), 1777–1794. https://doi.org/10.1111/cdev.12237

Ding, X. P., Wellman, H. M., Wang, Y., Fu, G., & Lee, K. (2015). Theory-of-mind training causes honest young children to lie. *Psychological Science*, *26*(11), 1812–1821. https://doi.org/10.1177/0956797615604628

Doumas, L. A. A., & Hummel, J. E. (2013). Comparison and mapping facilitate relation discovery and predication. *PLoS ONE*, *8*(6), e63889. https://doi.org/10.1371/journal.pone.0063889

Doumas, L. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*(1), 1. https://doi.org/10.1037/0033-295X.115.1.1

Edwards, B. J., Williams, J. J., Gentner, D., & Lombrozo, T. (2019). Explanation recruits comparison in a category-learning task. *Cognition*, *185*, 21–38. https://doi.org/10.1016/j.cognition.2018.12.011

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, *41*(1), 1–63. https://doi.org/10.1016/0004-3702(89)90077-5

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

Flavell, J. H., Green, F. L., Flavell, E. R., Watson, M. W., & Campione, J. C. (1986). Development of knowledge about the appearance-reality distinction. *Monographs of the Society for Research in Child Development*, *51*(1, Serial No. 212). https://doi.org/10.1016/0010-0285(83)90005-1

Forbus, K. D., Ferguson, R. W., Lovett, A., & Gentner, D. (2017). Extending SME to handle large-scale cognitive modeling. *Cognitive Science*, *41*, 1152–1201. https://doi.org/10.1111/cogs.12377

Friedman, O., & Leslie, A. M. (2005). Processing demands in belief-desire reasoning: Inhibition or general difficulty? *Developmental Science*, *8*(3), 218–225. https://doi.org/10.1111/j.1467-7687.2005.00410.x

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*, 155–170. https://doi.org/10.1207/s15516709cog0702_3

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

Gentner, D. (2003). Why we're so smart. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 195–235). Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/4117.003.0015

Gentner, D. (2005). The development of relational category knowledge. In L. Gershkoff-Stowe & D. H. Rakison (Eds.), *Building object categories in developmental time* (pp. 245–275). Hillsdale, NJ: Erlbaum. https://doi.org/10.2307/1130921

Gentner, D. (2010). Boostrapping the mind: Analogical processes and symbol systems. *Cognitive Science*, *34*(5), 752–775. https://doi.org/10.1111/j.1551-6709.2010.01114.x

Gentner, D. (2016). Language as cognitive tool kit: How language supports relational thought. *American Psychologist*, *71*(8), 650–657. https://doi.org/10.1037/amp0000082

Gentner, D., Anggoro, F. K., & Klibanoff, R. S. (2011). Structure-mapping and relational language support children's learning of relational categories. *Child Development*, *82*(4), 1173–1188. https://doi.org/10.1111/j.1467-8624.2011.01599.x

Gentner, D., & Christie, S. (2008). Relational language supports relational cognition in humans and apes. *Behavioral and Brain Sciences*, *31*, 136–137.

Gentner, D., & Christie, S. (2010). Mutual bootstrapping between language and analogical processing. *Language and Cognition*, *2*(2), 261–283. https://doi.org/10.1515/langcog.2010.011

Gentner, D., & Forbus, K. (2011). Computational models of analogy. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*, 266–276. https://doi.org/10.1002/wcs.105

Gentner, D., & Gunn, V. (2001). Structural alignment facilitates the noticing of differences. *Memory & Cognition*, *29*(4), 565–577. https://doi.org/10.3758/BF03200458

Gentner, D., & Holyoak, K. J. (1997). Reasoning and learning by analogy: Introduction. *American Psychologist*, *52*, 32–34. https://doi.org/10.1037/0003-066X.52.1

Gentner, D., & Hoyos, C. (2017). Analogy and abstraction. *Topics in Cognitive Science*, *9*(3), 672–693. https://doi.org/10.1111/tops.12278

Gentner, D., & Kurtz, K. (2005). Relational categories. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. W. Wolff (Eds.), *Categorization inside and outside the lab* (pp. 151–175). Washington, DC: APA. https://doi.org/10.1006/cogp

Gentner, D., & Kurtz, K. (2006). Relations, objects, and the composition of analogies. *Cognitive Science*, *30*, 609–642. https://doi.org/10.1207/s15516709cog0000_60

Gentner, D., Levine, S. C., Dhillon, S., Ping, R., Bradley, C., Isaia, A., & 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., Loewenstein, J., & Hung, B. (2007). Comparison facilitates children's learning of names for parts. *Journal of Cognition and Development*, *8*, 285–307. https://doi.org/10.1080/15248370701446434

Gentner, D., & Markman, A. (1994). Structural alignment in comparison: No difference without similarity. *Psychological Science*, *5*(3), 152–158. https://doi.org/10.1111/j.1467-9280.1994.tb00652.x

Gentner, D., & Markman, A. (1997). Structure mapping in analogy and similarity. *American Psychologist*, *52*, 45–56. https://doi.org/10.1037/0003-066X.52.1.45

Gentner, D., & Medina, J. (1998). Similarity and the development of rules. *Cognition*, *65*, 263–297. https://doi.org/10.1016/S0010-0277(98)00002-X

Gentner, D., & Namy, L. (1999). Comparison in the development of categories. *Cognitive Development*, *14*, 487–513. https://doi.org/10.1016/S0885-2014(99)00016-7

Gentner, D., & Namy, L. L. (2006). Analogical processes in language learning. *Current Directions in Psychological Science*, *15*(6), 297–301. https://doi.org/10.1111/j.1467-8721.2006.00456.x

Gentner, D., & Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman & J. P. Byrnes (Eds.), *Perspectives on thought and language: Interrelations in development* (pp. 225–277). London: Cambridge University Press. https://doi.org/10.1017/CBO9780511983689.008

Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science*, *10*, 277–300. https://doi.org/10.1016/S0364-0213(86)80019-2

Gerson, S. A. (2014). Sharing and comparing: How comparing shared goals broadens goal understanding in development. *Child Development Perspectives*, *8*(1), 24–29. https://doi.org/10.1111/cdep.12056

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*(1), 1–38. https://doi.org/10.1016/0010-0285(83)90002-6

Goldstone, R. L., Day, S., & Son, J. Y. (2010). Comparison. In *Towards a theory of thinking* (pp. 103–121). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-03129-8_7

Goldstone, R. L., & Son, J. Y. (2005). The transfer of scientific principles using concrete and idealized simulations. *The Journal of the Learning Sciences*, *14*(1), 69–110. https://doi.org/10.1207/s15327809jls1401_4

Gopnik, A., & Astington, J. W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development*, *59*(1), 26–37. https://doi.org/10.2307/1130386

Gopnik, A., & Wellman, H. (1994). The theory theory. In L. Hirschfeld & S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 257–293). New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511752902.011

Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory. *Psychological Bulletin*, *138*(6), 1085–1108. https://doi.org/10.1016/j.cognition.2009.12.001

Hale, C. M., & Tager-Flusberg, H. (2003). The influence of language on theory of mind: A training study. *Developmental Science*, *6*(3), 346–359. https://doi.org/10.1111/1467-7687.00289

Haryu, E., Imai, M., & Okada, H. (2011). Object similarity bootstraps young children to action-based verb extensions. *Child Development*, *82*(2), 674–686. https://doi.org/10.1111/j.1467-8624.2010.01567.x

Heyes, C. (2014). False belief in infancy: A fresh look. *Developmental Science*, *17*(5), 647–659. https://doi.org/10.1111/desc.12148

Higgins, E. J., & Ross, B. H. (2011). Comparisons in category learning: How best to compare for what. In L. Carlson, C. Holscher, & T. Shipley (Eds.), *Proceedings of the 33rd annual conference of the Cognitive Science Society* (pp. 1388–1393). Austin, TX: Cognitive Science Society.

Hofmann, S. G., Doan, S. N., Sprung, M., Wilson, A., Ebesutani, C., Andrews, L. A., … Harris, P. L. (2016). Training children's theory-of-mind: A meta-analysis of controlled studies. *Cognition*, *150*, 200–212. https://doi.org/10.1016/j.cognition.2016.01.006

Holyoak, K. J., & Thagard, P. (1997). The analogical mind. *American Psychologist*, *52*, 35–44. https://doi.org/10.1037/0003-066X.52.1.35

Hoyos, C., & Gentner, D. (2017). Generating explanations via analogical comparison. *Psychonomic Bulletin & Review*, *24*(5), 1364–1374. https://doi.org/10.3758/s13423-017-1289-5

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*, 427–466. https://doi.org/10.3758/BF03197236

Knoll, M., & Charman, T. (2000). Teaching false belief and visual perspective taking skills in young children: Can a theory of mind be trained? *Child Study Journal*, *30*(4), 273–304.

Kokinov, B., & French, R. (2006). Analogy-making, computational models of. In L. Nadel (Ed.), *Encyclopedia of Cognitive Science* (pp. 113–118). London: Nature. https://doi.org/10.1002/0470018860.s00098

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, *67*, 2797–2822. https://doi.org/10.1111/j.1467-8624.1996.tb01889.x

Kovács, A. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, *330*, 1830–1834. https://doi.org/10.1126/science.1190792

Kuehne, S. E., Forbus, K. D., Gentner, D., & Quinn, B. (2000). SEQL--Category learning as progressive abstraction using structure mapping. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 770–775). Mahwah, NJ: Lawrence Erlbaum.

Kuehne, S. E., Gentner, D., & Forbus, K. D. (2000). Modeling infant learning via symbolic structural alignment. In L. Gleitman & A. K. Joshi (Eds.), *Proceedings of the 22nd annual conference of the Cognitive Science Society* (pp. 286–291). Mawah, N. J.: Lawrence Erlbaum.

Kurtz, K. J., Miao, C., & Gentner, D. (2001). Learning by analogical bootstrapping. *Journal of the Learning Sciences*, *10*, 417–446. https://doi.org/10.1207/S15327809JLS1004new_2

Lecce, S., Bianco, F., Demicheli, P., & Cavallini, E. (2014). Training preschoolers on first-order false belief understanding: Transfer on advanced ToM skills and metamemory. *Child Development*, *85*(6), 2404–2418. https://doi.org/10.1111/cdev.12267

Legare, C. H., & Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. *Journal of Experimental Child Psychology*, *126*, 198–212. https://doi.org/10.1016/j.jecp.2014.03.001

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in 'theory of mind'. *Trends in Cognitive Sciences*, *8*(12), 528–533. https://doi.org/10.1016/j.tics.2004.10.001

Lewis, S., Hacquard, V., & Lidz, J. (2012). The semantics and pragmatics of belief reports in preschoolers. In A. Chereches (Ed.), *Proceedings of the 22nd semantics and linguistic theory conference* (pp. 247–267). Chicago, IL: SALT. https://doi.org/10.3765/salt.v22i0.3085

Liu, D., Sabbagh, M. A., Gehring, W. J., & Wellman, H. M. (2009). Neural correlates of children's theory of mind development. *Child Development*, *80*, 318–326. https://doi.org/10.1111/j.1467-8624.2009.01262.x

Loewenstein, J., & Gentner, D. (2001). Spatial mapping in preschoolers: Close comparisons facilitate far mappings. *Journal of Cognition and Development*, *2*(2), 189–219. https://doi.org/10.1207/S15327647JCD0202_4

Loewenstein, J., & Gentner, D. (2005). Relational language and the development of relational mapping. *Cognitive Psychology*, *50*, 315–353. https://doi.org/10.1016/j.cogpsych.2004.09.004

Loewenstein, J., & Heath, C. (2009). The repetition-break plot structure: A cognitive influence on selection in the marketplace of ideas. *Cognitive Science*, *33*, 1–19. https://doi.org/10.1111/j.1551-6709.2008.01001.x

Loewenstein, J., Raghunathan, R., & Heath, C. (2011). The repetition-break plot structure makes effective television advertisements. *Journal of Marketing*, *75*, 105–119. https://doi.org/10.1509/jmkg.75.5.105

Loewenstein, J., Thompson, L., & Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin & Review*, *6*, 586–597. https://doi.org/10.3758/BF03212967

Lohmann, H., & Tomasello, M. (2003). The role of language in the development of false belief understanding: A training study. *Child Development*, *74*(4), 1130–1144. https://doi.org/10.1111/1467-8624.00597

Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–332. https://doi.org/10.1016/j.cogpsych.2010.05.002

Lombrozo, T. (2012). Explanation and abductive inference. *Oxford Handbook of Thinking and Reasoning*, 260–276. https://doi.org/10.1093/oxfordhb/9780199734689.013.0014

Low, J., Apperly, I. A., Butterfill, S. A., & Rakoczy, H. (2016). Cognitive architecture of belief reasoning in children and adults: A primer on the two-systems account. *Child Development Perspectives*, *10*(3), 184–189. https://doi.org/10.1111/cdep.12183

Markman, A. B., & Gentner, D. (1993a). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*, 431–467. https://doi.org/10.1006/cogp.1993.1011

Markman, A. B., & Gentner, D. (1993b). 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., & Gentner, D. (1996). Commonalities and differences in similarity comparisons. *Memory & Cognition*, *24*(2), 235–249. https://doi.org/10.3758/BF03200884

Markman, A. B., & Gentner, D. (1997). The effects of alignability on memory. *Psychological Science*, *8*(5), 363–367.

Markman, A. B., & Stilwell, C. H. (2001). Role-governed categories. *Journal of Experimental & Theoretical Artificial Intelligence*, *13*(4), 329–358. https://doi.org/10.1080/09528130110100252

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. Cambridge, MA: MIT Press. https://doi.org/10.1017/S030500090001134X

McLure, M. D., Friedman, S. E., , & (2015). Extending analogical generalization with near-misses. In B. Bonet & S. Koenig (Eds.), *Proceedings of the twenty-ninth AAAI conference on artificial intelligence* (pp. 565–571). Austin, TX: AAAI.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*(2), 254–278. https://doi.org/10.3758/BF03197629

Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development*, *78*(2), 622–646. https://doi.org/10.1111/j.1467-8624.2007.01018.x

Moore, C., Pure, K., & Furrow, D. (1990). Children's understanding of the modal expression of speaker certainty and uncertainty and its relation to the development of a representational theory of mind. *Child Development*, *61*(3), 722–730. https://doi.org/10.1111/j.1467-8624.1990.tb02815.x

Munakata, Y., Snyder, H. R., & Chatham, C. H. (2014). Developing cognitive control: The costs and benefits of active, abstract representations. In P. D. Zelazo & M. D. Sera (Eds.), *Minnesota symposia on child psychology: Developing cognitive control processes: Mechanisms, implications, and interventions* (Vol. 37, pp. 55–90). Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/9781118732373.ch3

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 Psychology: General*, *131*(1), 5. https://doi.org/10.1037/0096-3445.131.1.5

Paynter, J., & Peterson, C. C. (2013). Further evidence of benefits of thought-bubble training for theory of mind development in children with autism spectrum disorders. *Research in Autism Spectrum Disorders*, *7*, 344–348. https://doi.org/10.1016/j.rasd.2012.10.001

Pellicano, E. (2007). Links between theory of mind and executive function in young children with autism: Clues to developmental primacy. *Developmental Psychology*, *43*, 974–990. https://doi.org/10.1037/0012-1649.43.4.974

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Brain and Behavioral Sciences*, *31*, 109–178.

Perner, J. (1991). *Understanding the representational mind*. Cambridge, MA: The MIT Press. https://doi.org/10.7551/mitpress/6988.001.0001

Perner, J., & Lang, B. (1999). Development of theory of mind and executive control. *Trends in Cognitive Sciences*, *3*(9), 337–344. https://doi.org/10.1016/S1364-6613(99)01362-5

Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology*, *5*(2), 125–137. https://doi.org/10.1111/j.2044-835X.1987.tb01048.x

Peterson, C. C., Wellman, H. M., & Slaughter, V. (2012). The mind behind the message: Advancing theory-of-mind scales for typically developing children, and those with deafness, autism, or Asperger syndrome. *Child Development*, *83*(2), 469–485. https://doi.org/10.1111/j.1467-8624.2011.01728.x

Pham, K., Bonawitz, E., & Gopnik, A. (2012). Seeing who sees: Contrastive access helps children reason about other minds. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th annual conference of the Cognitive Science Society* (pp. 2180–2185). Austin, TX: Cognitive Science Society.

Poulin-Dubois, D., & Yott, J. (2018). Probing the depth of infants' theory of mind: Disunity in performance across paradigms. *Developmental Science*, *21*(4), e12600. https://doi.org/10.1111/desc.12600

Powell, L. J., Hobbs, K., Bardis, A., Carey, S., & Saxe, R. (2018). Replications of implicit theory of mind tasks with varying representational demands. *Cognitive Development*, *46*, 40–50. https://doi.org/10.1016/j.cogdev.2017.10.004

Pyers, J. E., & Senghas, A. (2009). Language promotes false belief understanding: Evidence from learners of a new sign language. *Psychological Science*, *20*(7), 805–812. https://doi.org/10.1111/j.1467-9280.2009.02377.x

Rabkina, I., McFate, C. J., & Forbus, K. D. (2018). Bootstrapping from language in the Analogical Theory of Mind model. In C. Kalish, M. Rau, T. Rogers, & J. Zhu (Eds.), *Proceedings of the 40th annual meeting of the Cognitive Science Society* (pp. 924–929). Austin, TX: Cognitive Science Society.

Rabkina, I., Nakos, C., & Forbus, K. D. (2019). Children's sentential complement use leads the Theory of Mind development period: Evidence from the CHILDES corpus. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st annual conference of the Cognitive Science Society* (pp. 2434–2639). Montreal, QC: Cognitive Science Society.

Rai, R., & Mitchell, P. (2004). Five-year-old children's difficulty with false belief when the sought entity is a person. *Journal of Experimental Child Psychology*, *89*(2), 112–126. https://doi.org/10.1016/j.jecp.2004.05.003

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology*, *33*, 12–21. https://doi.org/10.1037/0012-1649.33

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

Rozin, P., Rozin, A., Appel, B., & Wachtel, C. (2006). Documenting and explaining the common AAB pattern in music and humor: Establishing and breaking expectations. *Emotion*, *6*, 349–355. https://doi.org/10.1037/1528-3542.6.3.349

Ruffman, T., & Perner, J. (2005). Do infants really understand false belief?: Response to Leslie. *Trends in Cognitive Sciences*, *9*(10), 462–463.

Ruffman, T., Perner, J., & Parkin, L. (1999). How parenting style affects false belief understanding. *Social Development*, *8*(3), 395–411. https://doi.org/10.1111/1467-9507.00103

Sabbagh, M. A., & Callanan, M. A. (1998). Metarepresentation in action: Children's theories of mind developing and emerging in parent-child conversations. *Developmental Psychology*, *34*, 491–502. https://doi.org/10.1037//0012-1649.34.3.491

Sagi, E., Gentner, D., & Lovett, A. (2012). What difference reveals about similarity. *Cognitive Science*, *36* (6), 1019–1050. https://doi.org/10.1111/j.1551-6709.2012.01250.x

San Juan, V., & Astington, J. W. (2012). Bridging the gap between implicit and explicit understanding: How language development promotes the processing and representation of false belief. *British Journal of Developmental Psychology*, *30*, 105–122. https://doi.org/10.1111/j.2044-835X.2011.02051.x

Scott, R. M., & Baillargeon, R. (2017). Early false-belief understanding. *Trends in Cognitive Sciences*, *21*(4), 237–249. https://doi.org/10.1016/j.tics.2017.01.012

Setoh, P., Scott, R. M., & Baillargeon, R. (2016). Two-and-a-half-year-olds succeed at a traditional false-belief task with reduced processing demands. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(47), 13360–13365. https://doi.org/10.1073/pnas.1609203113

Shao, R., & Gentner, D. (2016). Structural alignment in incidental word learning. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th annual conference of the Cognitive Science Society* (pp. 1038–1043). Austin, TX: Cognitive Science Society.

Shatz, M., Wellman, H. M., & Silber, S. (1983). The acquisition of mental verbs: A systematic investigation of the first reference to mental state. *Cognition*, *14*, 301–321. https://doi.org/10.1016/0010-0277(83)90008-2

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*(2), 125–140. https://doi.org/10.1007/BF02289630

Siegal, M., & Beattie, K. (1991). Where to look first for children's knowledge of false beliefs. *Cognition*, *38*, 1–12. https://doi.org/10.1016/0010-0277(91)90020-5

Siegler, R. S. (2002). Variability and infant development. *Infant Behavior and Development*, *25*(4), 550–557. https://doi.org/10.1016/S0163-6383(02)00150

Silvey, C., Gentner, D., Richland, L., & Goldin-Meadow, S. (2017). Children's specific comparisons from 26 to 58 months predict performance in verbal and non-verbal analogy tests in 6th grade. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th annual meeting of the Cognitive Science Society* (pp. 1072–1077). Austin, TX: Cognitive Science Society.

Slaughter, V., & Gopnik, A. (1996). Conceptual coherence in the child's theory of mind: Training children to understand belief. *Child Development*, *67*(6), 2967–2988. https://doi.org/10.1111/j.1467-8624.1996.tb01898.x

Slaughter, V., Imuta, K., Peterson, C. C., & Henry, J. D. (2015). Meta-analysis of theory of mind and peer popularity in the preschool and early school years. *Child Development*, *86*(4), 1159–1174. https://doi.org/10.1111/cdev.12372

Symons, D., McLaughlin, E., Moore, C., & Morine, S. (1997). Integrating relationship constructs and emotional experience into false belief tasks in preschool children. *Journal of Experimental Child Psychology*, *67*(3), 423–447. https://doi.org/10.1006/jecp.1997.2416

Thibodeau, P. H., Crow, L., & Flusberg, S. J. (2016). The metaphor police: A case study of the role of metaphor in explanation. *Psychonomic Bulletin & Review*, *24*(5), 1375–1386. https://doi.org/10.3758/s13423-016-1192-5

Todd, A. R., Hanko, K., Galinsky, A. D., & Mussweiler, T. (2011). When focusing on differences leads to similar perspectives. *Psychological Science*, *22*(1), 134–141. https://doi.org/10.1177/0956797610392929

Tomasello, M. (2014). *A natural history of human thinking*. Cambridge, MA: Harvard University Press. https://doi.org/10.1515/jso-2015-0041

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327. https://doi.org/10.1037/0033-295X.84.4

Walker, S. (2005). Gender differences in the relationship between young children's peer-related social competence and individual differences in theory of mind. *The Journal of Genetic Psychology*, *166*, 297–312. https://doi.org/10.3200/GNTP.166.3.297-312

Wang, S. H., & Baillargeon, R. (2008). Can infants be "taught" to attend to a new physical variable in an event category? The case of height in covering events. *Cognitive Psychology*, *56*(4), 284–326. https://doi.org/10.1016/j.cogpsych.2007.06.003

Waxman, S. R., & Klibanoff, R. S. (2000). The role of comparison in the extension of novel adjectives. *Developmental Psychology*, *36*, 571–581. https://doi.org/10.1037/0012-1649.36.5.571

Wellman, H. M. (1990). *The child's theory of mind*. Cambridge, MA: MIT Press. https://doi.org/10.1016/0010-0277(92)90004-2

Wellman, H. M. (2014). *Making minds: How theory of mind develops*, Oxford, U.K.: Oxford University Press.

Wellman, H. M., Baron-Cohen, S., Caswell, R., Gomez, J. C., Swettenham, J., Toye, E., & Lagattuta, K. (2002). Thought-bubbles help children with autism acquire and alternative to theory of mind. *Autism*, *6*, 343–363. https://doi.org/10.1080/15248372.2016.1205337

Wellman, H. M., & Bartsch, K. (1989). 3-year-olds understand belief. *Cognition*, *33*, 321–326. https://doi.org/10.1016/0010-0277(89)90033-4

Wellman, H. M., Cross, D., & Watson, J. (2001). A meta-analysis of theory of mind development: The truth about false belief. *Child Development*, *72*, 655–684. https://doi.org/10.1111/1467-8624.00304

Wellman, H. M., & Estes, D. (1987). Children's early use of mental terms and what they mean. *Discourse Processes*, *10*, 141–156. https://doi.org/10.1080/01638538709544666

Wellman, H. M., Hollander, M., & Schult, C. A. (1996). Young children's understanding of thought-bubbles and of thoughts. *Child Development*, *67*, 768–788. https://doi.org/10.1111/j.1467-8624.1996.tb01763.x

Wellman, H. M., & Liu, D. (2004). Scaling of theory of mind tasks. *Child Development*, *75*, 523–541. https://doi.org/10.1111/j.1467-8624.2004.00691.x

Wellman, H. M., & Peterson, C. C. (2013). Deafness, thought bubbles, and theory-of-mind development. *Developmental Psychology*, *49*, 2357–2367. https://doi.org/10.1037/a0032419

Wellman, H. M., & Woolley, J. D. (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition*, *35*, 245–275. https://doi.org/10.1016/0010-0277(90)90024-E

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*, 103–128. https://doi.org/10.1016/0010-0277(83)90004-5

Wolff, P., & Gentner, D. (2011). Structure-mapping in metaphor comprehension. *Cognitive Science*, *35*, 1456–1488. https://doi.org/10.1111/j.1551-6709.2011.01194.x

Wu, Y., Haque, J. A., & Schulz, L. E. (2018). Children can use others' emotional expressions to infer their knowledge and predict their behaviors in classic false belief tasks. In C. Kalish M. Rau T. Rogers & J. Zhu (Eds.), *Proceedings of the 40th annual meeting of the Cognitive Science Society* (pp. 1193–1198). Austin, TX: Cognitive Science Society.

Xu, F. (2016). Preliminary thoughts on a rational constructivist approach to cognitive development: Primitives, symbols, learning, and thinking. In D. Barner & A. S. Baron (Eds.), *Core knowledge and concept change* (pp. 11–28). Oxford, UK: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190467630.003.0002

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Appendix S1. Materials and Tasks.