Not a Pipe

STAND-FOR Relations in Human Cognition

by Barbu Revencu

Primary Supervisor: Gergely Csibra
Secondary Supervisor: Dan Sperber

Submitted to Central European University, Department of Cognitive Science
In partial fulfillment of the requirements for the degree of Doctor in Philosophy in Cognitive Science

Vienna, Austria
2023

Declaration of Authorship

I declare that this submission is my own work. It contains no previously published materials written by another person or accepted for the award of any other degree or diploma at Central European University or any other educational institution, except where due acknowledgment is made in the form of bibliographical reference.

The dissertation includes work that appears in the following publications:

Revencu, B., & Csibra, G. (2020). For 19-month-olds, what happens on the screen stays on the screen. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*. [link]

Revencu, B., & Csibra, G. (2021). For 19-Month-Olds, what happens on-screen stays on-screen. *Open Mind: Discoveries in Cognitive Science*, *5*, 71–90. [link]

Brody, G.*, Revencu, B.*, & Csibra, G. (2022). Images of objects are interpreted as symbols: A case study of automatic size measurement. *Journal of Experimental Psychology: General*, *152*(4), 1146–1157. [link]

Revencu, B., & Csibra, G. (preprint). Opening the black box of early depiction interpretation: From *whether* to *how* in the Theory-of-Puppets debate. *PsyArXiv*. [link]

*joint first authorship

Other parts of the dissertation will be submitted for publication with the following co-authors:

Chapter 1 with Gergely Csibra

Chapter 3 with Barbara Pomiechowska, Gabor Brody, and Gergely Csibra

Barbu Revencu

Acknowledgments

Gergő Csibra—for trusting me with this project, for his perpetual availability, for his breadth of knowledge and scientific integrity, for the contagious pleasure he takes in theoretical speculation, for never pulling rank, and for listening to me even when I am dead wrong

Dan Sperber—for vigorously criticizing this project while also encouraging me to keep going

Gabriel Săndoiu—for inspiring a desire to know more

Larisa Avram and Pusi Cornilescu—for introducing me to syntax

József Fiser—for prompting me to revisit mathematics

Basia Pomiechowska—for convincing me that stimuli should be psychologist-proof

Gabor Brody—for not allowing me to take anything for granted

Velisar Manea—for his outlandish views on everything

Dóri Mészégető—for the luxury of late mornings

Réka Finta, Edit Vitrai, Ági Volein—for ensuring that everything runs smoothly under the hood

George Soros and the ERC—for the cash money

Vjeran Kerić—for his warmth and humor

Bálint Varga—for his quirkiness and generosity

Thomas Wolf—for his subdued wisdom

Arianna Curioni—for her spirited conversation

Laura Schlingloff—for so many reasons

Martin Dockendorff—for some reason‡

‡ *for some reason, this dissertation is dedicated to him*

Abstract

Local assignments from visually available object symbols to entities under discussion underlie representational STAND-FOR relations and are ubiquitous across many forms of human communication, such as pretend play, puppet shows, diagrams, or animations (e.g., a banana stands for a phone, a puppet stands for an agent).

Chapter 1 lays out a cognitive architecture that can explain how humans represent STAND-FOR relations. The architecture consists of two representational layers—one for the perceptually available symbols (object indexes), one for the entities under discussion (discourse referents)—and an assignment function that maps the object indexes to the discourse referents. Once the mappings are established, the information conveyed through the symbol object is interpreted as applying to the discourse referent. I illustrate the architecture with early object substitution pretense and argue that it provides a better and more general account of pretend play than alternative views.

Chapter 2 asks whether 19-month-old infants take on-screen events to occur in the here and now or think that on-screen events are decoupled from the immediate environment. Across four experiments, I show that infants reject animation–reality crossovers but accept the depiction of the same animated environment on multiple screens. The results are consistent with the possibility that 19-month-olds interpret animations as external representations.

Chapter 3 tests several components of the cognitive architecture outlined in Chapter 1. I present evidence that 15-month-old infants can map arbitrary visual symbols onto familiar discourse referents based on predicative expressions (e.g., "Look! A duck!") applied to geometric shapes (e.g., a circle). Additional experiments show (i) that infants restrict the assignments to the speaker who stipulated them; (ii) that infants use their conceptual knowledge when interpreting subsequent events involving the symbols; and (iii) that alternative explanations cannot account for the central finding. The results show that the cognitive mechanism underlying the representation of STAND-FOR relations is easily activated and available early in human ontogeny.

Chapter 4 moves from infants to adults and asks whether photographs of objects undergo object recognition or symbol interpretation. I present evidence from a Stroop task indicating that adults interpret images of toys as the objects the toys are toys of—not as the toys themselves. A control experiment shows that the association between an image of a toy and the object the toy stands for is not automatic. When images of toys are displayed against the objects the toys represent, adults interpret them as depictions of toys. The results indicate that adults interpret images as symbols and compute what the images stand for even when this is irrelevant to the task at hand.

Chapter 5 provides an overall summary of the empirical findings in Chapters 2–4. I then discuss a recent debate in cognitive development on the use of symbols in research—Theory of Puppets—and link it to the theoretical framework laid out in Chapter 1 and to the experiments in Chapters 2–4. I end by presenting several avenues for future research and one long-term theoretical goal of the project.

Table of Contents

Glossary

Symbol objects — Perceptually available objects used and manipulated to convey information about entities currently under discussion.

Object indexes — Mental representations of perceptually available objects. The object indexing representation tracks identity over time.

Entities under discussion — Entities that are relevant to the current communicative context. They do not have to exist in the world (e.g., "Imagine a fair coin tossed ten times").

Discourse referents — Mental representations of entities under discussion. The discourse referent representation handles individuation, feature binding, and discourse-internal coreference.

STAND-FOR relations — Relations between a symbol object and an entity under discussion. The relations are mentally represented in a pointer architecture that links an object index to the corresponding discourse referent.

External representation — The union of the symbols in a scene and their spatiotemporal arrangement. The semantic contents of representations are propositions. The arguments are the discourse referents; the predicates are the properties ascribed to the discourse referents.

Assignment — Internal function that takes an object index as input and maps it to its corresponding discourse referent. It establishes the STAND-FOR relation.

Tokening — Internal function that takes a concept as input and returns a conceptual description as output. The conceptual description is attached to the discourse referent.

Chapter 1. STAND-FOR Relations: The Theoretical Framework

1.1. Introduction: The Explanandum

Fifty years ago, when NASA was about to launch the first spacecraft to exit the solar system, astrophysicist Carl Sagan thought it would be a good idea to place a message on board the spacecraft (Crane, 2003; Sagan, 1973). The message was supposed to carry information about the spacecraft's origin, should far-away intelligent beings intercept it. As language was out of the question, Sagan and his colleagues designed an aluminum plaque depicting two hydrogen atoms, a pulsar map, the spacecraft's trajectory relative to the solar system, and a pair of human adults (**Figure 1.1**). Despite its compactness, the engraving managed to synthesize an impressive amount of information about its source (Planet Earth), its creators (humans), and itself (the two circles on the top left are meant as a legend to the whole engraving). Science enthusiast that he was, Sagan himself was confident that any species advanced enough to intercept the space probe would successfully decode the message.

Sagan's optimism reflects a heavy dose of naïve realism and the curse of knowledge. These two illusions trick people into believing that the world is transparent, its interpretation straightforward. The obstacles engineers face in computer vision and natural language processing (e.g., Russell & Norvig, 2020) attest that this is false. The fluency with which one navigates the world belies how sparse, noisy, and ambiguous it actually is (e.g., Clark, 2015; Tenenbaum et al., 2011). In this case, the fact that humans excel at detecting and interpreting meaning-carrying objects (such as diagrams) does not imply that a galactic species will interpret the engraving in the same way.

For one, representations do not carry their communicative purpose on their sleeve. As Crane (2003) rightly points out, for the extra-terrestrials to engage with the engraving at all, they must figure out that this object carries information in the first place (as opposed to playing an instrumental part in the functioning of the spacecraft, for instance). Whether they would entertain this hypothesis is far from clear, as experiments with nonhuman animals indicate (e.g.,

pointing is not transparent: Kaminski & Nitzschner, 2013; Tomasello et al., 1997). In human communication, speakers sometimes accompany their information-carrying actions with ostensive signals—establishing eye contact, addressing the audience by name—that indicate to the audience that communication is taking place (Csibra, 2010; Scott-Phillips, 2014; Sperber & Wilson, 1995). But even in the absence of ostensive signals, humans retrieve communicative intentions based on action efficiency (in the case of gesture detection: Royka et al., 2022) or object and pattern recognition (in the case of inscription detection: Dehaene, 2009), presumably because they have learned which stimuli in the environment are likely to carry information.

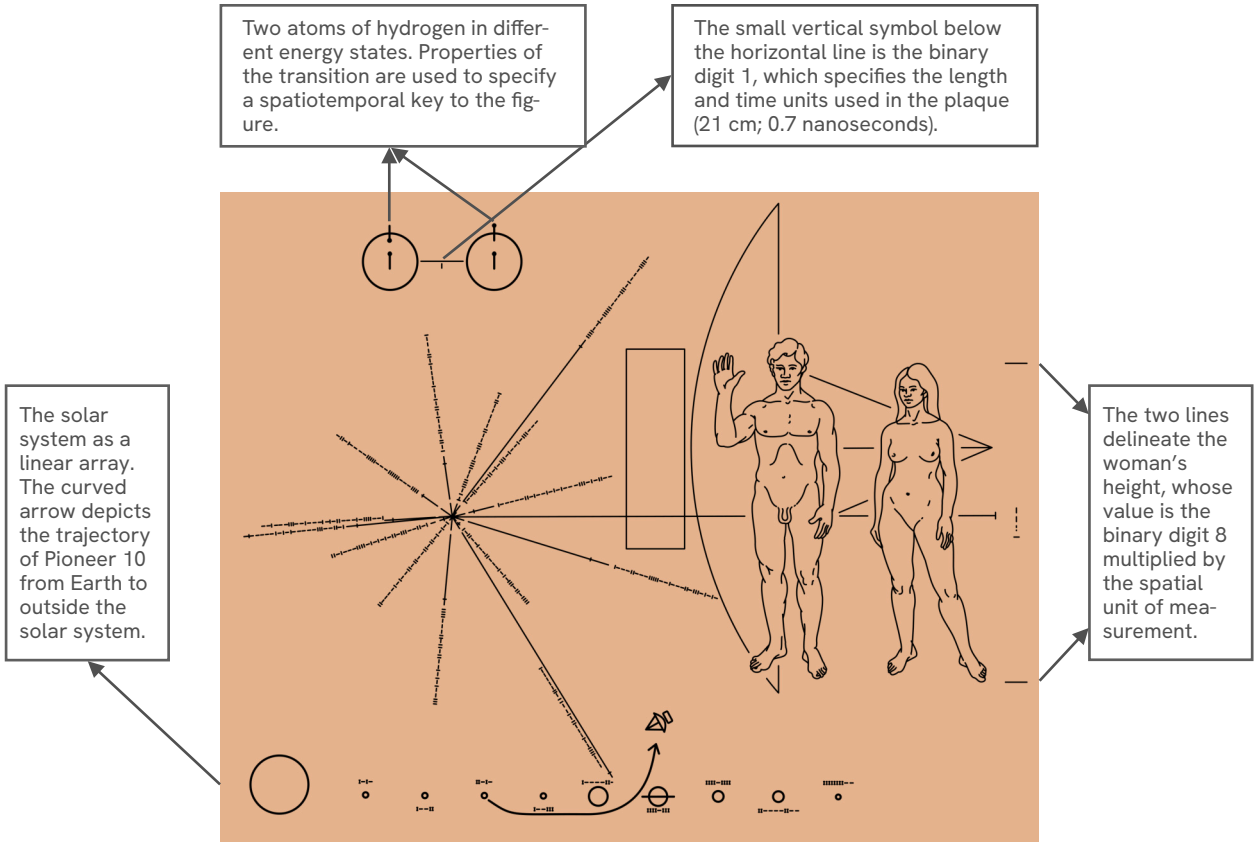

Figure 1.1. The engraving placed on Pioneer 10.

This barely scratches the surface of the problem. Even if one figures out that the engraving contains a message worth interpreting, a closer inspection reveals it to be much more equivocal than it may appear initially. Take the use of

horizontal lines, for instance. Sagan used these symbols in at least three different ways throughout the plaque. In the hydrogen diagram (**Figure 1.1**, top left), the horizontal line connecting the two circles illustrates the transition between two different energy states; in the linear solar system array (**Figure 1.1**, bottom center), another horizontal line depicts the rings of Saturn; and in the human adults drawing, the two horizontal lines are meant to delineate the woman's height. The ambiguity sharpens even more when considering the straight lines of all orientations. Nothing in the human body etchings would help a naïve observer distinguish joints from muscles, hairline from eyebrows, folds from contours.

Moving from symbols to symbolic relations, observe that the engraving contains a small fraction of the mappings humans make in communication: circles for atoms and planets, arrows for spatial trajectories, and relative sizes for actual size differences[1]. In addition, Sagan did his best to elucidate the mappings for the hypothetical addressee by providing a key for the spatiotemporal ones (see the upper left annotation to **Figure 1.1**). This illustrates the notion that external representations are governed by mappings that need to be recovered to extract the information contained in them (Millikan, 1984, 2017). Mapping rules can be highly conventionalized and general (e.g., widespread acronyms and traffic signs), domain-specific (e.g., number of stars for *p*-value bounds), or freshly minted, in which case they are often accompanied by a conventional device introducing the mapping (e.g., "henceforth referred to as X"; "suppose that X is Y"). However, nothing substantial hinges on this variation, as the interpreter's task is the same regardless of how conventionalized a mapping is: to figure out what each constituent symbol in the representation stands for.

Yet another source of variation comes from how the depicted content relates to the world. Sometimes, there will be a straightforward referential relation between a representation and the world (e.g., "the President of the United States"), but not always (e.g., "Once upon a time there was a queen"). Again, Sagan's plaque exhibits this divide. The two space probe drawings represent a particular object in the world (the space probe itself), but the two human figures do not. Mainstream thought on external symbolic representations, both theoretical and empirical, either shied away from analyzing drawings such as the human figures in Sagan's engraving (e.g., Greenberg, 2013) or subjected them to sepa-

[1] Note that relative size represents consistently in Sagan's engraving only if one partitions it into distinct subparts. Absent this invisible partition, the sun would be as large as a hydrogen atom and much smaller than humans.

rate treatment (Goodman, 1968; Rakoczy et al., 2005) on account of their lacking a world referent. But while the two human figures lack real-world counterparts and do not carry information about any world state, they are still representations of humans by any reasonable measure.

To a first approximation, this is the big picture of external symbolic representations that do not involve language. Humans use them to exchange information about objects, events, and states of affairs. Communicators provide interpretable evidence for their audience by externalizing information according to various mapping rules and in different representational formats. While representations are ultimately about the world, they can vary extensively in terms of how exactly they are related to it, which is dictated by the communicator's goals. On the receptive side, the audience has to recognize that a given stimulus is meant to carry information, parse the stimulus into its component constituents to individuate the symbols, figure out what each symbol stands for, analyze the relations between symbols to figure out what the content of the representation is, then decide whether and how to update their internal models of the world.

In this chapter, I offer a cognitive account for the interpretation of symbolic representations focusing on the semantics of their constituents—the individual symbols (Section 1.2). I model the relation between the symbols and the entities under discussion by introducing a simple formal operation, **assignment**, which creates local links between the mental representations that track them: **object indexes** for the external symbols (Pylyshyn, 1989) and **discourse referents** for the entities under discussion (Karttunen, 1976). I use **symbols**[2] to refer to the trackable physical entities that are the constituents of representations and **discourse referents** for mental representations of the entities under discussion. I invoke discourse referents, instead of referents, for two reasons. First, I want to clarify that symbols need not refer to anything in the world. Since the entities under discussion may be spatiotemporally undefined (e.g., "A kangaroo walks into a bar."), there is no reason to assume either that semantic completeness is a necessary condition for being a symbol or that symbols referring to

[2] I use "symbol" to cover not only representational objects with conventional meaning as, for instance, in Peirce's distinction between index, icon, and symbol (1897/1955), but any object that stands for an entity under discussion, regardless of how meaning is conveyed.

particulars in the world are the default in human communication[3]. Second, I want to emphasize that the referents, as well as their link to the symbols that represent them, are local to the current communicative context.

Once the links between the symbol and the discourse referent are established, information that applies to the discourse referent can be transmitted via symbol manipulation (sometimes literally). I use **STAND-FOR relations** to refer to the links between perceptually available symbol objects and discourse referents. Due to the mentally represented connection between the symbol object and the discourse referent, the predicates communicated via the object will be interpreted as properties belonging to the referent and not to the object that happens to stand for it temporarily.

I illustrate the ubiquity of this structure across many communicative devices (e.g., puppet shows, animations, drawings, graphs, memes) in which the interlocutor sets up object–discourse referent mappings as part of the interpretation process (Section 1.3). I then draw on early object substitution pretense to argue that the ability to interpret STAND-FOR relations develops early and reliably in human ontogeny (Section 1.4) and that such an ability accounts for the data better than alternative views (Section 1.5).

1.2. Architecture: The Explanation

The links between trackable object symbols and entities under discussion form the core of external representations. In this section, I sketch a cognitive mechanism that can explain how these links are set up during interpretation. Inspired by research on two cognitive mechanisms proposed in early vision and formal semantics, I hypothesize that two representational layers are involved in interpretation. The indexing layer tracks objects in a scene without necessarily conceptualizing them; the communicative layer tracks individuals that are currently being communicated about. I then introduce two functions: (i) **tokening**, which provides mental representations in the communicative layer with a conceptual (not linguistic) label; and (ii) **assignment**, which provides links between the two

[3] A distinction between the type and the token level is in order. What I mean here is not that external representations are disconnected from the world. Instead, I mean that there need not be a one-to-one mapping between depictions and particulars in the world. For instance, the drawings of humans in Sagan's engraving do not pick out any existing persons (token level), but they can still be used to illustrate how humans are shaped (type level).

layers. The output of these representations and processes consists of ordered pairs (X, Y), where X is a mental representation of an object that stands for an entity under discussion, which is mentally represented by Y.

1.2.1. Trackable Objects

To interpret the relation between a symbol and an entity under discussion, a system is needed that can represent, individuate, and track the objects that are temporarily used as symbols. Consider the ordinary object tracking system as an analog for the symbol tracking system. Humans extract distinct objects from visual scenes via an indexing system that provides pointers to these objects (Kahneman et al., 1992; Pylyshyn, 1989). These pointers individuate objects in the early stages of visual processing, maintain object identity across movement (Pylyshyn, 2001; Pylyshyn & Storm, 1988), and survive brief periods of occlusion (Scholl & Pylyshyn, 1999). While visual features can be bound to the object indexes (what the object looks like; Scholl & Leslie, 1999), the indexing system is limited in its ability to provide conceptual information about the objects (Carey, 2009) or downright unable to do so (Brody, 2020). The system is thus indifferent to the kinds of objects it indexes. In this sense, the class of things to which the indexing system is attuned consists of Spelke objects—connected and bounded entities that preserve cohesion in motion (Spelke, 1990). Even if pencils and dragons differ sharply in appearance, this is irrelevant to the indexing system: both objects will be assigned pointers all the same.

The object indexing system takes a visual scene as input and outputs a layer of representation $R(O)$ of individuated objects $\{o_1, \ldots, o_n\}$. At this level, the object in the here and now o_i causes the mental representation $R(o_i)$, while $R(o_i)$ refers to and carries information about o_i. Although the visual features of o_i can be bound to $R(o_i)$, conceptual information (e.g., the kind to which o_i belongs) need not be represented at this level. While the indexing system has physical objects in its proper domain, it can also be triggered by other types of input, such as marks on paper or pixel constellations. Indeed, most empirical evidence for visual indexes comes from experiments with computer-generated stimuli, so there is no question that density-less objects can trigger visual indexes.

While I am modeling the object indexing system for tracking symbols based on Pylyshyn's (1989) visual indexing system, I do not mean that the indexing system for tracking symbols is the same vision uses to index objects. First, symbols have to be individuated over much longer timespans. Second, the presence of an object input that causes a visual index is often not necessary for the link between a symbol and discourse referent to be set up, as the symbol itself can be created by stipulation (e.g., tracing the contours of an imaginary object by hand: Müller, 2013) or, even more abstractly, by a point to an empty location in space (So et al., 2009). As with props, the indexed location can be pointed at to pick up the discourse referent previously introduced. The object indexing system used to set up STAND-FOR relations must thus reside beyond early vision and may even constitute a separate system that subserves only the interpretation of symbols. For expository purposes, I will, however, continue to model the object indexing system on Pylyshyn's visual indexing model (1989) while being agnostic about whether or how they are related.

1.2.2. Discourse Referents

Beyond the ability to individuate objects in a visual scene, humans can also individuate entities under discussion in a conversation or discourse without visual support. To account for this ability, work in **discourse representation** theory advanced an internal mechanism that can track entities under discussion for as long as the discourse lasts. The mental representations created by the system, called **discourse referents**, enable discourse-internal co-reference (Karttunen, 1976) and the accumulation of information as the discourse unfolds (Heim, 1982; Kamp & Reyle, 1993). The properties ascribed to the entities under discussion are represented as predicates, which take the discourse referent as arguments. The following discourse, for instance, is internally represented as containing two discourse referents, **a** and **b**, where **a** = John and **b** is an object that has the property of being a red-scaled dragon:

(1) John met a dragon$_i$. Surprisingly, it$_i$ had red scales.

While discourse reference was initially introduced to deal with natural-language phenomena, it has recently been extended to cover communicative acts more broadly (Brody, 2020). For instance, pointing to an object behind an occluder (which impedes the audience from seeing it) may prompt the audience to

create a discourse referent and expect that subsequent communication will involve that object.

In sum, the discourse referent indexing system takes noun phrases (and possibly instances of pointing) newly introduced in a discourse or pointing gestures as input and outputs a layer of representation D, $\{d_1, d_2, \ldots, d_n\}$, denoting the set of entities under discussion, $\{e_1, e_2, \ldots, e_n\}$. At this level, whether there is a relation between a representation and the outside world depends on the discourse, not on the discourse referent layer itself, as discourse reference is a discourse-internal phenomenon. As such, it can only access the relation between a mental representation and an entity under discussion (Heim, 1982). Whether the concepts recruited for creating a mental representation pick out real-world entities is discourse-independent, as there are no existence restrictions on the entities humans can invent and exchange information about, as any joke, novel, or thought experiment attests. While example (1) provides a straightforward illustration in which a discourse referent picks out a member of a fictional kind, the same principles apply to fictional tokens of real kinds:

(2) Once upon a time, John met a koala$_i$. Surprisingly, it$_i$ had red fur.

1.2.3. Tokening

When reference to an object in the world does not hold in discourse, how is the meaning of a discourse referent established? In other words, how does the cognitive system distinguish between the meaning of "Once upon a time there was a young **queen**" and the meaning of "Once upon a time there was a young **boy**"?

One possibility is that discourse referents are assigned variables and are individuated like visual objects—based on their address in memory (Brody, 2020; Yu & Lau, 2023). For instance, "Once upon a time, there were a queen and a boy" introduces two discourse referents, each receiving a dedicated index: d_1 and d_2. Then, the information that one of them is a young queen and one is a young boy may be linked to the corresponding referents by feature binding: **queen**(d_1), **young**(d_1); **boy**(d_2), **young**(d_2). This aligns with **discourse representation** theory, which does not privilege any property over others: queenness and youth are on equal footing in the internal representation. The advantage of this representational format is that it allows multiple symbols to stand for multiple entities of the same type: **boy**(d_2) and **boy**(d_3). The disadvantage is that it fails to explain

why certain properties seem privileged over others. People want to know what kind of thing a symbol stands for (e.g., "Did you draw a balloon or a lollipop?"), and they use these descriptions for co-reference in discourse—compare (3a) vs. (3b). However, such preferences may stem from the organization of the conceptual system (Rosch et al., 1976) and not from the way discourse referents are individuated, so I will leave this possibility open.

(3) Once upon a time, there was a young queen.
  (a) The queen had so much privilege.
  (b) ?The young being had so much privilege.

A second possibility, which overcomes this hurdle, is that discourse referents are individuated by conceptual description: **queen** and **boy**. If noun phrases are instructions to fetch concepts (Pietroski, 2018) and if the human conceptual system is generative (Quilty-Dunn, 2020), it should be possible to create new conceptual tokens on the fly and attach them to the discourse referents. The tokens do not pick out a particular object unless the conceptual description defines them as having a real referent. In other words, the tokens are not spatiotemporally defined with respect to the actual world[4], and their referents are fictional. This does not mean that the tokens do not bear any connection to the world but rather that the conceptual system mediates this connection. Concepts are causally connected to the world but the world and the token are independent once the concept is selected. If, for instance, you were asked to imagine a fair coin tossed ten times, the representation you would use to keep track of it for this (probably pedagogical) episode would not pick out any existing coin. Instead, the invitation would prompt you to deploy your COIN concept to generate a new instance for this occasion. Under this view, discourse referents are individuated by conceptual description via a **tokening** operation, which takes as input a concept in the conceptual system, $c \in C$, and outputs a token of the concept mapped onto the discourse referent d, whose content might even inherit (a subset of) the properties of the class picked out by the concept c. In short, discourse referents can be considered concept tokens under this view.

[4] One could envisage a possible-worlds semantics account (Menzel, 2023), whereby truth values would be extended beyond the actual world, but I will not pursue this possibility here, given the relatively late emergence of modal concepts (Leahy & Carey, 2020). In any case, a possible-worlds account would not change the story radically, as discourse referents would not be evaluated with respect to the actual world under a possible-worlds account either.

This second possibility is formally similar to **mental file** theory (Perner et al., 2015; Perner & Leahy, 2015; Recanati, 2012), a proposal about the representational format used when encoding particular objects. In **mental file** theory, objects are always represented under a description, which provides the head of the file and which forms the individuation criterion internally. In the account put forth here, the proposal that discourse referents are individuated by description, not only by indexing, has the advantage of accounting for the same data that the mental file approach accounts for. For instance, this option explains why it is more difficult for children to give an object an alternative name (e.g., "bunny" after it has been labeled "rabbit") than to name an alternative property of the object (e.g., to say of a white–gray bunny that it is gray after the experimenter said it is white; Doherty & Perner, 1998). However, even if conceptual descriptions head discourse referents, these descriptions will have to be supplemented by indexes when different referent tokens belong to the same type. While I remain open to both possibilities, I will assume in what follows that discourse referents are individuated by conceptual description (± an index).

How one selects a conceptual description over another goes beyond the scope of this chapter, as the selection criteria are orthogonal to setting up the STAND-FOR relation itself. Symbolic representations can even be interpreted without possessing the requisite concept (e.g., children, for instance, routinely learn about animals never encountered before from picture books). That said, there are at least four types of cues one can use to establish the kind of entity that a discourse referent belongs to: (i) the visual features of the symbol (e.g., all else being equal, a drawing of a circle represents a circle); (ii) the behavior of the symbol (e.g., a drawing of a circle represents an agent when it appears to be self-propelled); (iii) linguistic stipulation (e.g., a drawing of a circle represents the median of a distribution if the legend explicitly says so); and (iv) convention (e.g., ♀ is a gender symbol).

1.2.4. Assignment

One last component needs to be added to the interpretive system sketched so far to allow for local connections between entities in the visual representational layer and those in the discourse layer. To depict the encounter of John with a red dragon in (1), instead of describing it in language, an external representation can be used, such as the one in Figure 1.2, where the stick figure (an indexed ob-

ject) stands for John (a discourse referent), and the remaining drawing (a second indexed object) stands for the red dragon (a second discourse referent).

A simple linking function connects the two mental representational levels by providing a pointer from a visual index to a discourse referent, thereby creating stable local relations throughout the discourse. While the linking function operates on internal representations, it is interpreted by the cognitive system as an assignment from an object to an entity under discussion.

Figure 1.2. Depiction of John meeting a red dragon.

The inputs to the assignment function are (i) a representation of a visual object $R(o) \in R(O)$; and (ii) a discourse referent $d \in D$. The output is a pointer from $R(o)$ to d such that the external symbol o, represented in the visual indexing system, stands for e, the entity under discussion represented in the discourse referent layer. Consequently, the entity under discussion can be referred back to within the current communicative episode via the symbol object.

To avoid confusion, the mappings need to be locally stable within any given discourse. In other words, the assignment operation should be a one-to-one function, such that a symbol can stand for only one entity under discussion and vice versa. Besides avoiding confusion, a one-to-one relation would also make it trivially easy to invert the mapping, thus enabling the move from the discourse referents to the symbols (e.g., “Give me the horse” when requesting somebody to hand a drawing of a horse). Finally, the mappings should be discarded once

the current discourse is over because the symbols need not represent the same entities across discourses. Figure 1.3 illustrates the components of the entire model.

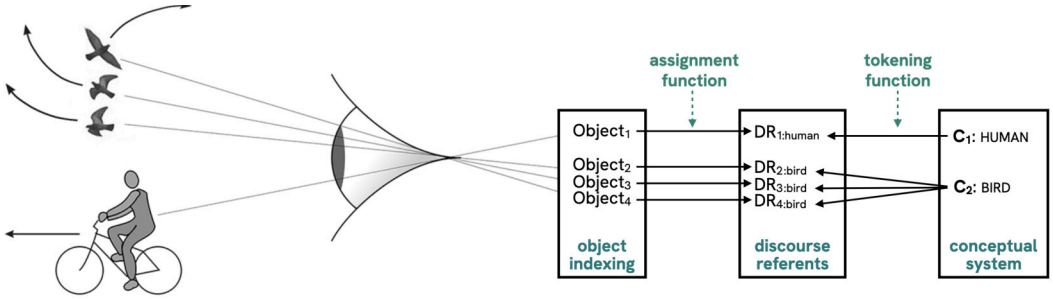

Figure 1.3. STAND-FOR relations between symbol objects and entities under discussion. The objects in the scene (in this illustration, the four individual drawings) cause mental representations of objects that index and track them. These indexes point further to a discourse referent layer, where descriptions are tokened from the conceptual system.

1.2.5. From Symbols to Representations: Predicate Application

The function of symbolic representations is to convey information to an audience. How do they achieve this? I have already laid out how symbols, the constituents of representations, get their interpretation. The perceptually available objects, represented internally by object indexes, are linked to a second representational layer, which indexes the entities under discussion. But this alone would not be very useful, as it would be equivalent to a language consisting only of nouns. However, once the assignment between an object and a conceptual token is in place, one can update one's internal model of the discourse referent according to the information generated by the communicator. The information accumulation process should consider the actions performed on the symbol by a communicator or by the symbol itself in the case of autonomously dynamic stimuli (e.g., animation) and turn them into predicates attributed to the discourse referent. While a complete account of how predicates are conveyed with symbol objects is outside the scope of the current chapter, I will point out that the spatiotemporal arrangements of symbols (spatial: drawings and graphs; spatiotemporal: animations, comic books) might be internally translated into predicates that apply to the discourse referent arguments that symbols stand for. While the relations the symbols enter and the actions the symbols engage in occur in phys-

ical space, the corresponding predicates need not have spatial meaning. While spatial relations in maps often preserve the spatial relations in the mapped territory, this is not always the case. In graphs, space often depicts various types of nonspatial magnitudes; in pretend play, when pretend-feeding toy animals, moving a pretend food item toward the toy's mouth corresponds to feeding that item to the animal.

In sum, nonlinguistic external representations are mentally represented in a propositional predicate–argument format. The symbols supply the arguments, and the spatiotemporal configurations the symbols are embedded in supply the predicates[5] (see Mussavifard, 2023, for a related proposal). In what follows, I will continue to focus on the link between symbols and discourse referents more than on the relation between representations as a whole and the propositions they convey.

1.3. Explanatory Scope: Representational Media

One needs to identify the relevant communicative contexts to establish the STAND-FOR relation between an external object and a discourse referent (instead of merely tracking objects for, say, hunting). Humans often preface communicative acts with ostensive signals (e.g., eye contact) that have this function precisely: they inform the audience that communication is taking place (Scott-Phillips, 2014; Sperber & Wilson, 1995). However, many communicative artifacts fulfill their function without ostensive signals or a communicator that is present (e.g., drawings, statues, movies, animations). Therefore, humans need experience with these classes of artifacts to be able to identify them as a class of objects which (i) have the function of conveying information; and (ii) establish their function partly by using objects to stand for discourse referents.

In all these cases, the objects, together with the actions performed on them, create physical scenes through which communicators depict events, relations, and properties of the discourse referents (Clark, 2016). Setting up the links between physical symbols and discourse referents is fundamental to all forms of communication where the object indexing system of the interlocutor is recruited in the interpretive process. I outline several communicative devices that exploit

[5] This is unlike language, which conveys both arguments and predicates by the same type of symbols. Some predicates receive their own symbols in nonlinguistic representations as well but this is rare and may require conventionalization (e.g., arrows for movement).

STAND-FOR relations below to illustrate the phenomenon's ubiquity and a glimpse at the contents that can be conveyed via these relations.

Consider narrative fiction. Both oral storytelling and written novels require the audience to set up multiple discourse referents to keep track of who does what to whom. But there are also many ways in which external symbols can be added to replace or supplement the linguistic narrative. Actors playing characters and props standing for objects give rise to theater and live-action movies; puppets operated from behind the stage give rise to puppet shows; static drawings displayed in rapid succession give rise to animations; and static drawings displayed next to one another give rise to comic books. In all these cases, mind-external entities (actors, puppets, animated figures) are temporary stand-ins for entities under discussion (fictional characters). The content of the "Snow White" fairy tale, for instance, can be rendered as a live-action movie, as a puppet show, or as an animation, as Figure 1.4 illustrates.

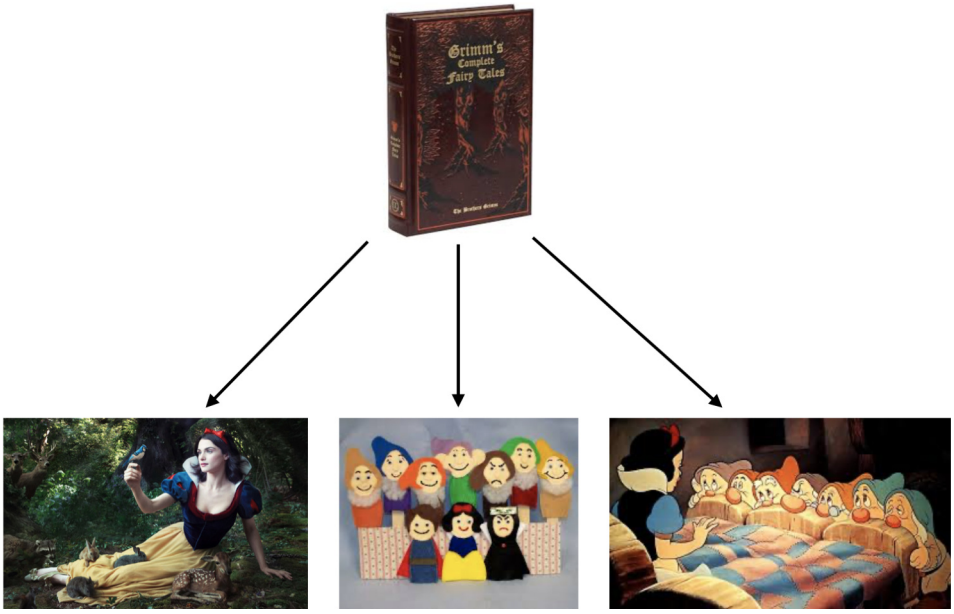

Figure 1.4. The propositional content of "Snow White" can be depicted in a live-action film (left), puppet show (center), or animation (right).

Beyond fiction, procedures and rules can also be depicted in external representations such as diagrams. The offside rule in European football, for instance, states that "a player is in an offside position if any of their body parts,

except the hands and arms, are in the opponents' half of the pitch, and closer to the opponents' goal line than both the ball and the second-last opponent" ("Offside Association Football", 2021). The rule is difficult to assimilate from language alone, at least not without parsing the above sentence several times. Figure 1.5 provides a better pedagogical tool by depicting an instance of an offside spatial configuration. In this diagram, the blue circles stand for the attacking team, the red circles for the defending team, and the small white circle for the ball. Because the leftmost blue player is beyond the second-last red opponent (marked by the dotted line), they are not allowed to play the ball should they receive it while in this position. While the entire diagram illustrates a real-world rule, the symbols it consists of do not pick out any particular footballer or ball in the world. As with Snow White, one can teach the offside rule through other media, such as animations or props (e.g., by using bottles for the players and a bottle cap for the ball and manipulating the objects on a table in front of the audience).

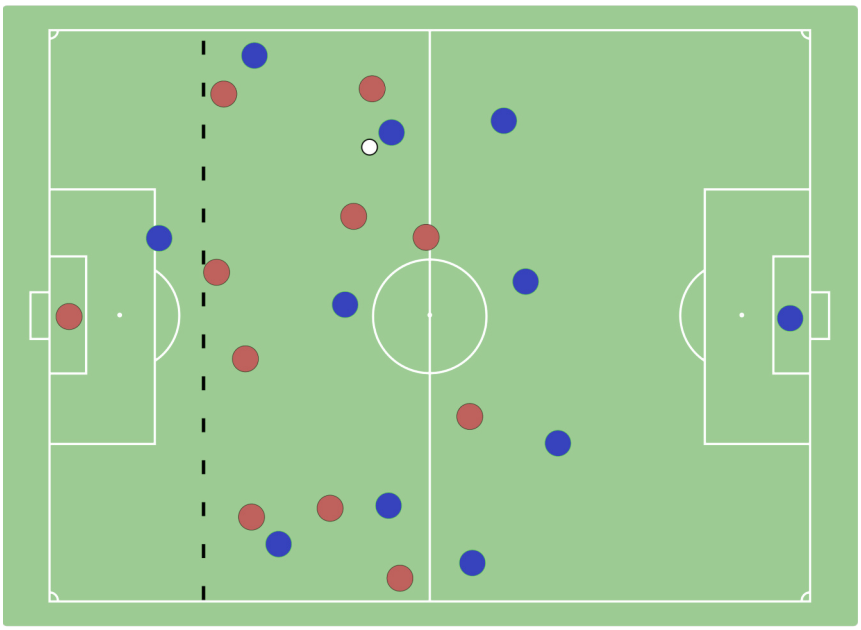

Figure 1.5. The blue dot/striker to the left of the dotted line is currently behind the second-last red dot/defender, and therefore in an offside position.

Scientific graphs, internet memes, and assembly instructions work in precisely the same way: visual objects are assigned a temporary conceptual identi-

ty (by linguistic stipulation or iconic features) and used to convey information about the entities they represent (Figure 1.6). While external representations occur in space, the information predicated about the entities under discussion need not be spatial. For instance, in the graph on the left of Figure 1.6, space is used to convey magnitude, as customary in the Cartesian coordinate system.

Figure 1.6. (Left to right) Scientific graph depicting the relation between length and width in three different flower species; internet meme depicting a relation between people and ideologies by exploiting a clichéd relationship dynamic; IKEA assembly instruction depicting how to mount the legs of a table.

Before looking into the development of understanding representations, I would like to make a few conceptual clarifications often glossed over in the literature. I grouped puppet shows, drawings, animations, and live-action movies as a unitary class of stimuli and put forth a cognitive architecture that can handle the basic requirements these stimuli have in common. In all these cases, I have claimed, communicators use perceptually available objects as temporary symbols for entities under discussion and convey information about the latter through the former, according to what the corresponding representational medium affords.

Alternative ways of classifying representations, while related, miss this defining criterion. In her discussion of human uniqueness, Millikan (2017) points out that humans may be the only species that can deal with **detached-content** stimuli—stimuli that carry information about states of affairs that are distant in space or time and do not indicate their relation to the perceiver. She cites sentences about absent entities, animal tracks, and video recordings as instances. According to Millikan's classification, a video taken by a surveillance camera is equivalent to a video of a theatre play because the video contents are detached from the perceiver in both cases. However, only the second one counts as a rep-

resentation involving STAND-FOR relations under my classification. What videos do is capture light patterns on film to allow humans to perceive the same stimuli from a temporal or spatial distance: devices for **tele-perception**. Their content is detached, in Millikan's sense, but there need not be any symbols or discourse referents involved. Consider the difference between watching a theater play live and watching a recording. While the recording does add a layer of detachment (because the actors and props are not in the here and now), it does not add a representational layer over and above the one that links the actors and props to the characters, objects, and places the play is about.

A further way of carving up the stimuli space has focused on the format in which the stimuli appear: three-dimensional versus two-dimensional (e.g., Pierroutsakos & Troseth, 2003; e.g., Snow & Culham, 2021; Troseth & DeLoache, 1998). While format may be statistically related to the property of being representational, it is not logically related to it. Two-dimensional stimuli need not be symbolic, as the surveillance video example in the previous paragraph illustrates, and conversely, three-dimensional stimuli can be symbolic, as the theatre example shows. In sum, while two-dimensional stimuli with detached content, such as videos and pictures, are prevalently used symbolically in communication, their use—not their form—turns them into communicative stimuli.

1.4. Development

1.4.1. STAND-FOR Relations in Infants

Little is known about infants' early grasp of STAND-FOR relations. This topic has been under-addressed in developmental research, even though representational stimuli are routinely used to tap into infants' cognitive processes (Packer & Moreno-Dulcey, 2022).

Moreover, the little that is known has been used as evidence that infants do not possess an early symbol concept. On the one hand, infants discriminate perceptually 2D representations from 3D objects before their first birthday (DeLoache et al., 1979) and correctly identify the objects represented, possibly even in the absence of experience (Hochberg & Brooks, 1962). This does not show that infants interpret the pictures as representations—that is, by establishing a link between an indexed object and a discourse referent. Instead, infants could respond to some other cue, such as surface similarity. On the other, while able

to discriminate between 2D and 3D stimuli, infants sometimes mistakenly treat photographs and videos as 3D objects (DeLoache et al., 1998; Pierroutsakos & Troseth, 2003). This is often cited as evidence against the notion that infants understand representations (DeLoache, 2004; Spelke, 2022).

This is a questionable conclusion for at least three reasons. First, infants' grasping behavior does not seem to be directed to the picture itself but to its content. When mistakenly treating photographs or videos as 3D objects, 9-month-olds adjust their grabbing behaviors to the depictions of objects whose life-size would afford picking up; in other words, they are less likely to pick up a picture of a bed than that of a bottle (DeLoache & Burns, 1994). This suggests that some interpretive process over the symbol is already taking place. Second, the inference conflates the distinction between 2D and 3D stimuli with the distinction between representational and nonrepresentational stimuli. The only inference that can be drawn is that infants do not yet understand what kind of thing 2D stimuli are, not that they do not understand symbols. Third, this conclusion glosses over the fact that experience with a representational medium might be necessary before engaging with it appropriately. There is an apocryphal story about people running away from the screen the first time they saw the short Lumiére brothers movie depicting a train approaching the station. It is unclear whether this happened, but if it did, it would be absurd to conclude that these adults lacked representational understanding. What they would have lacked was knowledge about the medium, in particular its affordances and its interaction with the immediate surrounding environment. Coming back to infants, this surely is a body of knowledge they must build over time, perhaps guided by adult scaffolding and aided by the ostensive signals that mark objects as symbols.

The earliest experimental evidence for the ability to represent symbolic relations comes from a neuroimaging experiment done on 5-month-old infants by Kabdebon and Dehaene-Lambertz (2019). Previous research had shown that young infants are sensitive to abstract syllabic structure, even in pseudo-words (Marcus et al., 1999). Still, there was as yet no evidence that the syllabic structures abstracted from the input (e.g., AAB) are mentally represented by a single variable (explicitly) or whether they are embedded in neural networks (implicitly). The authors reasoned that only if infants represent structures as a single variable will they be able to easily pair them with an arbitrary stimulus such as an image. In the familiarization phase of their experiment, 5-month-old infants were presented with successive pairings of trisyllabic pseudo-words conforming to

the same syllable structure followed by a picture (e.g., AAB-structured words such as "nonofe", "gagalu", "titina" were followed by a cartoon fish; ABA-structured words such as "rutaru", "kemike", "ladila" were followed by a cartoon lion). At test, infants were presented with pairings between images and novel pseudo-words that either conformed to the same structure as in familiarization (on congruent trials) or did not (on incongruent trials). EEG data revealed (i) that infants learned the correct mappings from the abstract rule governing the pseudo-words to the corresponding image (Experiment 1); (ii) that they can do so even when one of the images is replaced by an arbitrary label (Experiment 2); (iii) and that this mapping is not merely predictive but bidirectional—the effects hold even when infants are tested on pairings presented in the opposite order from familiarization (the image is followed by the pseudo-words; Experiment 3).

Kabdebon and Dehaene-Lambertz take the data to support the early emergence of cognitive operations that support symbolic relations—here, the relation between a picture and an abstract rule governing syllabic structure. While the study provides good evidence that infants represent abstract syllable structures with **internal** mental symbols (otherwise, it would not have been as easy to pair an external stimulus, such as a picture, with a syllabic rule), it remains unclear whether infants interpreted the pictures themselves using the architecture in **Section 1.2**. This would have been the case if infants took an abstract rule governing syllabic structure (e.g., AAB) as a concept of which the pseudo-words (e.g., "nonofe") are tokens. Under this account, infants learned during familiarization that a picture always stands for tokens of the same concept and detected a semantic violation when it did not. However, this requires the assumption that the pseudo-words themselves are conceived of as discourse referents, which seems far-fetched, at least at this age. In addition, two other possibilities seem more likely. First, infants may have taken the pseudo-words to be the symbols—and the pictures the referents—rather than the other way around (Spelke, 2022). In this case, the study would not show anything about visual symbols since the symbols would have been linguistic. Second, infants may have interpreted both the pseudo-words and the pictures as symbols (e.g., for an absent referent) and figured out the systematic relations between them. In this case, the data would be silent on the early emergence of STAND-FOR relations because the relation between a symbol and its referent would not have been tested.

1.4.2. STAND-FOR Relations in Toddlers and Children

If the infant literature on STAND-FOR relations as conceived in the current chapter is sparse, the literature on early pretend play[6] offers plenty of insight into the links between symbols and discourse referents. As object substitution pretense requires children to set up precisely such relations, early competence in this domain implies the early emergence of this capacity. Typically, a pretend play setup involves toys or neutral objects and an adult experimenter who stipulates pretend identities on props, gives children some information about them, then prompts the children to act on the objects or manipulates the objects herself according to simple scenarios.

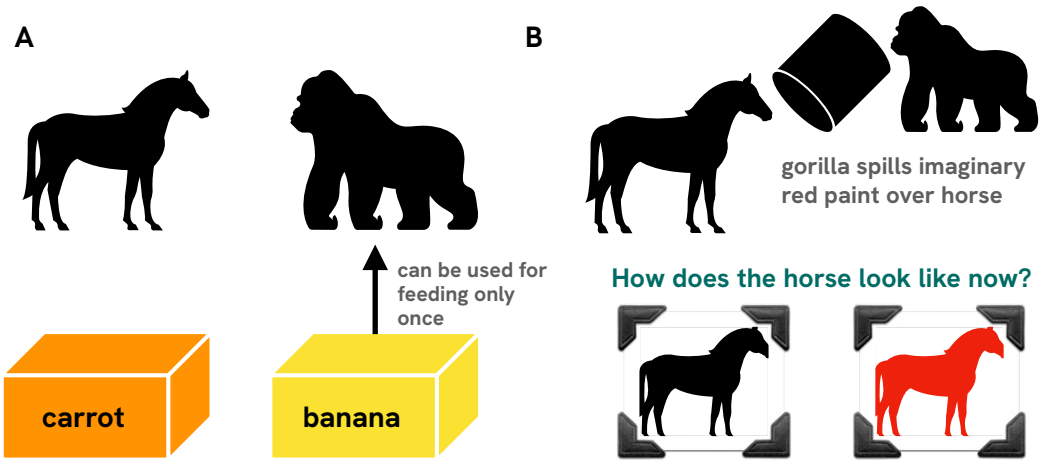

Figure 1.7. (A) Schematic representation of Harris and Kavanaugh (1993, Experiment 1). After mapping the orange blocks to carrots, and the yellow blocks to bananas, 2-year-olds perform pretend actions according to preference: carrots to horses, bananas to gorillas. A block is no longer a good candidate for feeding after having been used already, even though it is still physically present in the scene. **(B)** Schematic representation of Kavanaugh and Harris (1994, Experiment 2). If the experimenter pretends that the gorilla spilled paint over the horse and asks 2-year-olds how the horse looks now, they choose the picture depicting the imaginary transformation.

[6] For expository purposes, I will use "pretend play" and "object substitution pretense" interchangeably, but I will note here that there are cases of pretending that do not involve props in a straightforward way (e.g., when the caregiver pretends they are a wolf). These cases may also require receivers to set up STAND-FOR relations but the fine-grained distinctions between these cases and the object substitution ones go beyond the scope of the current chapter.

By their second year of life, children choose the correct prop for different pretense actions (Bosco et al., 2006; Harris & Kavanaugh, 1993). If they are faced with a toy gorilla and a toy horse and with orange and yellow blocks that they are told stand for carrots and bananas, respectively, they will feed the two animals their preferred food by choosing the appropriate blocks. Two-year-olds can also follow predicated causal transformations undergone by the discourse referents (not the symbols). They refrain, for instance, from using a banana block if the gorilla has eaten it already, even though the block is still physically present in the scene (**Figure 1.7A**).

Moreover, 2-year-olds can select pictures representing the imaginary transformations that an experimenter enacts. In one scenario (**Figure 1.7B**), the experimenter pretended that the gorilla spilled red paint over the horse (by prompting the child to pretend that the empty container contained imaginary red paint and tilting it over the horse). Children were then shown two pictures, one depicting a toy horse identical to the one in front of them, the other depicting a similar-looking horse colored red. When asked to show how the horse looks now, toddlers pointed to the picture depicting a red horse, ignoring the perceptual similarity between the other picture and the toy horse in front of them (Harris et al., 1997; Kavanaugh & Harris, 1994). In other words, they can answer questions based on what is being predicated about the discourse referents, which overrides the appearance of the physical objects in front of them.

It is possible, of course, that children at this age take the stipulated mappings as kind membership predication rather than a temporary assignment function ("orange block $\in$ {CARROTS}" instead of "orange block $\leftarrow$ carrot"). However, this does not seem to be the case, as young children keep symbol–discourse referent assignments distinct across both discourses and speakers (Andrasi et al., 2022; Harris & Kavanaugh, 1993; Wyman et al., 2009), suggesting that the relations created by the assignment are discourse-specific and temporary. To add to the evidence that symbols and discourse referents are kept apart, children rarely confuse imaginary and actual events (Bourchier & Davis, 2002; Weisberg, 2015), they keep make-believe identities distinct across contexts and speakers (Wyman et al., 2009), and they do not think that pretend objects materialize (Golomb & Galasso, 1995; Woolley & Phelps, 1994).

Finally, these experiments also support the reality of the tokening and predicate application operations. Otherwise, toddlers would be unable to follow the experimenter's stipulations (e.g., yellow block $\leftarrow$ banana) or the events de-

picted subsequently (e.g., the horse getting covered in red paint). These operations work in a top-down manner, as they are unrelated not only to any immediate percept but to any percept whatsoever. In typical object substitution pretense, there is no specific object that the props stand for—neither the horse nor the food morsels have real-world counterparts. Thus, there is no mental representation of a particular object the toddler can access when keeping track of what is happening. Instead, she will need to create these representations from scratch via her internal conceptual system that can generate descriptions based on the experimenter's stipulations. On top of that, toddlers' conceptual system seems to be at work even for predicate application, given that they compute the implications of the propositions explicitly conveyed by the experimenter. When the experimenter communicates that the gorilla tilted the bucket over the horse, the imaginary paint is assumed to behave like ordinary paint: it is subject to gravity and stains the unlucky individuals it intersects on its way down.

Early object substitution pretense thus suggests that the ability to set up local relations between symbols and discourse referents emerges early in development. Two-year-olds can assign a temporary label Y to a spatiotemporally trackable object X (X $\leftarrow$ Y) and do this in a way that preserves the distinction between X and Y. This is shown by the fact that assignments are discarded when the pretense is over, so that the object X becomes free to stand for new discourse referents. Glossed this way, pretend play is a subspecies of communicative exchange, much like drawings, animations, and other media that rely on the interface between object indexing and discourse referents. Thus, pretend play is not a standalone cognitive feat but a manifestation of a cognitive subsystem that handles a broad class of stimuli. I have dedicated an entire subsection to it because it provides an excellent case study for the architecture in **Section 1.2**, as it exhibits all the components I am advocating for. In addition, pretend play allows researchers to be sure that toddlers do not confuse symbols and discourse referents[7]. I am not the first to notice the parallels between pretend play and other representations (e.g., aesthetics: Gombrich, 1963, Chapter 1; Walton, 1990; linguistics: Clark, 2016; psychology: Piaget 1945/1962, Vygotsky, 1966/2015), but

[7] This, however, does not imply that symbols and discourse referents are not allowed to coincide in external representations. Nothing in the architecture stops a symbol from standing for itself (e.g., actors playing themselves in movies; illustrating what happened to my phone yesterday via my phone). In such cases, there will still be two mental representations involved—one for the symbol, the other for the discourse referent—but the objects in the world these representations pick out will coincide.

there are few accounts of the cognitive mechanism underlying the capacity to interpret objects as symbols of other entities writ large. I turn to them in the next section.

1.5. Alternative Explanations

1.5.1. The Decoupling Account

The first extended cognitive account of pretend play is Leslie's (1987) meta-representational model. Leslie was the first to point out that pretense shares critical features with the semantics of mental state verbs. One can thus believe/pretend that this cup is empty (even though it is full), that this cup is a seashell (even though it is not), or that there is a second cup on the table (even though there is none). Like BELIEVE, the PRETEND predicate also blocks inference to the truth of the embedded proposition in all these cases. Based on this parallelism, Leslie argued that toddlers should quarantine pretense representations to avoid their representations about the world becoming corrupted by pretense-induced misrepresentations. He assumed that toddlers start a pretend play game with a primary mental representation that accurately captures the perceived scene (e.g., "This object is a banana"). They then create a copy of this primary representation that is decoupled from its regular input–output relations and that can be modified based on the pretend play stipulations (e.g., "This banana is a phone"). Decoupling ensures quarantining, so toddlers will not leave the pretend play episode with phony ideas about bananas.

Subsequent theorists have added the claim that fictional propositions should also be stored away from representations about the real world. The behavioral account (Lillard, 2001; Nichols & Stich, 2000) postulated a Possible Worlds Box that separates pretend contents from the rest of the cognitive system. A prominent evolutionary model (Cosmides & Tooby, 2000) argued for a system of scope tags that travel along with representations and block upward inference accordingly. Regardless of implementation, there has been widespread agreement that pretend representations must be stored separately from reality representations (Lillard et al., 2011; Weisberg, 2015; Wyman et al., 2009).

However, I have reasons to think that decoupling and quarantining are not needed to account for object substitution pretense. Children could keep the ontology of bananas and phones straight not by decoupling but by not categorizing

the prop as a banana in the first place. Recall that the function of the indexing system is to track identity, which does not require attaching a conceptual description to the symbol. If the prop is not categorized as a banana to begin with, children will not come to believe that bananas are phones.

At the token level, though, children may come to falsely believe that **this prop** is a phone, even if they do not represent it as a banana. But is it the case that children need a representation with the content "this is a phone" when pretending that the banana is a phone? In language, it is evident that words are not their referents, so nobody assumes that there is a quarantined representation with the content "'banana' = banana". In graphs, this analysis does not make much sense either. Humans do not suppose that the diamond *is* the mean of the distribution; it merely stands for it. Analogously, in pretending with props, humans use physically available objects to convey information about discourse referents. Therefore, the relations one sets up in pretense only hold between an object index and a discourse referent, and they need not involve the IS-A predicate at all. The assignment from the symbol objects to the discourse referent is an instance of STAND-FOR stipulation, not of kind membership predication. As a result, the proposition "this banana is a phone" will never be entertained.

Moreover, if pretending that the banana is a phone involves a decoupled representation with the content "the banana is a phone", whatever predicates are ascribed to the phone will be attributed to an object that is a banana as well. But it does not make sense for a banana to be used to (pretend) call anybody. That predicate only applies to the phone that the banana currently stands for.

What about the discourse referent layer, considered separately from the symbols? The symbol–referent links are orthogonal to the way in which the discourse referent layer relates to the world. Humans can use props for depicting specific events (how Brutus killed Caesar), generic states of affairs (how corkscrews work), or fictional content (how Batman killed Joker, how dragons look like). If the content holds only in a fictional world, quarantining might start to play a role. But in this case, there is no need to come up with pretense-specific quarantining because the same problem arises for language, which can be used to talk about anything, factual or fictional. Therefore, the quarantining in pretense should be the same as for any communicative act. The parallels between pretend play and communication have been noted before (e.g., Friedman, 2013; Leslie & Happé, 1989), but the cognitive implications of the parallelism have not been fully spelled out, to the best of my knowledge.

On the other hand, the claim that decoupling is needed as an explicit operation rests on the tacit assumption that the default propositional attitude to mental representation is BELIEF and that the contents extracted from external representations must be quarantined lest they contaminate children's beliefs with false contents. But, as already noted, it could well be that infants do not automatically enter a BELIEF relation to the mental representations they create based on the external representations they interpret. If the contents of external representations are not in danger of being encoded as beliefs, there is no need for an additional decoupling operation.

In any case, one must not lump together the structure of pretense (X ← Y) with the ontological status of Y, failing to discriminate format and content. If these are, in fact, two separate issues, children should be able to learn from pretense despite the alleged need to quarantine. While there is not much data regarding toddlers' learning from pretense, recent studies have shown that preschoolers learn generic information from pretense (Baer & Friedman, 2016; Hopkins et al., 2015; Sutherland & Friedman, 2012). In these studies, 3–5-year-old children were taught new facts about a known category (that dogs are afraid of raccoons) or about a new category (that "nerps" like apples) with props. Children extended what they learned to new exemplars but not to the props used during learning. In addition, if a speaker who did not witness the pretense stipulation asked for the new category, they did not even consider the props as good candidates to offer to the new speaker, indicating that symbol–referent assignments are local (and thus automatically quarantined). When children were encouraged by Experimenter 1 to pretend that screwdrivers are "sprocks", they were less likely to give a screwdriver to Experimenter 2 when she asked them for a "sprock" (Hopkins et al., 2015). Preschoolers can thus learn new information about the world from object substitution pretense, as expected if pretend play is a subspecies of human communication.

In sum, the consensus surrounding the need to quarantine the mental representations involved in pretend play relies on the mistaken assumption that IS-A relations are the only possible way to link props to referents[8]. But if children have access to STAND-FOR relations too, the need for decoupling disappears—at

[8] This assumption may be prevalent because of a blind spot that sometimes surfaces in everyday usage too, as language often fails to mark STAND-FOR relations explicitly. For instance, the US anti-narcotics campaign poster showed a fried egg next to the slogan "This is your brain on drugs"—the intended reading was most likely not the literal one. In **Chapter 3**, I will present evidence that infants can interpret such predicative expression as stipulating STAND-FOR relations.

least when it comes to the format of object substitution pretense. When it comes to content, things become murkier. On the one hand, positing indiscriminate quarantining and decoupling leads to wrong predictions, as preschoolers should not be able to learn about Y based on the quarantined assignment X ← Y. Taking a step back from pretense, consider once again Figure 1.5, where the colored dots do not pick out any particular football players, yet the representation as a whole is used to teach a piece of real-world knowledge: what offside is in European football. Thus, the relation between what a representation depicts and what it is rational to learn from it is not as straightforward as to allow us to discern what exactly should be quarantined. On the other hand, quarantining and decoupling may not be needed as additional operations, as representations may not be interpreted in relation to the real world by default. In the following subsection, I offer an interpretation of children's seeming failure to understand pictures, videos, and scale models along these very lines and broaden the discussion by including other representational media.

1.5.2. The Dual Representation Account

Infants and toddlers, including those who already engage in pretend play, are often not credited with understanding nonlinguistic representations (e.g., DeLoache, 2004; Perner, 1991; Spelke, 2022). In this section, I will go through several reasons which underlie this skepticism and argue that they are not well-founded. I will revisit the data on how infants and young toddlers deal with visual representations in the lab as well as several alternative explanations that have been put forth to account for the discrepancy between interpreting and using representations.

As noted in Section 1.1, research on visual representations has focused on representations that stand in a one-to-one relation with things in the world (DeLoache, 1987, 1991, 2004; Tomasello et al., 1999). Empirical research revealed that 2-year-olds are at chance in object retrieval tasks if they are shown where the object is via pictures, video, or scale models (DeLoache, 1991; Troseth & DeLoache, 1998). In these studies, children must find a toy hidden by the experimenter in a room with which they had previously been familiarized. In picture studies, the experimenter shows children a photograph of the room, points to the corresponding location (e.g., to the chair), and tells them that she hid the toy there (DeLoache, 1991). In scale model studies, the experimenter does the same on a miniature version of the room (DeLoache, 1987). In video studies, children

watch the entire hiding event on TV (Troseth & DeLoache, 1998). Since it is not before their third birthday that children reliably pass all three tasks, DeLoache (2004) concluded that grasping and exploiting external representations undergoes a protracted development because children have to overcome a **dual representation** problem (see DeLoache, 2004, for a review): a nonlinguistic representation represents an object while being an object itself. This would make it harder for young children to perform adequately in these tasks because they do not have the executive resources to simultaneously represent a symbol as an object and as its referent.

I think this line of explanation is plausible but not for the age group that DeLoache had in mind. If attentional resources are limited in the way DeLoache suggests and if it takes time to allocate them properly (by learning or by maturation), testing infants' understanding of symbolic objects may underestimate their competence if such stimuli trigger other cognitive subsystems, such as the core object system (Spelke, 1990, 2022). Above 18 months, however, toddlers do not encounter any dual representation problem when engaging in pretend play, even though the representational requirements are the same: they deal with objects that stand for something else while being objects themselves. Thus, their failure at two years to retrieve an object from information received via an external representation cannot be attributed to a problem they have already overcome. In addition, children are aware of the dual nature of representational objects earlier than DeLoache's studies suggest, as they can reason about both symbol and referent when pragmatic context differentially highlights one of the two aspects of a representation (Preissler & Bloom, 2007). For instance, when 2-year-olds are shown a line drawing of a new object and told either "This is a dax" or "My brother keeps this in the wallet", they assume the former refers to the depicted object, while the latter refers to the picture.

One feature distinguishing DeLoache's tasks from object substitution pretense is their relation to the world. All her tasks require children to link the information they obtain via pictures, videos, or scale models to a state of affairs that holds in the here and now. However, as already noted, this does not have to be the only way representations are interpreted—not even the default. If representations of states of affairs involving particular objects are only a subclass of external representations, one cannot generalize toddlers' failures from a subset to the entire class. It is thus possible that understanding representations of particular objects and events requires an additional step of linking the discourse

referent layer to actual objects and events, which may go beyond 2-year-olds' cognitive repertoire. This can occur because children need to learn that external representations can be used to inform about a particular state of affairs, because there is an additional relation that needs to be computed (discourse referents–world), or both.

Counterintuitive though this may be, consider a scenario in which a child pretending that a banana is a phone is asked to go find the phone. She will be dumbfounded, as she has not assumed that there is a phone out there that the banana refers to. If this is the default interpretation, it might take children time to learn that representations can also be linked to actual states of affairs. Alternatively, the additional step of linking a representation to the real world may tax the child's cognitive resources. While I favor the first option, the second possibility is supported by the prevalence of perseveration errors in DeLoache's task. Suddendorf (2003) replicated 2-year-olds' failure across several trials in the retrieval task found by DeLoache but noticed that they retrieved the object above chance on the first trial. Hypothesizing that performance on later trials may be impeded because children go back to the location from the first trial, Suddendorf modified the paradigm and introduced separate target rooms for each trial. Consistent with the hypothesis, 2-year-olds were above chance in their average responses when the target room changed on each trial. But whichever cause underlies children's failures in DeLoache's tasks, neither implies a failure to understand representations.

1.5.3. The Conceptual Deficit Account

Alternatively, one could reject the claim that the phenomena under discussion involve representational understanding, from object substitution pretense to DeLoache's object retrieval tasks. This is the path Perner (1991) takes. He denies that pretend play should be treated as evidence of symbolic understanding. For him, one cannot attribute a cognitive system with understanding representations unless the representational relation between the medium and the content is itself represented. As young children do not possess this capacity yet, Perner opts for a behavioral account. For him, infants do not understand the banana as a symbol for a phone but create a separate representation with the content "the banana is a phone" and act **as if** the decoupled representation were accurate. I agree with Perner that the representational relation itself is not represented in early childhood, but this is not the litmus test for attributing a cognitive capacity

to an organism—depth perception does not require an explicit representation of depth. Analogously, STAND-FOR relations may be embedded in humans' cognitive architecture without being explicitly represented by a standalone concept. Thus, singling out the behavioral **as-if** account as the only alternative to representational understanding poses a false dilemma. The **explicit understanding** account and the **as-if** account are not the only possible options.

Moreover, the behavioral account has several drawbacks. The first two concern wrong predictions for pretense and have been rightly noted by Friedman and Leslie (2007). First, the account creates a double-edged sword for itself regarding pretense recognition. On the one hand, children driven by the perceptual similarity between actions will have difficulty distinguishing actual from **as-if** behavior. One might be tempted to invoke ostensive manner cues (e.g., smiling or exaggerated motion) as a guide to the distinction, but the behavioral account does not have this option. Manner cues will necessarily match the behavior to a **lesser** degree (e.g., smiling when pretending that the banana is a phone is not necessarily what one does when holding an actual phone), so pretense recognition should be hindered in these situations instead of aided. Second, pretense based on sound effects should not occur. A child observing an adult moving a pencil on the table while saying "vroom, vroom" could not interpret the pencil as a car because this is not what one would do if the pencil were a car.

The remaining two problems stem from not linking pretend play to the larger class of external representations of which pretend play should be considered a subspecies. While the **as-if** account may explain a subset of pretend play scenarios, it stops making sense once other representational media are considered. Suppose the legend of a graph stipulates that the circle on the figure represents the mean of the depicted distribution. In that case, there is no sensible way to construe an interpretation under the **as-if** account. What would it even mean to behave as if the circle were the mean? Finally, I suspect an analogous developmental procession also holds for language. Interpreting words as symbols and sentences as representations precedes metalinguistic awareness, yet there is no reason to postulate an analogous behavioral **as-if** account for early language understanding[9].

[9] I do not want to overstate the parallelism to language. STAND-FOR relations may not be needed in word understanding because words cannot be tracked spatially. Symbol objects, on the other hand, must be tracked so that the predicates can be applied to the right arguments.

Another argument pursued by Perner (1991) rests on toddlers' seeming failure to understand that a single representation can have multiple meanings and that representations can misrepresent. Regarding the multiple-meanings issue, Perner's claim is ambiguous between two different readings, depending on whether time is taken into account. On the first reading, infants should not be able to assign multiple meanings to a single symbol across discourses but they are: two-year-olds are happy to use props as stand-ins for different entities across discourses (Harris & Kavanaugh, 1993; Wyman et al., 2009).

On the second reading, which is probably closer to what Perner had in mind, infants should understand not only that a representation means something but also that it could have meant something else. Why this should be the case is far from clear. If communicators convey their intended meaning well, representations turn out unambiguous and thus have a single meaning, which the audience can retrieve. Whenever explicit stipulation is involved (e.g., in pretend play or in graph legends), one might not entertain multiple meanings because the inferential work required for retrieving the meaning would considerably reduce. In the case of ambiguous symbols (e.g., a drawing that looks both like a lollipop and a balloon), multiple meanings may have to be entertained **before** bringing the discourse referent under a conceptual description. In those cases, younger children may fail to consider multiple meanings for unrelated reasons. Children at this age seem to sample one answer from a distribution instead of simultaneously entertaining multiple possibilities (Leahy & Carey, 2020). But even in such cases, the idea that multiple meanings should be considered seems misguided, as the interpretive process must eventually converge on a single meaning.

As for misrepresentation, it may well be true that children realize only later that representations can misrepresent. This, however, might occur not because they do not understand representations but because the discourse referent layer (i.e., the content of the representation) is not linked to any real-world state of affairs, which would allow verification. At this point, representations cannot misrepresent by definition.

1.5.4. The Linguistic Account

Yet another possibility is to accept that pretend play, drawings, and scale models require representational understanding but to reject the notion that the underlying capacity is a primitive of the cognitive system. Rakoczy et al. (2005)

start by distinguishing referring from non-referring symbols and acknowledge that pretend play requires a proto-symbolic capacity—not a fully symbolic one—because the referents of props in pretense are not related to the world and because the symbolic operation that toddlers do when pretending is projecting a kind onto the prop. If semantic completeness is considered necessary for ascribing a fully symbolic capacity, this reduces to a terminological difference of little consequence. However, there are problems with the claim that pretend play amounts to projecting a kind property onto an object prop (e.g., in pretense, a banana does not represent a phone, but rather the generic phone-ness property). In some cases, this may be enough. If somebody tells you, "This is a great device", while holding up their newest smartphone, you know that the predicated property is not restricted to that particular device but to the generic model that the device instantiates. In such cases, it has been argued that objects are symbols of their kinds (Csibra & Shamsudheen, 2015). However, problems will crop up when two different symbols stand for the same type of thing, in which case the two symbols will have the same meaning. In other words, types are not suited for the individuation required by STAND-FOR relations; only tokens are (Brody, 2020). In addition, 2-year-olds understand predicates that can only be applied at the token level, suggesting that they are beyond mere property projection. When pretending that a yellow block is a banana that has been fed to a toy gorilla (Harris & Kavanaugh, 1993), infants refrain from offering the same yellow block to the gorilla a second time. This would not be possible if the yellow block stands for the kind, as the property of being eaten can only be ascribed to a token of the kind. In sum, one of the reasons behind qualifying young children's symbolic capacities as "proto" needs to be dropped.

Rakoczy et al. (2005) go on to offer a cultural learning account for understanding representations built around shared enculturation, intentionality, and natural language acquisition. While enculturation may play a role in dictating which types of stimuli are to be used and interpreted as representations, I do not see how enculturation could create a cognitive architecture from scratch. I will thus focus on the cognitive preconditions for interpreting objects as symbols that Rakoczy et al. (2005) put forth.

On the one hand, Rakoczy et al. (2005) argue that understanding intentions must precede understanding symbols because symbolic objects have meaning only by virtue of their user's intentions. While a thorough examination of the connection between symbol understanding and intention attribution goes be-

yond the scope of the present chapter, I would like to note two points. First, Rakoczy et al.'s observation is probably true as a metaphysical statement. However, metaphysics and psychology need not overlap, so the move from metaphysics to psychology cannot be made without further argument. While it is true that mass is a necessary condition for objecthood (as intentions may be for symbols), it does not follow that mass needs to be represented when reasoning about mid-sized objects, as indeed it is not (Spelke, 1990, 2022). Second, a self-consistent psychological account of interpreting communication as a type of action without resorting to intention ascription is both possible and developmentally plausible (Mussavifard, 2023). If interpreting external symbols only requires establishing links between an object layer and a discourse referent layer (**Section 1.2**), attributing intentions to the communicator will be superfluous, at least on some occasions. What will matter is learning which stimuli in the environment are to be interpreted this way, for which I expect ostensive signals and adult scaffolding to play an important role (Csibra, 2010; Lillard & Witherington, 2004).

On the other, Rakoczy et al. consider language a precondition for understanding objects as symbols. The central argument hinges on the fact that symbols must always be brought under some description when interpreted. While this fits nicely with the notion that discourse referents are headed by descriptions, the conclusion about the necessity of natural language is valid only if one denies preverbal infants the ability to generate such descriptions some other way. And there are at least two ways that bypass natural language. First, preverbal infants might possess a language of thought—an internal, open-ended, and language-like system that combines primitive units into arbitrarily complex expressions (Fodor, 1975). An innate language of thought would render natural language unnecessary by generating the conceptual descriptions itself. Second, if concepts in core systems could be fed as inputs to the tokening function, this would be enough for the referents to be brought under a description (e.g., AGENT, OBJECT) even in the absence of a language of thought[10].

An additional argument against the notion that linguistic input is required comes from studies of deaf children born to hearing parents who are not exposed to natural language in early development. However, these children invent their own gestures, which map neatly onto the distinctions in **Section 1.2.5**: they

[10] Linguistic descriptions (e.g., "suppose this is my car") may be required to connect symbols to real-world particulars, especially when the referent is not present in the scene.

invent gestures for discourse referents and for predicates that apply to the discourse referent arguments (Goldin-Meadow, 2005).

Finally, Rakoczy et al. (2005) draw on training studies to argue that language plays a crucial role in the development of pretense. In one study (Rakoczy et al., 2006), children exposed to explicit discourse for pretense (e.g., "I pretend that this stone is an apple, but really it is a stone") outperformed a control group exposed to pretense without explicit discourse. This evidence is used to argue that language also plays a causal role outside the lab. But if "causal" means "necessary", no such conclusion follows: words facilitate category formation (Waxman & Markow, 1995) without being necessary for category formation.

Rakoczy et al. (2005) are not alone in assuming that understanding objects as symbols rests on natural language. Spelke (2022, Chapter 10) also opposes the idea that infants interpret nonlinguistic stimuli as representational, citing infants' seemingly non-symbolic behavior toward pictures as evidence. I have noted the difficulty of interpreting these data (**Section 1.4.1**). Nevertheless, even if infants do not interpret pictures as symbols initially, the conclusion that the symbolic nature of pictures must be derived from language cannot be valid. By this token, if very young infants do not expect words to be symbolic, one should conclude that they derive this from language—not a promising avenue.

Given that infants seem to be quite open to the types of entities that carry meaning (e.g., spoken language, signed gestures), an equally plausible story is that infants have a broad symbol concept but need to figure out what classes of stimuli around them are used as symbols, which should take time. Finally, symbol objects may mask infants' competence with symbols because, unlike words, objects trigger other core domains. But above all, I fail to see any plausible story about how infants would draw on the symbolic character of language to understand the symbolic function of other kinds of stimuli.

1.5.5. The Flagging Account

The proposal closest in form to the one I have defended is Harris and Kavanaugh's (1993) flagging model of pretense. According to them, children engaging in object substitution pretense have two operations at their disposal. First, they can attach a flag to a mental representation of a prop to keep track of the make-believe identities (e.g., this yellow brick = make-believe banana). This corresponds to the assignment and tokening operations in the architecture outlined

in Section 1.2. Second, they can edit these flags to keep track of pretend action consequences on the action participants (e.g., this make-believe banana has been eaten). This corresponds to the predicate application procedure, which takes the actions performed on a given symbol and turns them into properties ascribed to the discourse referent.

In addition, Harris and Kavanaugh explicitly linked pretense to story comprehension, which the introduction of discourse referents in my account also implies. However, the flagging model has two shortcomings. On the one hand, the flagging operations were specifically introduced to account for object substitution pretense, not for external representations writ large. On the other, Harris and Kavanaugh (1993) failed to notice that the relevant relation is not a quarantined identity relation (e.g., brick = banana) but a representational one (e.g., brick STANDS FOR banana). Thus, the flagging model introduces mechanisms that do not generalize to other communicative media, thereby overfitting the data to pretend play. Moreover, Harris (2000) argued that pretend play is conducive to imagination, simulation, and counterfactual thought—capacities that need not be exercised in pretense, at least in setting up the relevant symbol–referent relations. Indeed, the creativity of children's pretend play in production is quite limited initially, as they mostly imitate what they see adults doing (Adair & Carruthers, 2022; Harris, 2021; Striano et al., 2001).

1.5.6. Summary

On the one hand, the accounts put forth specifically for pretense fail to acknowledge how similar pretend play is to other representational media (Harris, 2000; Leslie, 1987). Because of this, these accounts propose mechanisms such as decoupling, which become superfluous once recognizing that pretend play should be aligned with drawings, animations, and the like. On the other, the accounts that attempt to explain representational understanding writ large suffer from not taking pretend play seriously enough (DeLoache, 2004; Perner, 1991) or from glossing it as a derived competence based on theory of mind or language (Rakoczy et al., 2005; Spelke, 2022). Based on the current data and theoretical considerations, there is no good reason to conclude that understanding nonlinguistic representations is a late achievement in human childhood, nor that it depends on other cognitive faculties. Instead, the capacity emerges earlier than is usually assumed and covers a broader range of phenomena than is often acknowledged.

1.6. Conclusion

The ability to interpret STAND-FOR relations between object symbols and entities under discussion builds on two representational layers, **object indexes** and **discourse referents**, only one of which concerns objects in the here and now—the object indexing layer. On top of the two layers, two simple functions are at work: (i) **tokening**, which tags discourse referents with descriptions generated by the conceptual system; and (ii) **assignment**, which creates local relations between the symbol objects and the discourse referents. This cognitive structure is available early in development, at most by age two, when toddlers' pretend play exhibits all these architectural components.

Carving the explanandum this way is parsimonious, as it brings several superficially different phenomena under a single umbrella (drawings, diagrams, animation, movies, graphs, assembly instructions, and memes). An explanation that posits a specialized cognitive architecture provides a more robust account of pretend play by allowing a distinction between its format and contents. As a result, quarantining is no longer required. Because the system uses a pointer architecture, the need for quarantining stipulations is lifted away. Because the discourse referents are spatiotemporally undefined by default, the need to quarantine contents at the token level is also removed. And because pretend play is a representational activity, quarantining at the type level should not be expected either. Pretend play depictions can be used to teach and learn information about the world, just like language or any other representational system.

From a theoretical perspective, this line of investigation can inform several central discussions on human cognition, among which human uniqueness (Deacon, 1997), the language of thought (Dehaene et al., 2022), and the possibility of a species-specific core knowledge of symbols (Spelke, 2022; Spelke & Kinzler, 2007). From a methodological perspective, this research program is immediately relevant to psychology, as many experiments use stimuli that might be interpreted via STAND-FOR relations: animations, pictures, and puppet shows. If this is right, care must be exercised. If a cognitive subsystem different from ordinary perception underlies humans' interpretation of puppet shows, animations, and pictures, psychologists should explore the implications of using experimental stimuli that fall under its scope. This may have unintended effects both on participants' responses to such stimuli and on the scope of the theoretical models built on such data.

Chapter 2. For 19-Month-Olds, What Happens On-Screen Stays On-Screen

2.1. Introduction

Humans rely extensively on external representations in communication: drawings for objects, maps for space, calendars for time, and language for virtually anything they can think of. This capacity allows humans to transcend their immediate environment and gather information about distal states of affairs from proximal sources by decoupling incoming percepts, which necessarily reach the senses in the here and now, from the information carried by those percepts (Ittelson, 1996; Millikan, 2017).

Representations can carry information about at least two types of content. On the one hand, humans use representations to convey information about individuals in the world: the proper name Barack Obama refers to the former President of the United States; a map of London represents the spatial layout of the same city; a child's drawing of her teddy bear stands in for her favorite toy (while the toy itself does not represent any particular bear). On the other, the very same representational vehicles can be used to communicate about non-specific or fictional entities too: the proper name Batman picks out a well-known fictional character; a map of Hogwarts represents spatial relations of a place that can never be visited; a child's drawing of a house need not pick out any particular house outside her mind. By definition, these entities can be accessed via representations only.

Consider Heider and Simmel's (1944) short animations of geometrical shapes moving around. When adults are asked to describe such clips, they respond as if they talked about real agents. They attribute to them goals, desires, and intentions: the big triangle is chasing the small triangle, the circle wants to exit the enclosing, and the three shapes together form a love triangle (Heider & Simmel, 1944; Oatley & Yuill, 1985). Regardless, they are not fooled into believing that these shapes do form romantic bonds in front of them. Adults know these are not fully-fledged agents: they are not afraid that the big bully triangle will chase them, and they do not consider interacting with the shapes. In other

words, they know that the shapes and movement patterns stand for various agents and interactions among them, even if they do not expect these events to have actually happened (absent additional information). I take the link between a spatiotemporally trackable object (e.g., a triangle) and a conceptually defined entity (e.g., an agent) to be constitutive of representational relations.

Animations inspired by Heider and Simmel-like minimalist stimuli are routinely used in developmental research to tap into the emergence of conceptual understanding and, in many cases, there is substantive evidence that young infants interpret them in an adult-like manner: they attribute instrumental and social goals to simple shapes (Gergely et al., 1995; Kuhlmeier et al., 2003; Liu et al., 2017), they infer social relations from minimal interactions between these shapes (Powell & Spelke, 2013; Tatone et al., 2015), as well as ascribe mental states to them (Surian et al., 2007; Tauzin & Gergely, 2018). Undoubtedly, infants' inferences are prompted by the cues they would use to detect agents outside the lab, such as face-like features, self-propelled movement, and contingent reactivity (see Opfer & Gelman, 2011, for a review). But little is known about what infants make of these stimuli once the interpretive process has started.

Assuming that infants do not possess a concept of representation as a null hypothesis (e.g., Perner, 1991), how do they interpret animations? I delineate four hypotheses for infants' interpretation of animations as a broad stimulus category. First, infants might find the animations **fully opaque** (Hypothesis 1) because the information therein is too sparse to interpret (i.e., they cannot see a circle as an agent because agents are three-dimensional entities with whom one can interact contingently). On the opposite end, infants might be **naïve realists** with respect to animations and perceive them as spatiotemporally continuous with the surrounding environment (Hypothesis 2). If so, they should think that whatever is represented on-screen is happening here and now, in front of them. In between the two extremes, infants might believe that animations are temporally but not spatially continuous with the immediate environment. This will occur if infants know that screens have boundaries that objects cannot cross and perceive screens as (spatially self-contained) **aquaria** (Hypothesis 3). Finally, infants may interpret animations as **representations**, though not necessarily of particular objects or states of affairs (Hypothesis 4). This would imply that infants (i) can establish a link between a symbol object (e.g., a coherent pixel constellation on the screen) and a spatiotemporally undefined referent (i.e., a fictional object);

and (ii) dissociate symbols from referents in a way that shows they have learned how the representational medium works (here, on-screen 2D animations).

To test these hypotheses, I investigated whether 19-month-olds expect a ball falling on the screen to land in boxes below the screen (**Figure 2.1**). First, I obtained a baseline for infants' accuracy in tracking real balls falling (Experiment 1: Reality Baseline). Second, I tested whether infants expect on-screen falling animated balls to land in boxes below the screen (Experiment 2: Crossover). Third, I ran a control version, in which both the ball and the boxes were part of the animation, to ensure that infants can follow the trajectory of the animated ball when everything happens on the screen (Experiment 3: Animation). Finally, I tested whether infants think animations are tied to the screen on which they are presented (Experiment 4: Aquarium). The experiments were conducted with 19-month-olds because I targeted an age at which infants are known to fail DeLoache-type tasks but do not have problems understanding questions about objects' locations. Experiment 1 was piloted with 12- and 15-month-olds as well, but these infants mostly ignored the experimenter's questions.

Figure 2.1. Setup overview in Experiments 1–3: Reality Baseline, Crossover, Animation (left to right).

2.2. Experiment 1: Reality Baseline

2.2.1. Methods

TRANSPARENCY AND OPENNESS

The hypotheses and methods for all experiments were preregistered at the Open Science Framework (Experiments 1 and 2: https://osf.io/bwu9p; Experiment 3: https://osf.io/juerf/; Experiment 4: https://osf.io/gj5ys/). The experiments were approved by United Ethical Review Committee for Research in Psy-

chology (EPKEB) in Hungary, and informed consent was obtained from the participants' caregivers before the experimental session.

PARTICIPANTS

The final sample for Experiment 1 consisted of 16 typically developing 19-month-olds (M_{age} = 19 months 14 days, SD_{age} = 12.4 days). In the pilot run for Experiment 1, 10 out of 10 babies answered the question on the first trial correctly. Based on this data, I ran a power analysis for the binomial test against chance with an assumed effect size of 0.875. This effect is detected with 85% power with a sample size of 15, but 16 was chosen as the sample size for counterbalancing reasons. Based on this analysis, the sample sizes for Experiments 2 and 3 were selected to have equally sized samples across the three groups.

APPARATUS AND MATERIALS

Experiment 1 used a wooden seesaw (height = 40 cm; width = 60 cm) that could be inclined left and right (angle ≈ 25 degrees) via a 25-cm handle extending from the back of the seesaw, which allowed for the manipulation of the seesaw from behind a curtain (**Figure 2.1**, left). In addition, I used several identical-looking red sponge balls (radius = 2.5 cm) and two different-colored rectangular cardboard boxes (14 × 15 × 26 cm^3) as containers for the balls dropped from the seesaw. A secret compartment was added to each box, which ensured that the balls in the box were not accessible to infants even if they tried to open the boxes. (This ensured that infants could not receive any feedback at test.) The compartments were padded with soft cloth to remove the acoustic cues produced by the falling ball. In addition, two black rectangular cardboards were attached on top of the boxes in Experiments 1 and 2 to cover the edge of the screen in Experiment 2. The experimenter used two plush toys (a cat and a bird), which she hid in the boxes to familiarize infants with the task of pointing to object locations, and a canvas bag for storing the toys and balls throughout the procedure. Three ceiling-mounted video cameras recorded infants' behavior from different angles.

STIMULI

A small loudspeaker, placed behind the seesaw, played a 1-second jingle before each test trial to prompt infants to attend the ball-falling event. The experimenter talked to the participants using infant-directed intonation and following a pre-specified script (see **Procedure** below).

PROCEDURE

Infants were seated on their caregivers' laps on a chair, approximately 40 centimeters from the table on which the seesaw was placed. To familiarize the infants with the setup, the experimenter drew the infant's attention to the two boxes, showed them that they could be opened, and revealed their (empty) insides. She then took a plush toy cat from a canvas bag and allowed the infant to inspect the toy for 10 seconds. Meanwhile, she pushed the inner compartments backward to be able to drop the toy into the boxes. She then asked the infant to hand the toy, moved behind the seesaw, drew the infant's attention to herself ("[Infants' name,] look!"), and dropped the toy into one of the two boxes. She then slid the inner compartments back into place, pushed the boxes to the edge of the table, where the infant could reach them, and asked, "Where is it?". If the infant failed to respond within 3 seconds, she asked them, "Where is the cat?" two more times (at 10-second intervals) before retrieving the toy from the box herself. If infants picked the correct box, the experimenter congratulated the infant and took the toy from the box. If infants picked the wrong one, the experimenter showed them that the box they chose was empty and retrieved the toy from the box where it had been dropped. The next familiarization trial was identical, except that a toy bird replaced the cat and was dropped in the other box by the experimenter. When infants responded correctly for two trials in a row (out of a maximum of eight attempts), the experimenter put the toys away, pushed the boxes to the left and right of the seesaw, and pulled their inner compartments backward so the ball could fall from the seesaw into the boxes.

The test trials started with the experimenter drawing the infant's attention to the ball that had been placed in the middle of the seesaw before the session. While looking at the ball from behind the seesaw, she drew the infant's attention to the red ball in the middle of the seesaw ("[Name,] look at the ball!"). Immediately afterward, infants heard a 1-second jingle played by a loudspeaker behind the seesaw and saw the ball falling left or right into one of the boxes (the seesaw was manipulated from behind a curtain by a second experimenter). The experimenter did not follow the ball trajectory with her gaze but kept her eyes on the middle of the seesaw. After the ball fell, the seesaw was brought back into a horizontal position. The experimenter then pushed the boxes to the table's edge and asked the infant, "Where is it?". Like in familiarization, infants received two more prompts before the trial ended. Unlike in familiarization, infants were given neutral feedback by being congratulated regardless of their choice, and the ball

was not removed from the box. No infant tried to open the boxes after expressing their choice. A trial ended when infants chose a box or after the third prompt. Infants were then handed one of the two toys used in familiarization and encouraged to play with it. In the meantime, the experimenter set up the next trial by pulling the boxes backward and placing a new ball in the middle of the seesaw. Each infant received four test trials.

DESIGN

The box where the object was placed alternated across familiarization and test such that the toy in the last familiarization trial and the ball in the first test trial always ended up in opposite boxes (AB-ABBA). The side with which the AB-ABBA alternation started (left vs. right), the side of the boxes (orange box on the right and blue box on the left vs. orange box on the left vs. blue box on the right), and the experimenter's position during the test question (to the left vs. to the right of the seesaw) were counterbalanced.

CODING

There were two primary dependent measures: choice and correctness. Infants scored 1 for choice if they unambiguously reached, grasped, or pointed to one of the two boxes and 0 otherwise. Infants scored 1 for correctness if they chose the box on the same side as the falling event and 0 otherwise. One researcher recorded infants' responses during the testing session. Another researcher, naïve to the ball's location, double-coded them offline from the video recordings. Inter-rater reliability was very high (Cohen's κ = .86); inconsistencies were solved by discussion. Based on piloting data, I preregistered that I would also code how often children pointed to the center of the seesaw when not choosing one of the two boxes.

EXCLUSIONS

Based on preregistered criteria, four additional infants who did not make two consecutive correct choices across eight familiarization trials were excluded. One other infant was excluded due to experimenter error. Trials were also excluded if infants did not follow the ball trajectory with their gaze based on video recordings (n = 2, out of 64 trials). One additional trial was excluded due to experimenter error.

DATA ANALYSIS

Infants' raw scores for each trial (0 or 1 for choice, 0 or 1 for correctness if infants made a choice) were supplemented by two additional individual scores: the proportion of choices across valid trials and the proportion of correct responses across trials in which a choice has been made. Since the balls that fell into the boxes throughout the test trials were not removed from the boxes, infants' responses during later trials might be influenced by the fact that balls kept piling up in both boxes. Therefore, I also preregistered and ran a separate analysis for the first trial. All analyses were conducted in R 4.2.2 (R Core Team, 2022).

2.2.2. Results and Discussion

As expected, infants were able and motivated to solve the task. Most of them provided at least one response (87.5%, 14 out of 16 participants), and they did so in 72.1% of the trials (44 out of 61). When they chose, their responses were correct 83.3% of the time (median = 100%, Wilcoxon signed rank, V = 82.5, p = .007, r = 0.66), well above the 50% chance level. Ten of the 16 infants performed at ceiling, never choosing the wrong box. On the first trial, 11 of the 12 infants who chose a box were correct (binomial exact test, p = .006).

The purpose of Experiment 1 was twofold: (i) to make sure that infants can follow the trajectory of balls falling into boxes; and (ii) to get a quantitative baseline of this capacity when the entire setup consists of real objects. The results indicate that 19-month-olds can answer questions about displaced objects reliably and accurately. This allowed me to proceed to the study's central question and investigate whether infants would do the same when screen events appear to extend into the surrounding environment.

2.3. Experiment 2: Crossover

This experiment provided infants with the same visual information about the location of falling balls as Experiment 1, but the balls were animated and fell from a cartoon seesaw on a TV screen, while the target locations were the same real boxes as in Experiment 1.

2.3.1. Methods

Except where noted below, the methods were the same as in Experiment 1.

PARTICIPANTS

The final sample consisted of 16 typically developing 19-month-old infants (M_{age} = 19 months 7 days, SD_{age} = 13.9 days).

APPARATUS AND MATERIALS

The seesaw in Experiment 1 was replaced by an LCD TV screen (16:9, diagonal 110 cm) to play animations in which a ball on the screen fell either to the left or the right. The same boxes used in Experiment 1 were placed under the screen to create the illusion that the ball lands into the boxes (**Figure 2.1**, center).

STIMULI

The events from Experiment 1 were transposed in a 2D-animated format, using Adobe Animate CC 2018: a red ball (more precisely, a red circle) falling off a seesaw to the left or to the right. The dimensions of the animated ball and seesaw matched those of the real objects. To give the illusion that the animated ball fell into the box, the boxes were placed under the screen based on the ball's trajectory. Black sheets extending from the boxes were used to cover the screen bezels to make the endpoint of the ball falling event ambiguous.

PROCEDURE

The warmup phase was identical to Experiment 1: the experimenter dropped a (real) toy into one of the two boxes and asked the infant where the toy was. Test trials followed the same logic as those in Experiment 1. While behind the screen, the experimenter drew the infant's attention to the red ball on the screen ("[Name,] look at the ball!"), which then rolled to the left or the right of the seesaw. The experimenter then pushed the boxes toward the infant and asked them, "Where is it?". The trial ended if the infant chose one of the two boxes or if they did not respond to the third prompt.

CODING

The inter-rater reliability between the online and offline coders was substantial (Cohen's κ = .76); inconsistencies were solved by discussion.

EXCLUSIONS

Four additional infants, who did not make two correct choices in a row during familiarization, were excluded. Four trials were excluded because infants did not look at the falling event (n = 2) or due to experimenter error (n = 2).

2.3.2. Results and Discussion

Unlike in Experiment 1, only 50% of the infants chose a box at least once at test (8 out of 16 participants). Out of the 60 valid trials included in the final analysis, infants picked a box in 18 trials only (30%). This was not because infants were less motivated to provide an answer to the question in this version of the task. In 24 out of the remaining 42 trials (57%), infants pointed to the screen when asked where the ball was. When infants did make a choice, they chose the box that was on the side of the falling event 45.8% of the time (median = 0.5, Wilcoxon signed rank, $V = 3.5$, $p = .71$, $r = 0.20$). On the first trial, 4 of the 8 infants who chose a box were correct (binomial exact test, $p = 1$).

In Experiment 2, infants behaved in a way that is inconsistent with the belief that animations are spatiotemporally continuous with reality. In contrast to Experiment 1, they were less likely to choose a box when asked where the ball was and often preferred to point to the screen. When they did respond, however, they picked a box at random instead of basing their answers on the side where they saw the ball falling.

2.4. Experiment 3: Animation

It is possible that infants did not get the intended referent of the question "Where is the ball?" in Experiment 2 because they did not see the red animated circle as a potential candidate for "the ball", and they may have pointed to the screen to request another animation. Experiment 3 moved the two boxes into the animated world to test this alternative explanation. If infants understand the question as intended, they should be able to point to the correct location when asked about the ball's whereabouts.

2.4.1. Methods

Except where noted below, the methods were the same as in Experiment 1.

PARTICIPANTS

The final sample consisted of 16 typically developing 19-month-old infants (M_{age} = 19 months 3 days, SD_{age} = 12.8 days).

PROCEDURE

The procedure was the same as in Experiment 2, except for the boxes, which were also part of the animation (**Figure 2.1**, right). The familiarization trials were identical to the ones in Experiments 1 and 2, except that the cardboard boxes were removed from the table once infants passed the familiarization phase with the two plush toys. Unlike in the first two experiments, the animated boxes were not brought closer to the infant after the test question was asked.

CODING

The inter-rater reliability between the online and offline coders was substantial (Cohen's κ = .80); inconsistencies were solved by discussion.

EXCLUSIONS

Eight additional infants were tested but not included because they did not provide two consecutive correct responses at familiarization (n = 5), because they did not look to the screen in any trial (n = 1), and because of experimenter error (n = 2). Two trials were excluded because the infant did not look at the screen during the falling event (n = 1) or because of experimenter error (n = 1).

2.4.2. Results and Discussion

Comparable to Experiment 1, 81.3% of infants chose a box at least once (13 out of 16 participants). Out of the 62 valid trials included in the final analysis, infants chose a box in 30 trials (48.4%). As for accuracy, infants chose the box that was on the same side of the animated ball far from chance levels: they pointed to the correct box in 93.6% of the trials in which they made a choice (median = 1, Wilcoxon signed rank, V = 78, p < .001, r = 0.86). On the first trial, 10 of the 11 infants who chose a box (out of 15: one participant's first trial was excluded) were correct (binomial exact test, p = .012).

While they made fewer choices compared to Experiment 1, infants overwhelmingly pointed to the box into which they saw the ball fall on the trials where they made a choice. This suggests that the random pattern of pointing in Experiment 2 was due neither to infants' inability to link the animated red circle to the intended referent of "the ball" nor to other differences between Experiments 1 and 2 (e.g., the fact that the experimenter could not herself see the ball because she was standing behind the TV screen in Experiment 2).

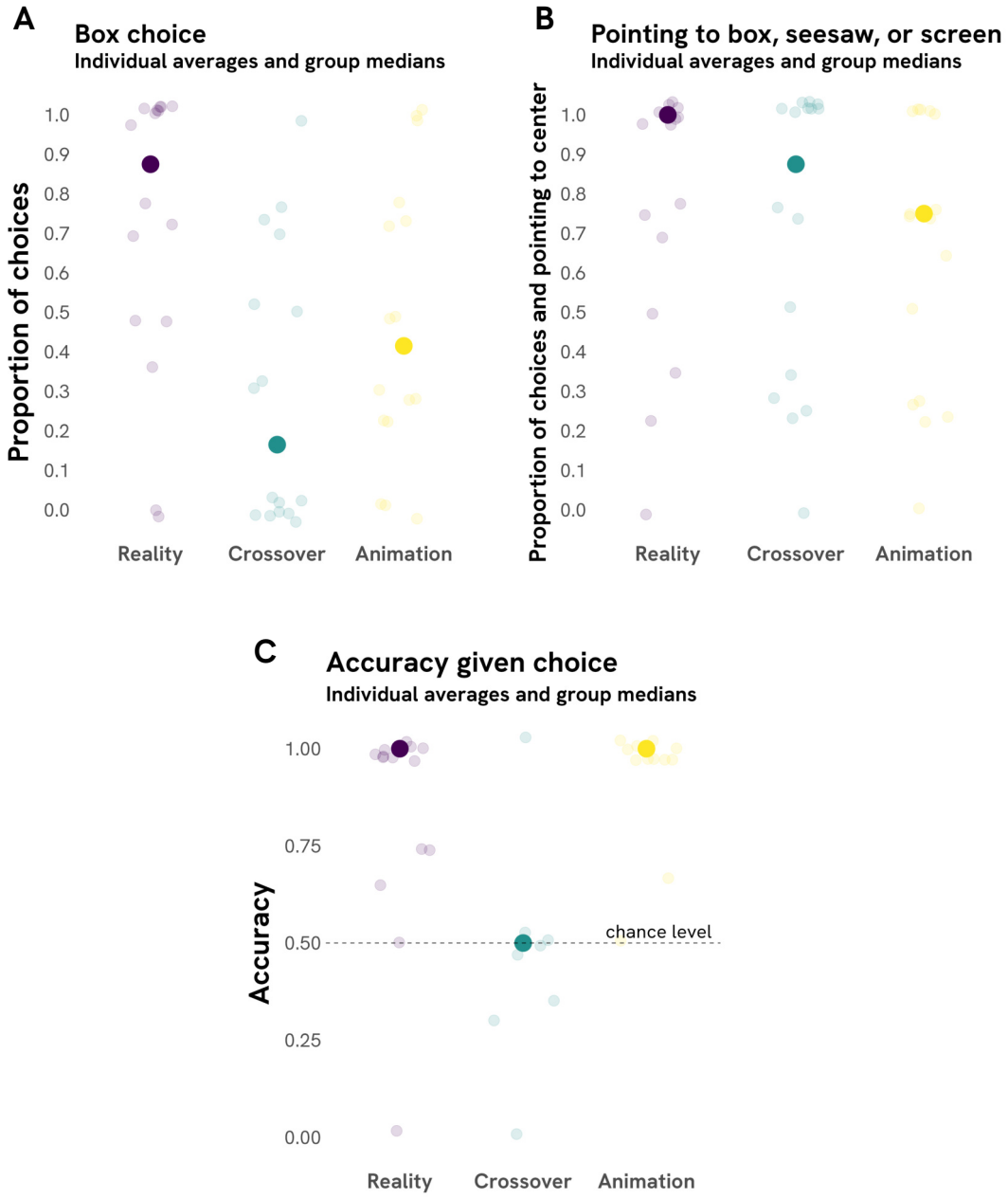

Figure 2.2. Results of Experiments 1–3. Transparent dots indicate individual proportions across the four trials; opaque dots represent group medians. **(A)** How often infants pointed to one of the two boxes in response to the test question. **(B)** How often infants pointed either to one of the two boxes or to the center of the seesaw/screen. **(C)** How often infants responded correctly in the trials in which they chose one of the two boxes.

2.5. Comparisons Across Experiments 1–3

2.5.1. Frequentist Analyses

The experiment-wise box choice rates are shown in Figure 2.2A. Nonparametric analyses revealed that the frequencies with which infants chose a box in the three experiments were unlikely to come from the same distribution (Kruskal-Wallis, $\chi^2(2) = 9.36$, $p = .009$). Planned pairwise comparisons with Holm's correction indicated that the contrast between Experiments 1 and 2 drove this difference (Dunn's test, $z = 3.05$, $p = .007$).

However, the frequency of responses to the question "Where is the ball?" did not differ across the three experiments. If infants' pointing to the center of the display is taken into account, the difference between response rates disappears (Kruskal-Wallis, $\chi^2(2) = 1.9$, $p = .386$). Infants in Experiment 2 chose to point to the screen instead of the two boxes, even though they were made salient by the experimenter pushing them toward the infant before asking them where the ball was (Figure 2.2B). This strengthens the interpretation that they did not think the animated ball could have landed in the boxes below the screen.

Like choice rates, accuracy rates across the experiments (Figure 2.2C) were unlikely to come from the same distribution (Kruskal-Wallis, $\chi^2(2) = 13.66$, $p = .001$). This difference was driven by Experiment 2, where infants were at chance when choosing between the two boxes (Experiment 1 vs. 2, Dunn's test, $z = 2.88$, $p = .008$; Experiment 2 vs. 3, Dunn's test, $z = 3.61$, $p < .001$). When infants chose a box in Experiments 1 and 3, they chose it based on the falling event they had just seen. By contrast, in Experiment 2, they completely disregarded the animated falling event and picked a box at random.

2.5.2. Bayesian Analysis

To model both choice and accuracy rates, I built a hierarchical Bayesian multinomial mixture model in STAN (Carpenter et al., 2017; Kruschke, 2014; McElreath, 2020), which models both measures simultaneously (Figure 2.3). Using infants' responses (no choice, correct choice, or incorrect choice), the model infers both (i) whether infants believed that falling balls ended up in the boxes; and (ii) whether their beliefs differed across experiments. I use $b_{\text{EXPERIMENT}}$ (ranging from 0 to 1) to denote infants' beliefs about ball location in each experiment. The prior on the three b-values is centered on 0.5 and skeptical of extreme values. I make three assumptions as to how beliefs and responses are linked. First,

I assume that infants are more likely to make a choice and to choose correctly if they believe that the ball is in one of the two boxes (indicated by the skewed priors on the left side of the tree). Second, I assume infants are equally likely to refrain or pick a box (at random) when they do not think the ball is in either of the two boxes (as shown by the balanced priors on the right side of the tree). Third, I assume that the b-parameters are sampled from the same underlying beta-distribution (parameterized by ω and κ) to avoid overfitting. The script for the analysis can be found on the Open Science Framework project page, accessible at https://osf.io/s83qn/files/osfstorage.

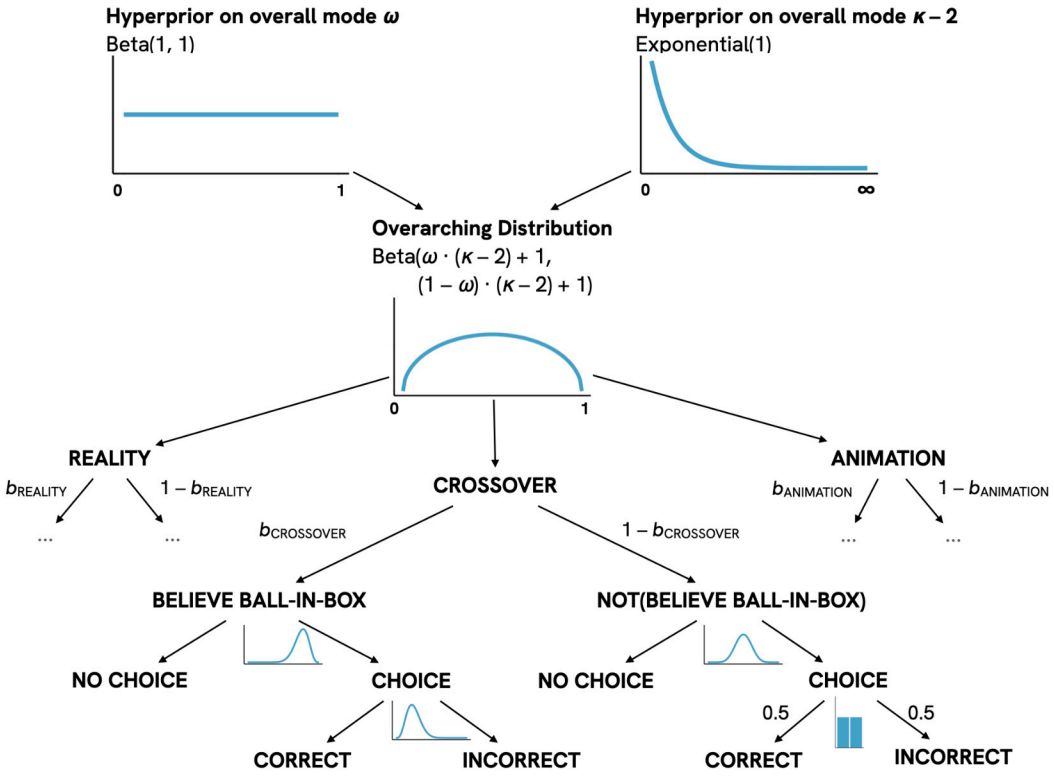

Figure 2.3. Schematic representation of the data-generating process assumed to underlie infants' choice and accuracy rates in Experiments 1–3. Infants' beliefs that the ball is in the box are generated from the same overarching distribution parameterized by ω and κ. In each of the three experiments and in each trial, infants can either choose a box or not and, if they do, they can choose it correctly or not. From the observed behavior, the tree can be inverted via Bayes' rule to obtain infants' beliefs in each of the three experiments.

Having constructed the data-generating model (from infants' beliefs to their responses), I use Bayes' rule and invert it to infer infants' beliefs from their responses. If infants always choose and choose correctly, they probably think that the ball is in the box; if infants make a choice only half the time and are at chance when choosing, they probably do not believe that the ball is in the box. Thus, large b-values (closer to 1) would indicate that infants believe there is a ball in the box into which they last saw it fall; conversely, small b-values (closer to 0) would indicate that they do not entertain this belief.

The posteriors on the overarching parameters $b_{\text{EXPERIMENT}}$ are displayed in Figure 2.4. For Experiment 1, b_{REALITY} peaks close to 1, mode = 0.85, 89% credible interval [0.56, 0.97], suggesting that infants relied on the ball falling event when answering the test question. Similarly, $b_{\text{ANIMATION}}$ also peaks toward the right end of the [0, 1] interval, but the estimate has higher uncertainty because infants made fewer choices than in Experiment 1, mode = 0.92, 89% credible interval [0.41, 0.97]. By contrast, $b_{\text{CROSSOVER}}$ shows the opposite trend toward 0, indicating that infants did not think that the animated ball ended up in the boxes, mode = 0.04, 89% credible interval [0.01, 0.31].

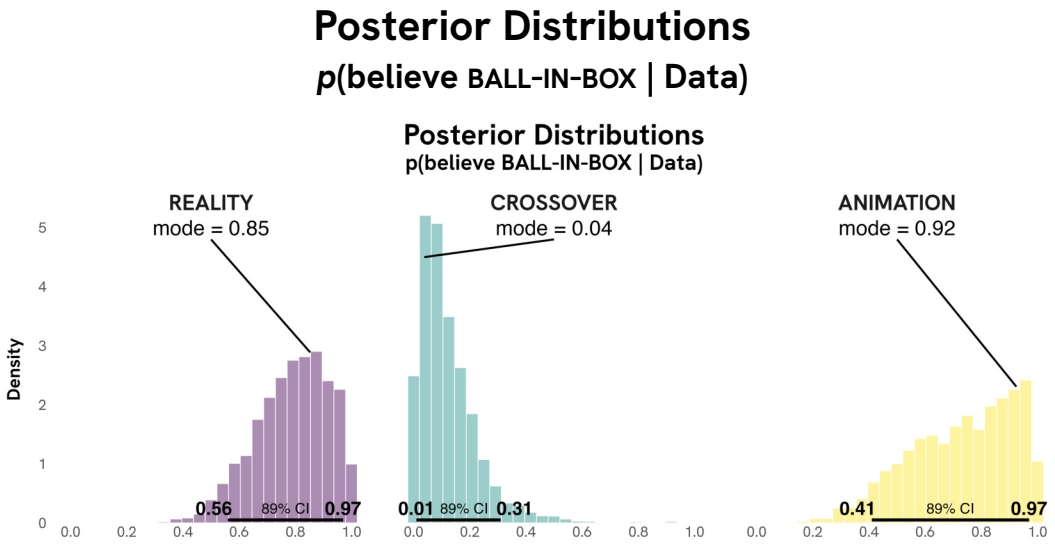

Figure 2.4. Posterior distributions for the belief parameter in Experiments 1–3. Bold horizontal lines above the *x*-axis give the 89% credible interval of the distributions.

2.5.3. Discussion of Experiments 1–3

The results obtained in Experiments 1 to 3 rule out two of the three hypotheses outlined in the **Introduction**. On the one hand, infants did not behave as the **naïve realism** account would predict (Hypothesis 2). When asked where the ball was in Experiment 2, they either pointed to the screen or chose one of the boxes at chance, indicating that they did not expect animated balls falling on-screen to end up in boxes below the screen. However, this was not because the animation was so **opaque** that it did not allow them to link the red circle on the screen to the noun phrase "the ball" (Hypothesis 1). Otherwise, they would have failed in Experiment 3, where everything was on the screen.

However, it remains an open question whether infants have just learned that screens are spatially disconnected from their surroundings while still believing that the events depicted on the screen are happening in the here and now, just like in an **aquarium** (Hypothesis 3). If this is the case, infants should not accept that an event displayed on one screen can move to a different screen—unlike adults, who can start watching a movie in the theater and end it on their laptops at home without losing track of narrative continuity. This potential explanation was tested in Experiment 4.

2.6. Experiment 4: Aquarium

Experiment 4 asked how infants would identify the protagonist of an animation when they get potentially conflicting information about its location. Infants were shown two animations on two screens placed side-by-side on a table. Each animation consisted of an animal (a bear and a rabbit, respectively) leaving its house and entering back in. The houses were identical, but the animation backgrounds were different. After making sure that infants learned which animal lived on which screen, the two backgrounds were surreptitiously swapped, and the experimenter asked infants about the animals' location again. Do infants individuate the protagonists by their physical locations (the house in the screen on which the animation was presented) or by their virtual locations (the house in the animation scene of which the protagonist was part)? If they opt for the virtual location, the **aquarium** hypothesis can be ruled out: screens are not merely spatially bounded physical containers for infants.

2.6.1. Methods

PARTICIPANTS

The final sample consisted of 32 typically developing 19-month-old infants (M_{age} = 19 months 17 days, SD_{age} = 7.6 days).

STIMULI

Experiment 4 used two 15-second animations featuring two protagonists, a rabbit and a bear (Figure 2.5). In each animation, the protagonist came out of its house, walked around, fetched a piece of fruit, then went back inside. Crucially, the backgrounds of the two animated worlds were chosen to contrast as much as possible, but the animals' houses were identical. In addition, there were also two 5-second backup animations, which showed the two animals exiting their house and entering back in (see **Procedure** below).

DESIGN

The experiment consisted of two between-subject conditions[1] and a single trial. The two conditions differed in whether the animation backgrounds were swapped (Swap condition) or not swapped (**No Swap** condition) between monitors from familiarization to test. A single trial was administered because subsequent trials would have been tainted by evidence (from the first trial) that animations can move from one screen to another.

APPARATUS

Two LCD monitors (16:9, diagonal 61 cm) were used to play the two animations. A VESA dual-mount arm held the monitors suspended above a table (Figure 2.5). To help infants keep track of the physical monitors, tapes of different colors were attached to the bezels of the two monitors. Two curtains were glued to the monitors to cover the monitors between familiarization and test. Because each monitor had its own curtain, infants could track the movement of both monitors individually.

[1] A slightly different version of the same study was preregistered at the Open Science Framework because I initially thought that running the Swap condition only would suffice to test the **aquarium** hypothesis (n = 32). After the preregistration, I realized that the results from the experimental condition (Swap condition) would not be interpretable without a control condition, so I decided to split the sample into two equal (n = 16) groups.

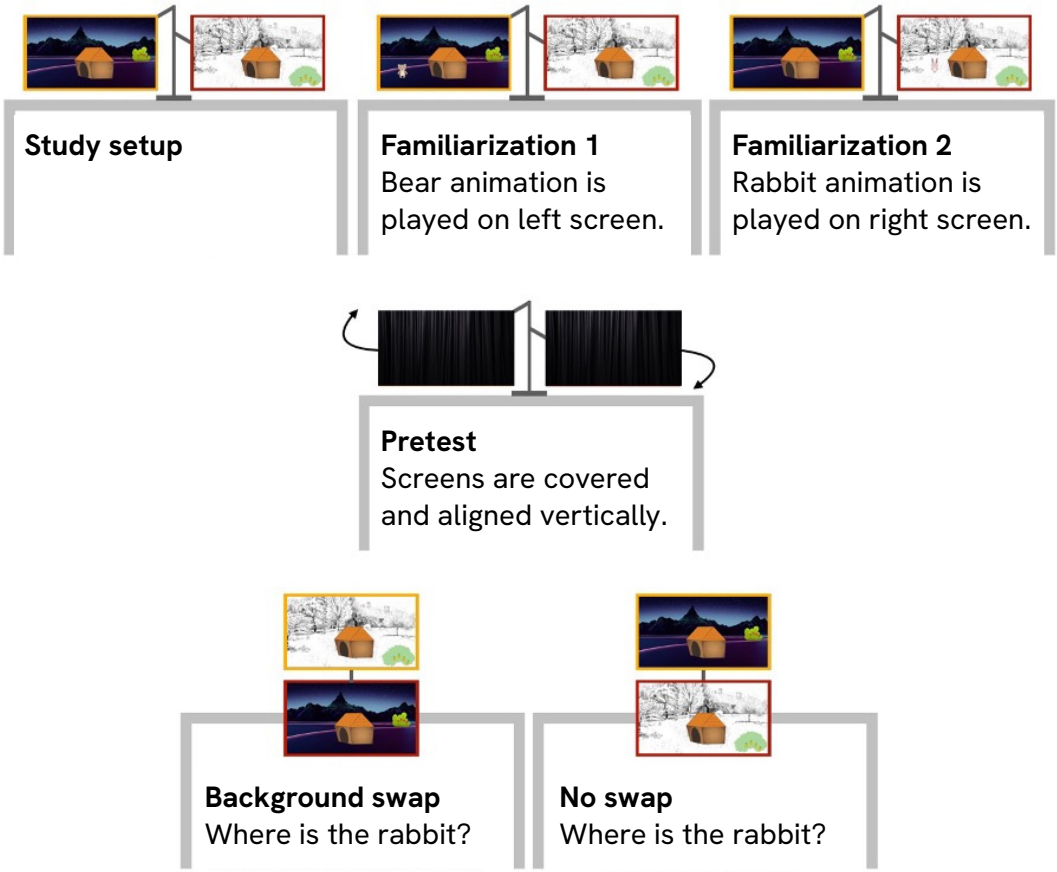

Figure 2.5. Schematic representation of Experiment 4.

PROCEDURE

Before the familiarization phase, infants were seated on their caregivers' laps on a chair, approximately 40 cm from the table on which the monitors were placed. Caregivers were instructed at the beginning of the session to close their eyes during the test phase. The experimenter moved next to the infant and (in infant-directed intonation) drew their attention to one of the two screens ("Oh, look, here is the bear's house. Let's see what's going to happen!"). Then corresponding animation started playing, and the experimenter narrated the events unfolding on the screen (e.g., "Wow, look a bear! The bear comes out of the house. And look, he's walking! Oh, and now the bear is collecting a raspberry and then he's going back! He's entering the house again!"). After the animation was over, the experimenter went behind the screen and asked infants where the animal was (e.g., "Where is the bear? Can you show me?"). If infants did not answer within 3 seconds, the experimenter repeated the question twice. If infants did not an-

swer or answered incorrectly, they were shown a 5-second clip showing the bear (rabbit) coming out of the house and going back in. The question was repeated, and the short clip was shown again if infants did not respond. If they answered by pointing to the screen on which they had just seen the animation, the experimenter congratulated them and repeated the same process with the second screen and animation.

After passing the second familiarization question, infants were asked about the first animal again to ensure they had stored both animals' locations. The familiarization phase was repeated once if they failed to answer the question correctly. If infants responded correctly, they were congratulated, and the test phase started.

The experimenter drew the curtains over the monitors and brought them from horizontal to vertical alignment (**Figure 2.5**, bottom row). This manipulation was meant to eliminate side and perseveration biases. During the rearrangement, the two monitors always remained visible so infants could track the individual screens through space. In the Swap condition, the two backgrounds were surreptitiously swapped while the screens were covered; in the No Swap condition, nothing changed. Once the monitors were vertically aligned, the experimenter unveiled them by pulling the curtain backward, moved next to the child, and asked about one of the two animals' whereabouts: "Look what's happening! Let's find the animals! Where is the bear (rabbit)?" If the infant did not provide a response within 3 seconds, the experimenter asked two more questions ("Can you show me the bear (rabbit)? In which house is the bear (rabbit)?"). Once infants pointed to one of the screens, they were asked the same question about the remaining animal. The answers to the second question were not analyzed because they were not independent of the answers to the first one. Still, they were included in the procedure to ensure that infants answered the location questions consistently (if they think the bear is on screen A, they should also think that the rabbit is on screen B). Otherwise, it would be unclear whether their pointing was related to the test question (e.g., it could mean "I want to see that animation again"). Infants with inconsistent answers were therefore excluded.

Which animation went on which screen (left vs. right), which animation was played first (bear vs. rabbit), the content of the test question ("Where is the bear?" vs. "Where is the rabbit?"), and the experimenter's position during the test question (right vs. left) were counterbalanced across participants in both conditions.

CODING

Responses (upper vs. lower screen choice) were recorded by one researcher during the testing session and double-coded from video by a second researcher naïve to the animals' locations. Inter-rater reliability was very high (Cohen's κ = .812); inconsistencies were solved by discussion.

EXCLUSIONS

There were two main criteria for inclusion in the final sample. First, infants had to provide three consecutive alternating answers during familiarization to make sure that they stored both animals' locations in memory before the screens were covered. Second, infants had to provide contrastive answers at test. If they pointed to one screen in response to the bear question, they had to point to the other screen in response to the rabbit question. Even though only the first answer entered the analysis, I wanted to ensure they answered the test question based on location and not something else. If they point to the same screen when asked about the two different animals, this might express a preference for one of the two animations instead of reflecting their beliefs about the animals' locations.

Despite almost no dropout during piloting, 28 infants had to be excluded from the final sample based on these preregistered criteria because they did not pass familiarization (n = 14: six in the Swap condition, eight in the No Swap condition), did not provide a contrastive answer at test (n = 10), or did not answer at all (n = 4). In addition, 13 infants were excluded due to experimenter and technical errors (n = 8), fussiness (n = 4), or parental interference (n = 1).

2.6.2. Results

Before data analysis, infants' up-down scores for each trial were recoded to represent infants' strategy at test. They received a score of 1 if they pointed to **the same physical screen** they pointed to at familiarization and 0 if they pointed to the other screen than the one chosen in familiarization. By this coding scheme and assuming that virtual location is the correct answer, 1 is the correct response in the No Swap condition, while 0 is the correct response in the Swap condition.

In the Swap condition, 12 out of 16 babies pointed to a **different** screen from the one they pointed to at familiarization—they went for the virtual loca-

tion instead of the physical one. In the No Swap condition, where there were no conflicting location cues, 13 out of 16 babies pointed to the **same** screen as during familiarization. The observed effect of condition was unlikely under the null hypothesis (Fisher's exact test, p = .004). When adding the responses of infants who were excluded because they had not provided a contrastive answer to the control question (n = 14), the effect of condition does not change: 14 of 20 participants chose the different screen in the Swap condition, and 17 of 21 participants chose the same screen in the No Swap condition (Fisher's exact test, p = .002).

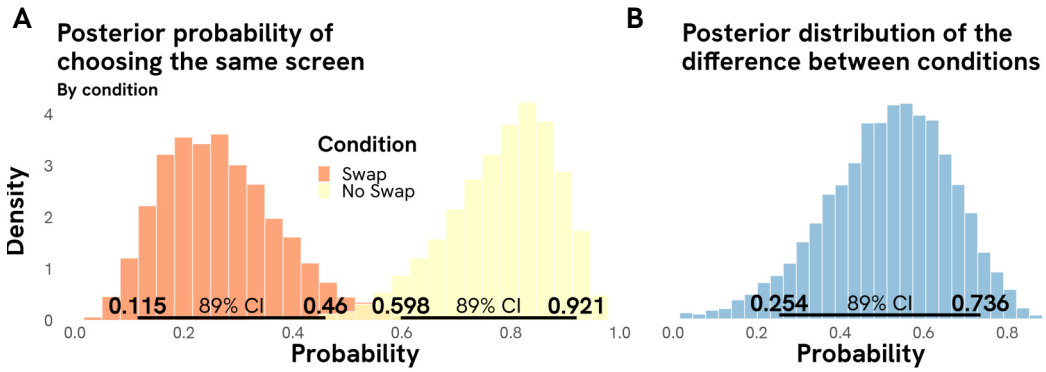

Figure 2.6. (A) Estimated probabilities of choosing the same screen as in familiarization, by condition. In the No Swap condition, infants chose the same screen as in familiarization. In the Swap condition, infants chose the other screen, indicating that they individuated the animated animals by background. **(B)** Posterior difference between the distributions in (A). Black horizontal lines above the *x*-axis give the 89% credible interval of the distributions.

I used a Bayesian logistic regression model to obtain the probability of choosing the same screen in each condition separately and the difference between the two conditions. The details of the model and the scripts to replicate the analyses can be found on the Open Science Framework project page (https://osf.io/s83qn/files/osfstorage). The posterior distributions of the probability of choosing the same screen in the Swap vs. No Swap condition are shown in **Figure 2.6A**. In the No Swap condition, the posterior mean for this parameter was .78, 89% credible interval [0.6, 0.92], while in the Swap condition, it was .27, 89% credible interval [0.11, 0.46]. **Figure 2.6B** depicts the posterior of differ-

ences between conditions as estimated by the model. The histogram indicates that infants' responses are influenced by whether the backgrounds are swapped (89% credible interval excludes 0 as a plausible value).

2.6.3. Discussion

The results of Experiment 4 rule out a potential explanation (Hypothesis 3) for the results in Experiments 1 to 3. According to this explanation, infants rejected the apparent screen–reality crossover in Experiment 2 because screens are containers with rigid boundaries that do not allow objects to pass through. This account predicts that infants should identify animated characters based on the screen on which they are presented. However, infants linked the two protagonists to their virtual environments, not physical locations, when the two possible locations were pitted against each other.

2.7. General Discussion

Investigating how infants interpret animated stimuli is relevant for both theoretical and methodological reasons. One way in which humans communicate is by using symbols to represent entities they want to communicate about. Symbols and the actions performed on them are used to create physical scenes through which events, relations, and properties of distal objects are depicted (Clark, 2016). Beyond animations, the same setup can be found in graphs, assembly instructions, and joint pretend play—representations where the visual and conceptual systems of the interlocutor are recruited for interpretation. The capacity to set up these links is central to gathering information about distal states of affairs from proximal sources, enabling humans to widen the range of things they can learn about without firsthand experience. Thus, the ability to grasp and exploit representations lies at the intersection of communication and learning, and therefore understanding how it develops can inform debates on both topics.

In addition, representations are relevant to methodology because they are pervasively used to elicit infants' and children's inferences (e.g., animations, puppet shows, games) in the lab. Moreover, experimental setups involving TV–reality crossovers are used in developmental research, under the assumption that infants are naïve realists (e.g., Kinzler et al., 2007; Lucca et al., 2018; Ma & Lillard, 2006). The experiments reported here brought this assumption to the

surface and provided a straightforward way of testing it. The results do not necessarily invalidate studies that use screen–reality crossovers in their designs but highlight the need to test any assumption underlying methodological decisions.

	FULL OPACITY	NAÏVE REALISM	AQUARIUM	REPRESENTATION	OBSERVED
EXPERIMENT 1: REALITY BASELINE	✓	✓	✓	✓	✓
EXPERIMENT 2: CROSSOVER	×	✓	×	×	×
EXPERIMENT 3: ANIMATION	×	✓	✓	✓	✓
EXPERIMENT 4: AQUARIUM	×	×	×	✓	✓

Table 2.1. An overview of the predictions made by the four different accounts for Experiments 1–4 and the observed results. Checkmarks represent above-chance performance (Experiments 1–3) or a difference between conditions (Experiment 4); crosses represent chance levels (Experiments 1–3) or no difference between conditions (Experiment 4). The observed results support the **Representation** Account.

I outlined four hypotheses in the Introduction (Table 2.1), three of which are incompatible with the results in Experiments 1–4. The **full opacity** account predicts that infants would not be able to understand animated falling events as such and would thus fail in both Experiments 2 and 3. But infants had no problems tracking the trajectory of animated balls within the confines of the screen (Experiment 3). If the **naïve realism** account were true, infants would represent animation and reality as a spatial continuum, and the first three experiments would have produced the same pattern of results. This is not what happened. When infants faced an animation that appeared to continue beyond the screen, they were not fooled into thinking that the boundary could be crossed (Experiment 2). When asked where the ball was, infants either picked a box at random or ignored the boxes and pointed to the screen instead. Finally, while the **aquarium** account can accommodate the results from the first three experiments, it cannot explain why infants identified animated characters by the background of the animation, as opposed to their physical location in Experiment 4.

While the results reported here do not provide direct evidence that infants at this age interpret animations as representations of (real or fictional) states of affairs, the findings are compatible with an early concept of **representation**. To recapitulate, I do not think representations are defined by reference to the world but by the STAND-FOR relation between a physical symbol—unitary pixel constellation on the screen, marks on paper, props—and a conceptually defined entity about which information is conveyed. This formulation renders the format of representation (X stands for Y) independent of the content (Y may or may not exist in the world). In a typical pretend play scenario, for instance, when 2-year-olds pretend that a wooden block is a carrot (Harris & Kavanaugh, 1993), they do not take the block to stand for a particular carrot in the world. Instead, they use their conceptual system to generate a new carrot token for the occasion. By contrast, in the tasks used by DeLoache and colleagues (reviewed in DeLoache, 2004), relying on this mechanism will not do, as the symbol object represents another particular object in the world, not merely a conceptually defined entity.

The dual representation explanation (DeLoache, 2004) cannot account for the contrast between infants' behavior in the present experiments (or their early proficiency with pretend play) and their failures in the tasks used by DeLoache and colleagues. The **dual representation** account attributes the failures to a deficiency in the representation of the symbol object (both an object and a stand-in for something else). However, animations and pretend play build on the same duality (both a 2D circle and a stand-in for a ball; both a block and a stand-in for a carrot), yet infants and young toddlers respond appropriately in these scenarios. I speculate that the nature of the referent underlies this difference instead. When the referent is not a particular object, infants set up the appropriate STAND-FOR relation between a physical symbol and a conceptually defined entity. When the referent is a particular object, they struggle with the tasks because they fail to make the additional link from the conceptually defined entity to the particular object they need to retrieve.

This observation can also be used as an argument against the possibility that infants interpreted the on-screen events not as representations but as events they perceived from a distance via the screen, similar to videos captured by surveillance cameras. If infants were capable of interpreting screens in this way, their performance in DeLoache's tasks should be much better.

If infants set up STAND-FOR relations between a visual object and a conceptually defined entity, their responses in these experiments would be naturally

accounted for. In Experiments 2 and 3, infants linked the definite noun phrase "the ball" to the on-screen red circle (without explicit instruction). They were then able to answer questions about the ball by tracking the trajectory of the red circle. Since animated objects do not exit screens, infants' responses diverged from the crossover to the fully animated setup. However, this was not merely due to the physical screen boundary, or else they would have rejected the possibility that the bear and rabbit swapped locations in Experiment 4. But since animated bears and rabbits are not actual agents, infants did not individuate them based on physical location. Instead, they tracked the visual cues to the symbols presented in familiarization instead (i.e., the animated backgrounds).

It goes without saying that experience with animations and screens, with which the infants tested here had extensive contact before their lab visit, is necessary to understand (i) that this particular class of stimuli is (potentially) representational; and (ii) how the representational medium works. There is no reason to expect that sampling from a population of infants without experience with animations would have produced the same results as the ones presented here. Participants' prior experience with animations was a precondition for testing whether infants interpret certain classes of stimuli as representations of entities belonging to familiar classes (balls, animals). Note, however, that experience with screens is not enough to pass the task. The Crossover Experiment 2 was recently replicated in a population of parrots, *Nestor notabilis,* well familiarized with screens (Bastos et al., 2021). Nevertheless, despite extensive screen experience, the parrots behaved in line with the **naïve realism** hypothesis and expected to find the virtual objects in the real boxes under the screen.

Finally, I would like to highlight two questions the present study does not answer. The first open question concerns the role of the experimenter. While she did not explicitly link symbols (e.g., the red circle) and referents (e.g., the ball), the experimenter did scaffold infants' interpretations by providing labels that could be mapped onto the visual objects on the screen (e.g., "ball", "rabbit", "bear"). It is thus unclear whether infants would interpret animations in the same way if left to their own devices, and it remains an open question what scaffolding elements infants need to interpret animations as they did here.

The second open question concerns the interpretation infants would give to other classes of stimuli, such as videos, which were not tested in the current studies. I cannot exclude the possibility that a setup like the one in Experiment 2, but with video recordings instead of animations, might fool infants into accept-

ing the screen–reality crossover. However, a direct comparison with other classes of stimuli would go beyond the scope of the current project. The findings should be taken as a proof of concept that the interpretation of certain stimuli (i.e., animations) is compatible with an early understanding of representations, not as evidence that infants have mastered the entire ontology of their environment. Even if infants were to reject a video–reality crossover in a setup such as the one in Experiment 2, virtual reality or realistic holograms would most likely lead them into error. The goal, however, was not to fool infants but to investigate their responses to stimuli that do not fool them.

2.8. Conclusion

The data in this chapter point to several conclusions. First, the world of infants is not a continuous spatiotemporal hodgepodge, as they do not confuse animations with their immediate environment. By 19 months, they have figured out that what happens on-screen stays on-screen and can answer questions about the location of objects appropriately based on this knowledge. Second, they have also figured out that animations are independent of the physical location they are presented at. That is, they dissociate medium and content, just like adults do. I take this as evidence for the claim that infants of this age—and, it goes without saying, from an industrialized population—might already interpret animated objects and events as representations.

Chapter 3. 15-Month-Olds Know That Arbitrary Objects Can Stand For Familiar Kind Tokens

3.1. Introduction

Humans often set up arbitrary and local mappings between visual objects and entities they want to communicate about. These objects are sometimes immediately available in the environment, in which case one can manipulate them to depict a relevant scene (e.g., bottles on the table rearranged to convey a distal spatial configuration). Alternatively, the objects themselves can be created for the occasion, as in graphs or maps (e.g., circles used to represent the size of the average lion).

What these objects represent often cannot be retrieved from their visual or behavioral features. In such cases, the identity of the referent must be conveyed via linguistic stipulation, either orally (e.g., "This pencil is the car, and the pin is the pedestrian") or in a legend appended to the representation (e.g., ◯ = lion). Taken literally, these predicative expressions would give rise to confusion since pencils are not cars and nor are circles lions. Yet the literal interpretation does not even seem to be even considered. In such cases, human adults intuitively infer that "is a" and the equality sign are shorthand for "stands for" and follow the content of the depiction without being confused about the literal falsity of the predication. Moreover, adults are also aware that these mappings are local. Outside of the current communicative context, they do not assume that the objects will continue to stand for the referents they happened to stand for previously.

In this chapter, I ask, first, whether human infants can set up STAND-FOR relations between arbitrary visual objects (e.g., a triangle) and discourse referents[1] belonging to kinds that infants are familiar with (e.g., a dog) based on linguistic stipulation (e.g., "This is your spine": **Figure 3.1**).

[1] Strictly speaking, discourse referents are mental representations of entities under discussion. For ease of exposition, I will use "discourse referents" to refer to both the entities under discussion and their associated mental representations, except when context does not allow disambiguation.

Figure 3.1. US anti-drugs campaign poster from the 1980s (left) and internet meme (right), illustrating that predicative expressions can be used to introduce STAND-FOR relations.

Second, I test whether the relation between object indexes and discourse referents is a one-to-one function. In Chapter 1, I hypothesized that each symbol in a representation should stand for only one referent within a discourse. If true, there arises an alternative account for children's mutual exclusivity inferences (Halberda, 2003; Markman & Wachtel, 1988) that does not involve postulating inbuilt lexical assumptions. When presented with two objects, a familiar one (e.g., a toy car) and an unfamiliar one (e.g., a cocktail strainer), children may interpret both objects as symbols. For the familiar one, they rely on iconicity to infer what it stands for (e.g., a car). They assume that the unfamiliar one stands for something as well. Still, they do not yet represent the concept under which it falls, so its referent is semantically unspecified. When asked about the "blicket", children infer that they are supposed to point to the strainer because of the assumption that symbols and referents are in a one-to-one relation and because the strainer is the only symbol that is currently unsaturated.

Third, I test whether the STAND-FOR relations between object indexes and discourse referents are local to the discourse. Recall that STAND-FOR relations are restricted to the discourse in which they are embedded. While there are no precise criteria that specify how discourses are individuated, it is reasonable to assume that discourses are individuated by speakers, all else equal. This as-

sumption is driven by evidence research on both pretend play, in which toddlers do not assume that pretend stipulations generalize across speakers (Andrasi et al., 2022; Wyman et al., 2009), and discourse referents in infancy (Brody, 2020). Testing the locality of the mappings also allows me to tease apart the IS-A and the STAND-FOR interpretations, as only the latter should be local and speaker-dependent. Presumably, a dog is a dog in a context-independent way.

Fourth, I test whether infants are sensitive to the conceptual identity of the discourse referents. In **Chapter 1**, I hypothesized that the interpretation of symbols relies on the conceptual system to generate descriptions (e.g., this object stands for **a dog**; that object stands for **a house**) that contribute to the individuation of discourse referents. If this is correct, infants should distinguish not only between different referent types (e.g., between a dog and a spoon) but also between the predicates that can be felicitously applied to the discourse referents (e.g., dogs move, but spoons do not).

3.2. Experiment 1: Different Symbols

The primary motivation for the current series of experiments was to explore whether 15-month-old infants understand STAND-FOR relations between perceptually available objects and discourse referents. I chose this age range to target infants who are yet to engage in object substitution pretense themselves (this starts around 18 months: Leslie, 1987; Piaget, 1945/1962). Even though there are a couple of studies suggesting that 15-month-olds understand some aspects of pretense (Bosco et al., 2006; Onishi et al., 2007), there is, as far as I know, none that tested the STAND-FOR relations directly. In addition, unlike pretend play studies, the present paradigm gives infants no cue that what they are shown is not to be taken literally. The second reason I tested 15-month-olds was to increase the number of basic-level nouns known by the participants. This way, multiple distinct conceptual identities could be assigned to the discourse referents without repeating them across trials. At the same time, this reduces the risk that infants interpret the linguistic stipulation events as a word-learning game. If infants presented with a picture of a geometric shape are told, "A book!" long after learning that "book" refers to books, they would be less likely to assume that a geometric shape **is a** book than before understanding the meaning of "book". Finally, I chose a wider age range (14 months 0 days–16 months 30 days) because

I did not think that the capacity to set up STAND-FOR relations would develop significantly in this interval[2].

Experiment 1 tested whether infants can represent STAND-FOR relations between visual objects and discourse referents in a looking-while-listening paradigm, based on Pomiechowska et al. (2021). If they can, they should be able to accept objects as symbols for discourse referents even when those objects do not belong to the kind to which the discourse referent belongs—just like they do in pretense. Infants were exposed to geometric shapes (e.g., an octagon and a triangle), one of which received a label familiar to infants of this age (e.g., "car" applied to an octagon). Infants were then asked a question containing the same word used at stipulation (e.g., "Where is the car?") or a different word, not heard previously in the trial (e.g., "Where is the spoon?").

I predicted that, upon hearing the same word at test, infants (i) would look above chance at the labeled object; and (ii) would look longer at it than when hearing a different word. In this sense, trials in which infants hear a different word act as a control condition, ensuring that infants do not orient to the object merely because their attention is drawn to it. In another sense, however, these trials are also mutual exclusivity tests. If STAND-FOR relations are one-to-one, and if one of the visual objects already stands for a discourse referent, only the remaining object qualifies as a potential symbol for the different-word referent. This hypothesis predicts that infants will look at the labeled object at below-chance levels on trials in which they hear a different word at test.

3.2.1. Methods

TRANSPARENCY AND OPENNESS

The hypotheses and methods for Experiment 1 were preregistered at the Open Science Framework (https://osf.io/5gs48). In the **Data Analysis** section below, I note where and why I deviated from the preregistration in the primary analyses. The stimuli, sample trial videos, anonymized data, and analysis scripts are available on the project's online Open Science Framework repository, accessible at https://osf.io/x3naq/. The local ethical committee approved all experiments reported here, and informed consent was obtained from the participants' caregivers before the testing session. Participants were gifted a toy at the end of the testing session.

[2] The results support this conjecture (see **Appendix A**).

PARTICIPANTS

The final sample consisted of 32 typically developing German-speaking 14–16-month-old infants (M_{age} = 15 months 9 days, SD_{age} = 22.3 days). An additional four infants were tested and excluded due to fussiness (n = 2) or due to not providing sufficient valid data (n = 2; see **Data processing and exclusion** below). The sample size was set based on a pilot experiment with 10 participants analyzed with a growth curve analysis model that I eventually discarded. However, the sample size is large enough to detect a mid-sized effect with 80% power and is above average for infant studies.

APPARATUS

Infants' gaze was recorded using a Tobii Pro Spectrum Eye Tracker with an integrated 23.8-inch-diagonal monitor (resolution: 1920 × 1080; refresh rate: 60 Hz). External speakers delivered the sound. A custom-made Python program building on PsychoPy 2021.1.3 (Peirce et al., 2019) was used to calibrate the infants, present the stimuli, and collect the eye data.

Figure 3.2. Visual stimuli used in Experiment 1. Top: 12 photographs of objects belonging to kinds that are familiar to 15-month-old infants. Bottom: eight pairs of symbol stimuli. In each pair, the stimuli differ in both shape and color for maximum discriminability. Possible color pairings are orange–green and blue–red.

STIMULI

Two sets of visual stimuli were used: color photographs representing 12 kinds of objects (Figure 3.2, top) that are familiar to German-speaking infants of this age

(Grimm & Doil, 2019) and eight pairs of geometric shapes (Figure 3.2, bottom). The objects' bounding boxes were matched in height and, whenever possible, in wdith; their display size was approximately 330 × 330 pixels (Figure 3.3). The stimuli also included an image of a pointing hand, displayed at 213 × 366 pixels.

Audio stimuli were the 12 nouns corresponding to the familiar kinds depicted by the photographs, embedded in different carrier phrases: "Hi baby! Look! An X! Here's an X! Wow, an X!", "Where is the X? X!". The sound stimuli were recorded by a female native speaker of Austrian German in infant-directed speech.

PROCEDURE

Infants were shown animated clips while seated on their caregivers' laps. The caregivers wore opaque glasses that did not allow them to see what was shown to the infants on-screen. The experiment consisted of 14 trials, split into four **Training**, four **Experimental**, and six **Word Knowledge** trials.

Training trials (Figure 3.3, top) were meant to familiarize infants with the general procedure and to give them evidence (i) that the voice they heard is connected to what is happening on the screen; and (ii) that the speaker speaks the same language as them. A Training trial consisted of three parts: **baseline**, **highlight**, and **test**. All trials started with a blue curtain covering the entire display. An attention-getter appeared in the center of the screen and rotated until the infant oriented to it for 500 ms. The curtain then went up to reveal a static display of two object photographs, one on the left and one on the right of the screen (e.g., a cat and a spoon). In the baseline part, the static display was shown to infants for 2 seconds in silence. After this period, the highlight part started. An animated hand appeared above one of the two objects (e.g., the cat), pointing to it. The hand moved up and down while the infant was greeted ("Hi baby! Look!") to draw their attention to the object. The hand stopped above the object, and infants heard the word typically used to refer to it three times in different carrier phrases ("A cat! Here's a cat! Wow, a cat!"). The highlight part lasted approximately 10 seconds and was followed by a 750-ms break, in which nothing happened.

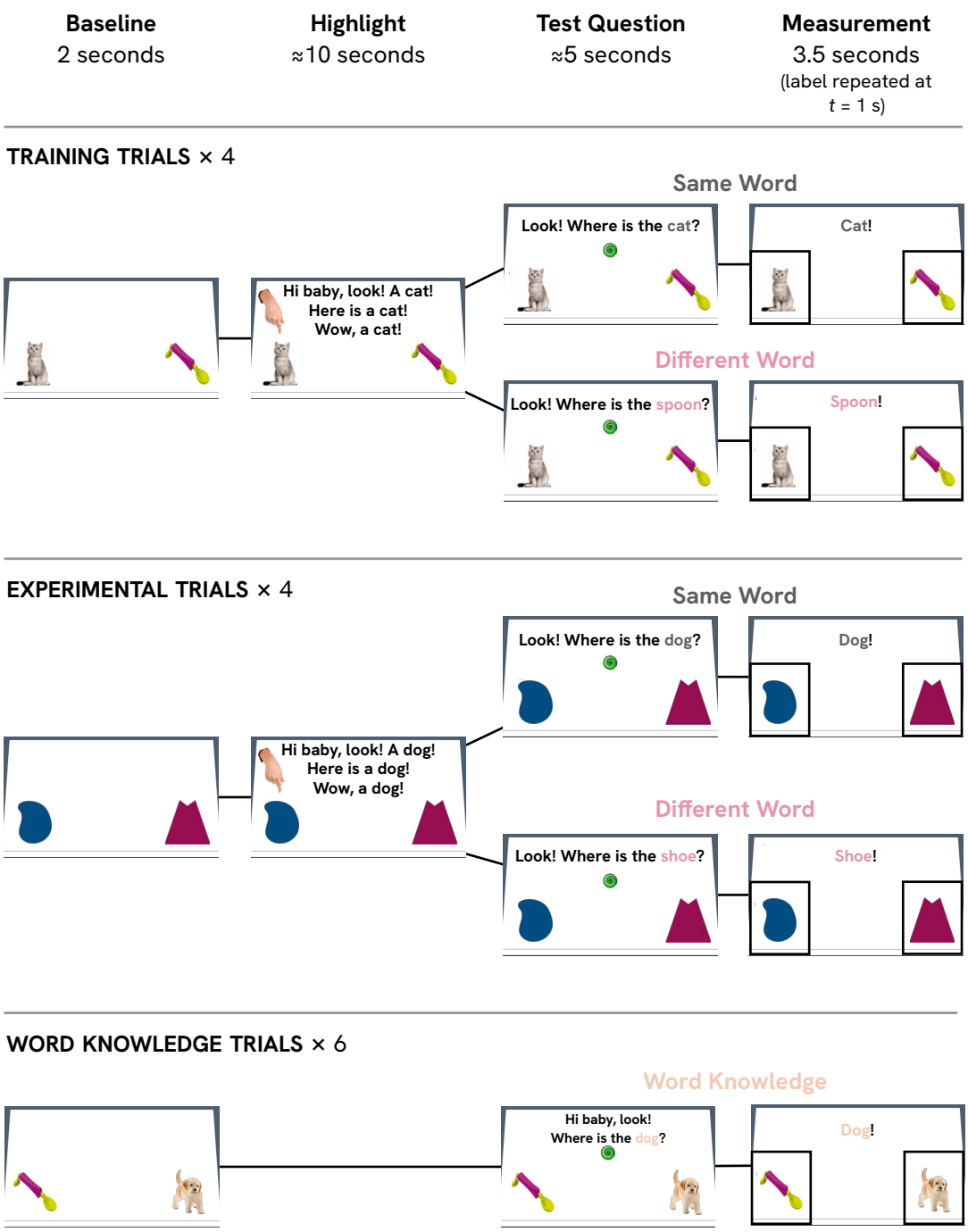

Figure 3.3. Trial sequence for each phase and trial type. Bounding boxes around the objects in the last column indicate areas of interest. Each trial started with a 2-second baseline, in which the objects were presented in silence. Training and Experimental trials contain a highlight event, in which one of the objects was pointed to and labeled. The highlight event was followed by a 750-ms break (not displayed). In all trial types, an attention-getter drew infants' attention to the center of the screen. The test question was played, then the attention-getter disappeared, and infants' looking behavior was measured. The test word was repeated one second into the measurement period.

The test event started with a colored rotating spiral in the center of the screen to draw infants' attention to the middle of the display. Once the infant oriented to it for 500 ms, the spiral started expanding and contracting cyclically while the test question was played. Depending on the trial type, the test question contained the same word used during the highlight event (e.g., "Where is the cat?") or the word typically referring to the other object on-screen (e.g., "Where is the spoon?"). The attention-getter disappeared at the offset of the test question, which coincided with the start of the measurement period. One second into the measurement period, infants heard the test word one more time (e.g., "Cat!"/ "Spoon!"). The measurement period ended after 3.5 seconds, when the blue curtain went down to cover the entire display. Each trial lasted approximately 22 seconds.

Experimental trials were identical to Training trials except that the two object images were replaced by geometric shapes (**Figure 3.3**, middle). Word Knowledge trials were identical to Training trials, except that the highlight event was removed. The two-second silent period at the beginning of the trial was immediately followed by the test question (**Figure 3.3**, bottom).

DESIGN

The experiment had a within-subjects design with one independent variable, **Trial type**, which was two-leveled: **Same Word** and **Different Word**. For each infant, the first eight trials alternated according to the ABAB-ABBA pattern (Training–Experimental), with type of first trial (Same Word or Different Word) counterbalanced across subjects. The side of the object that was pointed to and labeled during the highlight event was alternated according to an ABBA-ABAB pattern, with first side (left or right) counterbalanced across subjects. For the six Word Knowledge trials, the side of the correct response followed the ABBABA structure, with first side (left or right) counter-balanced across subjects.

The object pairings were randomly sampled for each subject. First, eight of the 12 object photographs were sampled and grouped into four pairs to create the Training trials (1–4). In each pair, the two objects were not allowed to belong to the same superordinate kind (e.g., agents: dog–duck), and the two words referring to them were not allowed to start with the same phoneme (e.g., "Buch [book]"–"Banane [banana]"). Within each pair, one object was the target of labeling in the highlight phase. Note that six words are needed to create four trials (one for each Same Word trial, two for each Different Word trial).

For the Experimental trials (5–8), I sampled four of the eight geometric shape pairings and assigned the remaining six words for the highlight phase. There was thus no overlap between the words used in the Training trials and those used in the Experimental trials. After the Experimental trials, infants saw a 30-second animated movie.

The Word Knowledge trials (9–14) tested whether infants knew the words used during the Experimental trials. In each trial, an object photograph denoted by a noun previously applied to the geometric shapes was paired with an object photograph from the Training phase. See **Table A1** in Appendix A for an example of a randomization list for one participant.

DATA PROCESSING AND EXCLUSION

At the end of each testing session, the Python script outputted a data file containing the infant's gaze information on each sample (every 16.67 ms). The gaze coordinates were averaged across eyes along both horizontal and vertical axes. The screen was divided into three regions, depending on whether the infant was looking at the left-object area of interest (AOI), the right-object AOI, or elsewhere on the screen (**Figure 3.3**, last column). A gaze data point was considered valid if the eye-tracker registered the gaze for at least one eye. As preregistered, an Experimental trial was excluded from the analysis if infants provided less than 60% valid data during baseline (n = 3 trials) or test (n = 6 trials). Infants were excluded if they failed to provide at least one valid trial of each type in the Experimental phase (n = 2).

After this preprocessing step, two new variables were derived. The first variable, Highlighted Object, which applies only to Training and Experimental Trials, received a score of 1 if infants' gaze fell into the AOI of the highlighted object, 0 if infants looked to the AOI of the other object, and NA if infants looked elsewhere on the screen. The second variable, Target Object, applied only to Word Knowledge trials (where no object was highlighted) and measured whether infants looked at the object they were asked about at test. The samples obtained during the test period of each trial were grouped into 50-ms bins (70 bins in total, corresponding to the 3.5-second test period). For each bin, I computed the variable **PLH**, representing the proportion of gaze samples that fell into the region of the highlighted object (in the Training and Experimental trials) or target object (in Word Knowledge trials) out of the samples that fell into the regions of both objects.

MISSING DATA

Due to the structure of the task, there were several types of missing data: (i) missing gaze data due to eye tracker signal loss (e.g., because infants looked away from the screen)—these data were removed from the analysis; (ii) missing AOI gaze data (e.g., because infants looked to other parts of the screen during a particular bin)—these data were removed from the analysis; (iii) missing Training and Experimental trials (e.g., because the infant did not provide enough valid data for that trial)—these trials were removed from the analysis; and (iv) missing Word Knowledge trials (because the infant did not provide data in a Word Knowledge trial)—I used the joint posterior probability distribution of a Bayesian model to impute the missing values (n = 2; see **Appendix A** for a complete specification of the model). Finally, for the Training and Word Knowledge trials in which infants' baseline preference between the two objects could not be estimated (n = 5 trials), their preference was imputed as neutral (i.e., as 0.5).

DATA ANALYSIS

I initially preregistered a Bayesian growth curve analysis model on infants' proportion of looks at the highlighted object in each test time bin, with participant, trial type, baseline PLH, test label knowledge (as measured in Word Knowledge trials), and time polynomials as predictors. Ultimately, I decided against this model due to the high false-positive rate of growth curve models (Huang & Snedeker, 2020) and the lack of interpretability of time polynomials. Instead, I report two-tailed paired t-tests on PLH trial averages for all predictions. Even though this analysis deviates from the preregistration, the model considering only trial averages is more conservative than a growth curve model. This makes it more likely to falsify the predictions than to confirm them.

I also conducted a Bayesian generalized linear model analysis, identical in structure to the preregistered one, except that trial average PLH scores were used instead of PLHs at each time point. Again, this analysis is more conservative than the preregistered one. To mitigate overfitting, I ran a single model on Experiments 1–3, whose results I report in **Section 3.6**.

The data for all experiments were analyzed using R 4.2.2 (R Core Team, 2022), and the packages *ggplot* 3.4.2 (Wickham, 2016), and *rethinking* 2.01 (McElreath, 2020). The reproducible code for the data analysis has been uploaded to the Open Science Framework repository of the project and can be accessed at https://osf.io/x3naq/files/osfstorage.

3.2.2. Results

PROPORTION OF LOOKING AT THE HIGHLIGHTED OBJECT AT TEST

Figure 3.4 plots infants' proportions of looking at the highlighted object at test (PLH), averaged by trial type and phase. In the Training phase, infants looked at the highlighted object above chance on Same Word trials, M = .70; $t(31)$ = 4.68, p < .001, Cohen's d = .83, 95% CI [.61, .78]. On Different Word trials, infants' PLH scores were not different from chance, M = .5; $t(31)$ = 0.003, p = .997, Cohen's d = 0.001, 95% CI [.394, .606]. Infants looked longer at the highlighted object on Same Word than on Different Word trials, $M_{difference}$ = 0.2, $t(31)$ = 2.72, p = .011, Cohen's d = 0.48, 95% CI [0.05, 0.34].

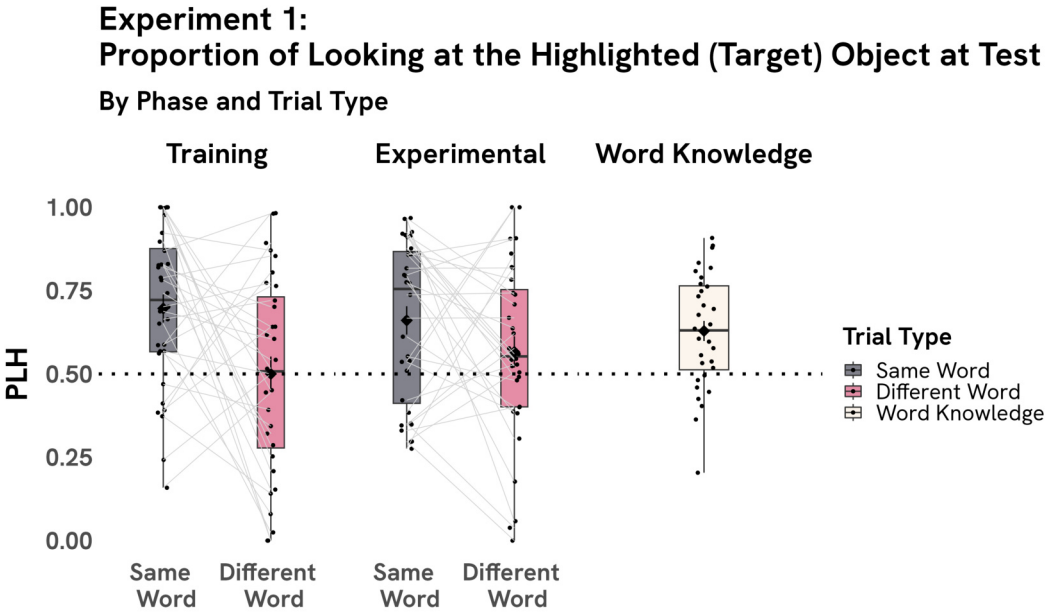

Figure 3.4. Results of Experiment 1, split by trial type and phase. The *y*-axis plots the proportion of looking at the highlighted object (Training and Experimental phases) or at the target object (Word Knowledge phase) at test. Black circles and the lines connecting them represent individual averages as a function of trial type and phase; black diamonds depict group averages ±1 SEM; boxplots indicate the median and interquartile range. The dashed horizontal line marks the chance level.

In the Experimental phase, infants behaved similarly. They looked at the highlighted object above chance on Same Word trials, M = .66; $t(31)$ = 3.78, p = .001, Cohen's d = 0.67, 95% CI [.57, .75]. On Different Word trials, infants'

PLH scores did not differ from chance, M = .57; $t(31)$ = 1.41, p = .169, Cohen's d = 0.25, 95% CI [.47, .66]. While they looked longer at the highlighted object on Same Word than on Different Word trials, the extent of the difference does not exclude the null hypothesis, $M_{difference}$ = 0.09, $t(31)$ = 1.55, p = .132, Cohen's d = 0.27, 95% CI [–0.03, 0.22].

Finally, the Word Knowledge phase results show that infants were familiar with the labels that were applied to the geometric shapes in the Experimental phase, M = .63; $t(31)$ = 4.28, p < .001, Cohen's d = 0.76, 95% CI [.57, .69].

EXPLORATORY: DIRECTION OF THE FIRST SACCADE AT TEST (PREREGISTERED)

As an exploratory measure, I analyzed which of the two objects infants oriented first at test. For each trial, infants received a score of 1 if their first gaze in one of the two object AOIs went to the side of the highlighted object and 0 if it went to the one not highlighted. Infants oriented to the highlighted object at above-chance levels on Same Word trials, $t(31)$ = 3.26, p = .003, Cohen's d = 0.58, 95% CI [.58, .86], but not on Different Word trials, $t(31)$ = 1.94, p = .062, Cohen's d = 0.34, 95% CI [.49, .74].

EXPLORATORY: BILINGUALS (NOT PREREGISTERED)

Because I did not expect bilingualism to play a role in this experiment, the sample included both monolinguals and bilinguals. However, I noticed upon analyzing the data that the five bilingual participants responded differently from their monolingual counterparts in all experiment phases. Notably, their proportions of looking at the target object on Word Knowledge trials were not different from chance, M = .53, $t(4)$ = 0.34, p = .75, Cohen's d = 0.14, 95% CI [.30, .77]. While this could have been a fluke, I decided to restrict the sample to monolingual children for the following set of experiments[3]. If bilingual infants do not recognize the words used in the Experimental phase, the measurement will be confounded because these infants might interpret the labeling events as kind membership predication. In this case, they would look longer at the highlighted object at test not because they take the shape as a symbol of a kind token but because they think the shape belongs to the kind denoted by the label.

[3] The effect of trial type in the Experimental phase is stronger in the monolingual subsample, $t(25)$ = 1.92, p = .066. This was an additional reason for which I decided to restrict the sample to monolingual participants in Experiments 2–4.

3.2.3. Discussion

To test the hypothesis that 15-month-old infants understand that objects can be symbols of various things, I asked whether they accept the assignments of familiar conceptual identities (e.g., dog) to arbitrary visual objects (e.g., a triangle) based on language (e.g., "Look! A dog!" while pointing to the triangle). The results partly show that they do. Infants looked at the highlighted object reliably—and oriented to it first—only when asked about the word applied to the object during the highlight phase. By contrast, when asked about a different word, infants did not distinguish between the two objects either in their looking times or in the direction of the first saccade. As for the lack of a statistical difference between Same Word and Different Word trials, this may be driven by two factors. First, the sample size was set based on a growth curve model, which probably overestimated the magnitude of the effect. Second, the inclusion of bilinguals may have increased measurement noise.

Experiment 1 also tested whether the relations between symbols and discourse referents are one-to-one. If this assumption is embedded in the cognitive system dealing with STAND-FOR relations, infants should look at the non-highlighted object when hearing a different word at test. This was not the case. Infants were at chance between the two objects when asked about a different word in the Experimental phase. However, infants showed the same pattern with photographs of familiar objects, which suggests that the labeling and pointing have an additive effect on infants' gaze behavior at test. This interpretation is strengthened by the differences in infants' scores between Training Same Word trials and Word Knowledge trials (70% vs. 63%) and in line with previous findings on the effect of labeling on attention (Baldwin & Markman, 1989).

Thus, two possibilities cannot be teased apart with the current design. On the one hand, infants may not make any assumption concerning the one-to-one nature of STAND-FOR relations. If a triangle stands for a dog, infants will not assume that the remaining octagon must stand for a shoe if a speaker asks about one. On the other, infants may have made the mutual exclusivity inference, but this was masked by the asymmetric effect that pointing to and labeling an object have on infants' subsequent attention to that object.

However, testing the assumption that STAND-FOR relations are one-to-one was secondary to the main goal of the present study. After all, different Word trials also serve as a control for the study's primary question, which asks

whether infants can set up assignments between arbitrary objects and discourse referents. If they are, they should be able to map a familiar noun phrase onto a geometric shape, which they are. At the same time, there are several alternative accounts that can explain the data in Experiment 1 (even though, as far as I can tell, none that would have predicted it).

The first alternative to the STAND-FOR interpretation is an **associative** account according to which infants did not interpret the labeling event at all. Instead, infants associated a meaningless phonological string (e.g., "dog") with the shape that was highlighted while they heard the string (e.g., the blue blob).

The second family of alternatives is that infants interpreted the predicative expression literally, that is, as referring to a property that truly applies to the labeled shape. In turn, this version comes in two flavors. On the one hand, infants may have accepted that the blue blob is, despite appearances, actually a dog and recategorized it as such (e.g., Jaswal & Markman, 2007). I will refer to this as the **recategorization** account. On the other hand, infants may have accepted that the blob is a "dog" not because it is a dog but because "dog" is a homonym that means either dog or blob. Older children seemingly struggle with incorporating homonyms into their lexicon (e.g., Mazzocco, 1997), but this remains a logical possibility. I will call this the **new lexical entry** account.

The third family of alternatives involves attributing a property to the speaker to accommodate the fact that she labeled a blue blob as a "dog". Again, there are several versions in which this can be construed. First, infants could have taken the speaker to be incompetent, as is typically assumed in mislabeling studies (e.g., Koenig & Echols, 2003), either ontologically (e.g., she falsely believes the blob is a dog) or linguistically (e.g., she falsely believes "dog" means blob; she speaks a foreign language in which "dog" means blob; she has an idiosyncratic lexicon in which "dog" means blob). I will refer to this option as the **incompetent speaker** account. Second, infants may have taken the speaker to intentionally mislead them into believing that the blob is a dog. Although this seems a stretch, given that children seem to understand lying much later (e.g., Mascaro & Sperber, 2009), I list it as a possibility worth considering. I refer to this as the **malevolent speaker** account.

One final possibility is that infants interpret the labeling event as a **referential pact** (Brennan & Clark, 1996; Matthews et al., 2010). In other words, they interpret that "dog" applied to the blue blob as a stipulation, just not of the

STAND-FOR kind. Instead, infants take the stipulation to introduce a local convention between themselves and the speaker, whereby "dog" is the label that they will now use to refer to the blue blob.

To tease apart between the **associative** account and the other alternatives, Experiment 2 exposed infants to identical-looking geometric shapes (e.g., two blue blobs). Like in Experiment 1, only one of the two shapes was pointed to and labeled. If infants passed the task in Experiment 1 by associating a noun to a feature set, there should be no difference between Same Word and Different Word trials. By contrast, if infants interpret the labeling event as connecting only the highlighted object to the referent denoted by the label, infants should look longer at the highlighted object at test in Same Word trials, as in Experiment 1.

3.3. Experiment 2: Identical Symbols

Experiment 2 was identical to Experiment 1, except that the two shapes on the screen looked identical in the Experimental phase. If infants associated a noun with visual features in Experiment 1, they should look at the highlighted object equally and at chance levels in both trial types. If, on the other hand, the stipulation induces infants to connect one visual object to one discourse referent, Experiment 2 should produce the same pattern of results as Experiment 1. This is based on the hypothesis that STAND-FOR stipulations operate at the level of individual objects, not object kinds. In turn, this rests on the assumption that the object indexing system used for tracking symbol identity uses location as the primary individuation criterion.

3.3.1. Methods

Except where noted, the methods were identical to those of Experiment 1. The Open Science Framework preregistration can be accessed at https://osf.io/m9uer.

PARTICIPANTS

The final sample consisted of 32 typically developing monolingual German-speaking 14–16-month-olds (M_{age} = 15 months 14 days, SD_{age} = 26 days). An additional nine subjects were tested and excluded because of fussiness (n = 6), be-

cause of a technical error (n = 1), or because they did not provide enough valid data (n = 2).

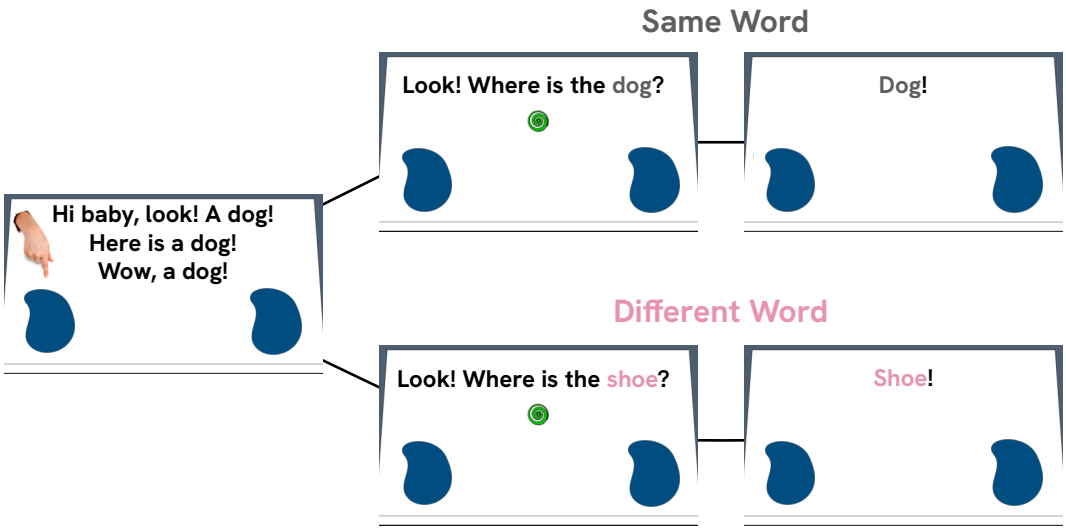

Figure 3.5. Schematic structure of **Experimental** trials in Experiment 2.

DESIGN

Experiment 2 differed from Experiment 1 only in the Experimental phase, where infants were presented with two identical-looking shapes (Figure 3.5). This could not be done in the Training phase, as the Different Word trials would have become infelicitous (e.g., asking about a spoon when there are two duck images on-screen), and infants were not supposed to infer that the speaker is unreliable.

DATA ANALYSIS

The same preregistered exclusion criteria as in Experiment 1 were used. Experimental trials were excluded if infants provided less than 60% on-screen data during baseline (n = 2 trials) or test (n = 6 trials). Infants who did not provide at least one trial for each trial type (n = 2) were also excluded. In addition, there were five trials in the Training and Word Knowledge phases in which infants' preference between the two objects could not be estimated (imputed as neutral preference), and two Word Knowledge trials in which infants did not provide data at test (imputed via the Bayesian vocabulary model).

3.3.2. Results

PROPORTION OF LOOKING AT THE HIGHLIGHTED OBJECT AT TEST

Figure 3.6 plots infants' proportions of looking at the highlighted object at test, averaged by trial type and phase. The Training phase replicated Experiment 1. Infants looked at the highlighted object above chance on Same Word trials, M = .64; $t(31)$ = 3.82, p = .001, Cohen's d = .68, 95% CI [.57, .72]. On Different Word trials, infants' PLHs did not differ from chance, M = .46; $t(31)$ = −0.87, p = .391, Cohen's d = −0.15, 95% CI [.38, .55]. Infants looked longer at the highlighted object on Same Word than on Different Word trials, $M_{difference}$ = 0.18, $t(31)$ = 3.29, p = .003, Cohen's d = 0.58, 95% CI [0.07, 0.29].

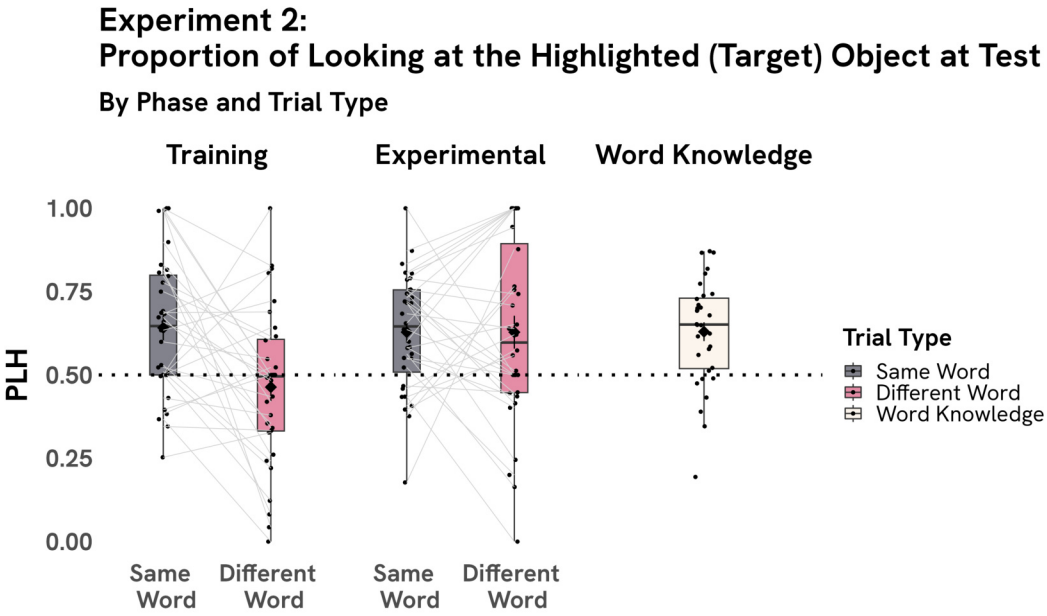

Figure 3.6. Results of Experiment 2, by trial type and phase.

In the Experimental phase, infants' looking behavior was markedly different. They looked at the highlighted object above chance on both Same Word trials, M = .63; $t(31)$ = 4.11, p < .001, Cohen's d = .73, 95% CI [.57, .69], and Different Word trials, M = .63; $t(31)$ = 2.73, p = .014, Cohen's d = 0.46, 95% CI [.53, .73]. There was no effect of trial type, $M_{difference}$ = 0, $t(31)$ = 0.003, p = .997, Cohen's d = 0.001, 95% CI [−0.11, 0.11].

Finally, the Word Knowledge phase results show that infants were familiar with the labels that were applied to the geometric shapes in the Experimental phase, M = .63; $t(31)$ = 4.61, p < .001, Cohen's d = 0.814, 95% CI [.57, .69].

EXPLORATORY: DIRECTION OF THE FIRST SACCADE AT TEST (PREREGISTERED)

As in Experiment 1, I also analyzed which of the two objects infants oriented first at test. Similarly to Experiment 1, infants oriented to the highlighted object above chance in Same Word trials, $t(31)$ = 3.70, p = .001, Cohen's d = 0.65, 95% CI [.61, .86], but not in Different Word trials, $t(31)$ = 0.68, p = .5, Cohen's d = 0.12, 95% CI [.41, .69].

3.3.3. Discussion

Infants looked at the highlighted object in both Same Word and Different Word trials at above-chance levels. At first glance, these results seem uninterpretable, as they seem to imply that infants shown two identical-looking shapes, one of which is given a familiar label (e.g., "dog"), consider that shape as a good candidate for being the referent of any label (e.g., "dog" and "shoe").

However, there are at least two other possibilities that are more sensible. First, infants could have rejected the stipulation to begin with, due to the lack of contrast between the two shapes. In that case, the above-chance levels in both trial types would have been driven only by the asymmetric highlighting of one of the two objects. In other words, the additive effect of pointing and labeling may have operated independently of the test label, which infants ignored altogether, and may have been so strong as to raise infants' preference for the highlighted object in both conditions. The drawback of this interpretation is that it cannot explain the effect on the direction of the first saccade.

Alternatively, infants may have accepted the labeling but generalized it to the non-highlighted same-looking shape. On Same Word trials, infants looked at the highlighted object at test not based on the noun only (e.g., "duck") but on the definiteness of the noun phrase (e.g., "the duck"), which could have only selected the previously introduced discourse referent. On Different Word trials, infants looked at the non-highlighted object below chance because they rejected it as a candidate for a new noun and returned to the highlighted object, possibly expecting that more events will happen on the side where communication occurred. This interpretation explains the effect on the direction of the first sac-

cade but requires positing an early sensitivity to definiteness—for which there is no evidence, as far as I know—and offers different explanations for the same behavior (i.e., looking at the highlighted object above chance) in the two trial types.

Whichever explanation is true, Experiment 2 failed to rule out alternative explanations for Experiment 1. If infants rejected the mappings altogether, Experiment 2 was not a good test of the **association** and **literal interpretation** accounts against the STAND-FOR hypothesis. If infants generalized the mapping from one shape to the other and were sensitive to the definiteness of noun phrases, this would be evidence against the **association** account—since infants do interpret the linguistic input—but not against the **literal interpretation** account. Infants could have extended their kind representations to include the new shapes (e.g., "dog" includes blobs in its extension) or created a new lexical entry for the known words (e.g., "dog" is a homonym and also means blob). In this case, infants would have generalized the word applied to the highlighted object to the other object because it had the same shape (Landau et al., 1988).

At the same time, Experiment 2 does not, on its own, falsify the STAND-FOR account. It remains possible that infants interpreted the highlight event as stipulating a STAND-FOR mapping, but one that they rejected or generalized to the other-looking shape based on reasons that the model in Chapter 1 failed to take into account (e.g., stipulation may operate at the level of symbol kinds, not tokens). What Experiment 2 did falsify is the assumption that the system used for tracking symbols prioritizes location over visual features. I return to this in the General Discussion.

To further look into alternative accounts for the results in Experiment 1, Experiment 3A introduces a second speaker, which tests the **associative** account and the **literal interpretation** accounts simultaneously. If infants represent the STAND-FOR relations as assignments local to a discourse, and if discourses are indexed by speakers, they should not generalize them to a different speaker. On the other hand, if infants associate a string and a visual object, recategorize the shape based on the label, or create a new lexical entry for the familiar noun, the identity of the speaker probing the association or representation should be irrelevant. Infants in the age range tested here should be able to distinguish between different speakers (Werchan et al., 2015) and generalize new words to new speakers (Buresh & Woodward, 2007) if they interpret the situation as a word-learning one.

3.4. Experiment 3A: Different Speakers

3.4.1. Methods

Except where noted, the methods were identical to Experiment 1. The Open Science Framework preregistration can be accessed at https://osf.io/ys57v.

PARTICIPANTS

The final sample consisted of 32 typically developing monolingual German-speaking 14–16-month-olds (M_{age} = 15 months 14 days, SD_{age} = 28.7 days). An additional seven subjects were tested and excluded due to fussiness (n = 4), parental intervention (n = 1), or failure to provide sufficient valid data (n = 2).

STIMULI

The audio stimuli recorded by the female speaker in Experiment 1 were doubled by a new set of stimuli recorded by a male native speaker. The recordings from the two speakers were closely matched in duration (r = .98; $M_{difference}$ = 33.8 ms; $SD_{difference}$ = 120.9 ms).

DESIGN

Experiment 3A differed from Experiment 1 in the introduction of a new speaker. Training trials were identical to Experiments 1 and 2, except that two were delivered by the female speaker and two by the male speaker. This was done to accustom infants to both speakers. Because infants were not supposed to infer that the speakers were part of the same scene, within the Training trials, the same speaker delivered both the stipulation and the test question.

To replicate Experiment 1, Experiment 3A contained two Experimental blocks. The blocks were presented in a fixed order. In the **Different Speaker** block (Trials 5–8), one speaker delivered the stipulation, and the other asked the test question. In the **Same Speaker** block (Trials 15–18, after the Word Knowledge block), the same speaker delivered the stipulation and test question. This block replicates Experiment 1. To further minimize the evidence that both speakers are present in the scene in the Different Speaker block, the test question was prefaced by another greeting of the infant (e.g., Speaker 1: "Hi baby! Look! A dog! (...)"; Speaker 2: "Hi baby! Look! Where is the dog/spoon?"). The same was done for the Same Speaker block (e.g., Speaker 1 or 2: "Hi baby! Look! A dog! (...) Hi baby! Look! Where is the dog/spoon?").

For the Training and Different Speaker trials (Trials 1–8), the speaker delivering the stipulation alternated in an ABBA-ABAB pattern for half the subjects and in an ABAB-ABBA pattern for the other half (Male vs. Female on the first trial counterbalanced). Trial type, speaker identity, and side of the highlighted object were counterbalanced across subjects. In the Word Knowledge trials (Trials 9–14), half the subjects were tested by the male speaker and half by the female speaker. The Same Speaker block (Trials 15–18) was identical to the Different Speaker block, except that the same speaker delivered both the stipulation and the test question.

DATA ANALYSIS

The same preregistered exclusion criteria as in Experiment 1 were applied to the Different Speaker block. The trials in which infants provided less than 60% on-screen data during baseline (n = 0 trials) or test (n = 12 trials) were excluded. Infants were excluded from the analysis if they did not provide at least one valid trial for each trial type in this block (n = 2). The same exclusion criteria were applied in the Same Speaker block, which came last (n = 17 trials). In addition, three infants fussed out before reaching this phase and did not provide any data in the Same Speaker block. These infants were included in the analysis because they provided sufficient data in the Different Speaker block.

There were four trials in the Training and Word Knowledge phases in which infants' preference between the two objects could not be estimated (imputed as neutral preference) and five trials in which infants' word knowledge could not be assessed (imputed from the Bayesian vocabulary model).

3.4.2. Results

PROPORTION OF LOOKING AT THE HIGHLIGHTED OBJECT AT TEST

Figure 3.7 plots infants' average proportions of looking at the highlighted object at test by trial type and phase. In the Training phase, as in Experiments 1 and 2, infants looked at the highlighted object above chance on Same Word trials, M = .70; $t(31)$ = 5.08, p < .001, Cohen's d = .90, 95% CI [.62, .78]. On Different Word trials, infants' PLH scores did not differ from chance, M = .51; $t(31)$ = 0.37, p = .72, Cohen's d = 0.07, 95% CI [.44, .59]. Infants looked longer at the highlighted object on Same Word than on Different Word trials, $M_{\text{difference}}$ = 0.19, $t(31)$ = 3.69, p = .001, Cohen's d = 0.65, 95% CI [0.09, 0.29].

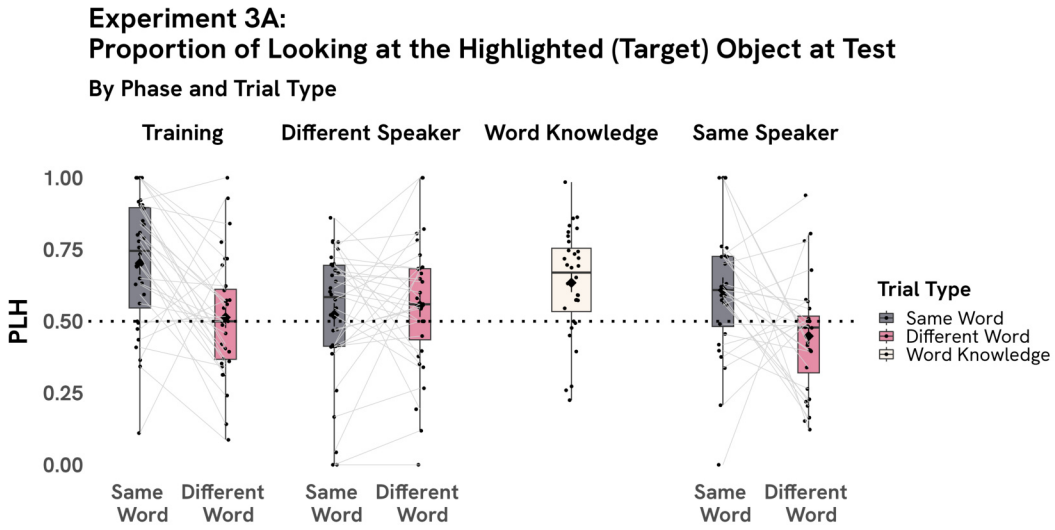

Figure 3.7. Results of Experiment 3A, by trial type and phase.

In the Different Speaker block, infants' behavior differed. They did not look at the highlighted object above chance either in Same Word trials, M = .52; $t(31)$ = 0.55, p = .584, Cohen's d = 0.1, 95% CI [.44, .61], or in Different Word trials, M = .56; $t(31)$ = 1.38, p = .179, Cohen's d = 0.24, 95% CI [.47, .64]. Infants looked equally at the highlighted object in Same Word and Different Word trials, $M_{difference}$ = −0.03, $t(31)$ = −0.7, p = .489, Cohen's d = −0.12, 95% CI [−0.13, 0.06].

By contrast, in the Same Speaker block, infants looked at the highlighted object above chance in Same Word trials, M = .61; $t(27^4)$ = 2.44, p = .022, Cohen's d = 0.46, 95% CI [.52, .70], but not in Different Word trials, M = .45; $t(27)$ = −1.37, p = .18, Cohen's d = −0.26, 95% CI [.37, .53]. The effect of trial type was higher than in Experiment 1 and incompatible with the null hypothesis, $M_{difference}$ = 0.15, $t(26)$ = 2.21, p = .036, Cohen's d = 0.43, 95% CI [0.01, 0.28].

A 2 × 2 repeated-measures Anova on the subset of subjects who provided at least one valid trial of each type in both blocks (n = 27) revealed no main effects of block or trial type (block: $F(1, 26)$ = 0.24, p = .63, η_p^2 = .01; trial type: $F(1, 26)$ = 1.33, p = .259, η_p^2 = .05) but an interaction between block and trial type, $F(1, 26)$ = 5.63, p = .025, η_p^2 = .18.

Finally, the Word Knowledge phase results demonstrate that infants knew the labels that were applied to the geometric shapes in the Experimental phase, M = .63; $t(31)$ = 4.12, p < .001, Cohen's d = 0.73, 95% CI [.57, .70].

[4] There are fewer degrees of freedom in this block because three infants fussed out before the Same Speaker block started.

EXPLORATORY: DIRECTION OF THE FIRST SACCADE AT TEST (PREREGISTERED)

I also analyzed which of the two objects infants oriented first at test. In the Different Speaker block, infants did not orient to the highlighted object above chance in any trial type (Same Word trials: $t(31) = 1.49$, $p = .147$, Cohen's $d = 0.26$, 95% CI [.46, .76]; Different Word trials: $t(31) = 1.22$, $p = .23$, Cohen's $d = 0.22$, 95% CI [.45, .71]). By contrast, in the Same Speaker block, infants oriented to the highlighted object above chance in Same Word trials, $t(27) = 2.08$, $p = .048$, Cohen's $d = 0.39$, 95% CI [.502, .82], but not on Different Word trials, $t(27) = 1$, $p = .326$, Cohen's $d = 0.19$, 95% CI [.43, .72], like in Experiments 1 and 2.

3.4.3. Discussion

In Experiment 3A, infants' looking behavior at test varied systematically depending on whether the same or a different speaker asked the test question. In the Same Speaker block, infants looked at the highlighted object above chance only on Same Word trials and longer on Same Word trials than on Different Word trials. In the Different Speaker block, infants did not look longer at the highlighted object, irrespective of trial type. This is evidence against both the **association** account and the **literal interpretation** accounts, as associations, as well as recategorization and lexical entry creation, are speaker-independent processes.

Incidentally, the results also suggest that the effect that pointing and labeling have on infants' attention to the object at test does not stem from a rudimentary attentional mechanism. In the Different Speaker block, infants were at chance between the two objects. This suggests that, in Experiment 2, they oriented to the highlighted object not merely because it was highlighted but due to an expectation targeting the speaker that drew their attention to that side.

Even though I preregistered that I would only run a version with the reverse order of blocks if there were no effect of trial type in the Same Speaker block in Experiment 3A, I wanted to make sure that the effect is robust and to test for a possible order effect (e.g., perhaps infants take longer to adjust to the task when there are two speakers involved, thereby showing a trial type effect in the second but not in the first block). Therefore, I ran a variant of Experiment 3A that switched the order of the Experimental blocks.

3.5. Experiment 3B: Different Speakers Reversed

3.5.1. Methods

The methods were identical to those in Experiment 3A, except that the Same Speaker block (Trials 5–8) came before the Different Speaker block (Trials 15–18). Because of the high similarity, I did not create an additional preregistration.

PARTICIPANTS

The final sample consisted of 32 typically developing monolingual German-speaking 14–16-month-olds (M_{age} = 15 months 5 days, SD_{age} = 25.4 days). An additional nine subjects were tested and excluded due to fussiness (n = 5) or for not providing enough valid data (n = 4).

DATA ANALYSIS

The criteria for trial exclusion were identical to Experiments 1–3A. Five trials were excluded in the Same Speaker block because infants provided less than 60% on-screen data during baseline (n = 1) or test (n = 4). Infants were excluded from the analysis if they did not provide at least one valid trial for each trial type in this block (n = 4). The same exclusion criteria were applied in the Different Speaker block, which came last (n = 19 trials). In addition, two infants fussed out before reaching the Different Speaker block but were included in the analysis because they provided sufficient data in the Same Speaker block. There were four trials in the Training and Word Knowledge phases in which infants' preference between the two objects could not be estimated (imputed as neutral preference) and nine Word Knowledge trials in which infants did not provide data at test (imputed via the Bayesian vocabulary model).

3.5.2. Results

PROPORTION OF LOOKING AT THE HIGHLIGHTED OBJECT AT TEST

Figure 3.8 plots infants' average proportions of looking at the highlighted object at test by trial type and phase. In the Training phase, similarly to Experiments 1–3A, infants looked at the highlighted object above chance on Same Word trials, M = .69; $t(31)$ = 5.99, p < .001, Cohen's d = 1.06, 95% CI [.63, 76], but not on Different Word trials, M = .59; $t(31)$ = 1.92, p = .056, Cohen's d = 0.35, 95% CI [.497, 0.69]. The effect of trial type was smaller than in Experiments 1–3A and not

large enough to reject the null hypothesis, $M_{difference}$ = 0.10, $t(31)$ = 1.90, p = .067, Cohen's d = 0.34, 95% CI [−0.007, 0.21].

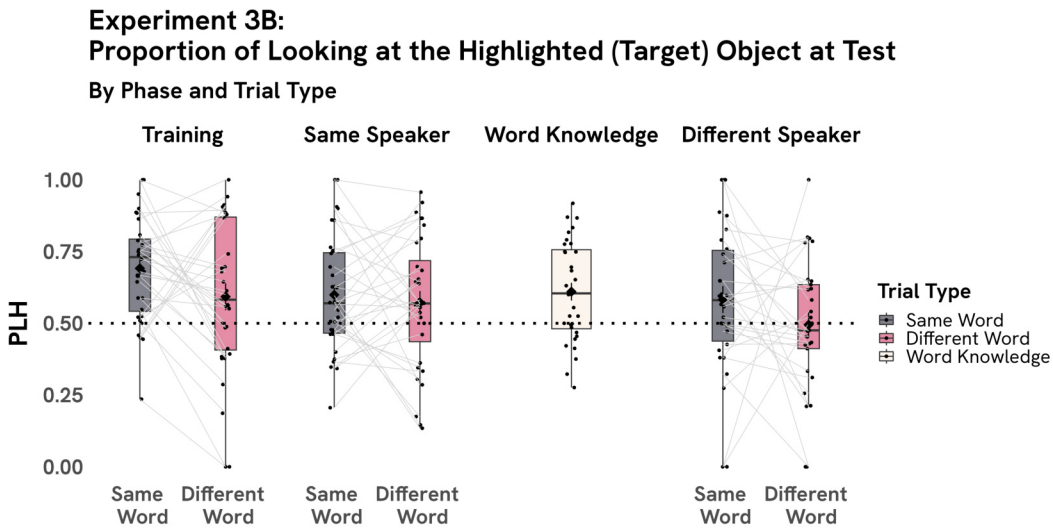

Figure 3.8. Results of Experiment 3B, by trial type and phase.

In the Same Speaker block, the results replicate Experiment 1 and, in part, Experiment 3A. Infants looked at the highlighted object above chance on Same Word trials, M = .60, $t(31)$ = 2.77, p = .009, Cohen's d = 0.49, 95% CI [.53, .67], but not on Different Word trials, M = .57; $t(31)$ = 1.80, p = .083, Cohen's d = 0.32, 95% CI [.49, .654]. The times spent looking at the highlighted object did not differ between Same Word and Different Word trials, $M_{difference}$ = 0.03, $t(31)$ = 0.57, p = .571, Cohen's d = 0.10, 95% CI [−0.07, 0.12].

In the Different Speaker block, the results are similar to the results in the Different Speaker block in Experiment 3A. Infants did not look at the highlighted object above chance either on Same Word trials, M = .58; $t(29)$ = 1.76, p = .088, Cohen's d = 0.32, 95% CI [.49, .68], or on Different Word trials, M = .50; $t(28)$ = −0.12, p = .91, Cohen's d = 0.02, 95% CI [.41, .58]. Also as in Experiment 3A, trial type had no effect, $M_{difference}$ = 0.07, $t(28)$ = 1.07, p = .295, Cohen's d = 0.20, 95% CI [−0.07, 0.21]. A 2 × 2 repeated-measures Anova on the subset of subjects who provided at least one valid trial of each type in both blocks (n = 29) revealed no main effects of block or trial type (block: $F(1, 28)$ = 2.82, p = .104,

η_p^2 = .09; trial type: $F(1, 28) = 1.65$, $p = .21$, η_p^2 = .06) and no interaction between block and trial type, $F(1, 28) = 0.008$, $p = .684$, η_p^2 = .006.

The results in the Word Knowledge phase indicate that infants knew the words applied to the geometric shapes in the Experimental phase, $M = .61$; $t(31) = 3.56$, $p = .001$, Cohen's $d = 0.63$, 95% CI [.55, .67].

EXPLORATORY: DIRECTION OF THE FIRST SACCADE AT TEST (PREREGISTERED)

As in Experiments 1, 2, and 3A, I also analyzed which of the two objects infants first oriented to at test. In the Same Speaker block, infants oriented first to the highlighted object above chance in Same Word trials, $t(31) = 3.09$, $p = .004$, Cohen's $d = 0.55$, 95% CI [.56, .78], but not on Different Word trials, $t(31) = 0.94$, $p = .354$, Cohen's $d = 0.17$, 95% CI [.43, .70]. In the Different Speaker block, infants did not orient first to the highlighted object above chance in any of the two trial types (Same Word trials: $t(29) = 1.53$, $p = .136$, Cohen's $d = 0.28$, 95% CI [.47, .73]; Different Word trials: $t(28) = 0$, $p = 1$, Cohen's $d = 0$, 95% CI [.35, .65]).

EXPERIMENTS 3A AND 3B TOGETHER

Analyzing the data from Experiments 3A and 3B together (henceforth, Experiment 3) reveals no three-way interaction (Trial Type × Block × Order: $p = .058$), no two-way interactions (ps > .144), and no main effects (ps > .091). Note, however, that interactions require much higher sample sizes than main effects (Gelman, 2018), especially as the effect in the Different Speaker block is not expected to flip but to be at chance.

Splitting the results by speaker identity (**Figure 3.9**) reveals that infants looked at the highlighted object above chance only on the Same Word trials of the Same Speaker block, $M = .60$, $t(59) = 3.70$, Cohen's $d = .48$, 95% CI [.55, .66] (all other trial types: ps ≥ .1). Trial type played a role in the Same Speaker block, $M_{difference} = 0.08$, $t(58) = 2.03$, $p = .046$, Cohen's $d = .27$, 95% CI [0.001, 0.16][5], but not in the Different Speaker block, where there was virtually no difference between the two trial types, $M_{difference} = 0.02$, $t(60) = 0.43$, $p = .666$, Cohen's $d = 0.06$, 95% CI [−0.06, 0.10].

[5] Experiment 3 replicates the trial type effect results in Experiment 1, $M_{difference} = 0.09$, $t(31) = 1.55$, $p = .132$, Cohen's $d = 0.27$, 95% CI [−0.03, 0.22], almost perfectly. This is evidence for the conjecture, raised in the **Discussion** to Experiment 1, that the reason there was no significant effect in Experiment 1 was insufficient power.

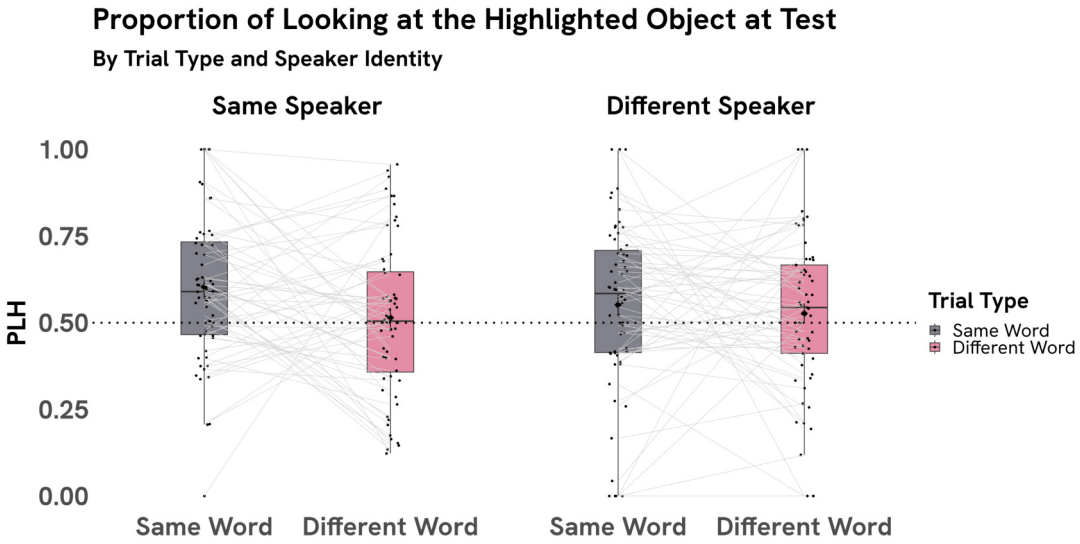

Figure 3.9. Results of Experiment 3, by trial type and speaker block (Experimental phase).

RELATION BETWEEN SPEAKER BLOCKS IN EXPERIMENT 3

As an additional analysis meant to investigate whether infants behaved in any way similarly in the Same Speaker and Different Speaker blocks, I obtained the trial type effects (Same Word – Different Word) for each infant in each of the two blocks. Figure 3.10 shows that even though the trials in the two blocks were identical (except for the identity of the speaker at test), there is no correlation between them, not even at the level of individual subjects (r = .09, p = .497). This provides additional evidence that infants did not treat the two blocks as equivalent.

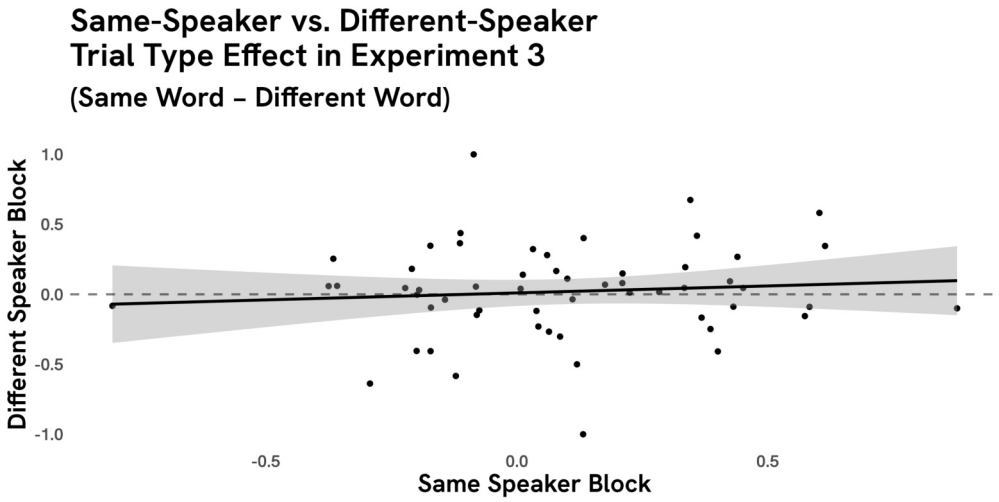

Figure 3.10. Correlation between trial type effects in Experiment 3 (*x*-axis: Same Speaker effects; *y*-axis: Different Speaker effects). Each dot represents trial type effects produced by individual subjects. The gray-shaded area represents the 95% confidence interval around the regression line.

3.5.3. Discussion

Experiment 3B, which presented infants with the Same Speaker block first, partly replicated Experiment 3A in that infants looked above chance at the highlighted object overall and in terms of their first saccade only in Same Speaker–Same Word trials. Where Experiment 3B differed was the absence of a difference between trial types in the Same Speaker block, suggesting that the order of blocks may have played a small role in infants' behavior. However, order alone did not drive the different results in Experiments 3A and 3B. If it had, infants' PLH scores in Experiment 3B would have been above chance on Same Word trials in the Different Speaker block, and there would have been a trial type difference in the second block, neither of which was the case. When analyzed together, the Same Speaker block in Experiments 3A and 3B replicated Experiment 1 almost exactly. By contrast, there was no difference between trial types in the Different Speaker block.

This pattern of results speaks against several alternative accounts for infants' interpretation of the predicative expression applied to the geometric shapes. First, if infants associated an uninterpreted phonological string with a shape, they should not have restricted this association to a single speaker. Second, if infants interpreted the predicative expression literally and learned, for instance, that the blue blob was, despite appearances, a dog, they should have had access to this information regardless of the person who queried it. Finally, the same argument can be leveraged against the hypothesis that infants created a new lexical entry for the known nouns. Infants did not infer, for instance, that "dog" is a homonym because there is no reason not to generalize a new lexical entry to a new speaker.

3.6. Comparisons Across Experiments 1–3

3.6.1. Bayesian Analysis

Bayesian modeling has two main advantages compared to frequentist approaches. First, it loses less information since there is no need to average over multiple trials within a subject. Second, because the models are generative, the posterior distributions can be queried indefinitely, and, therefore, the multiple comparisons problem in frequentist statistics no longer arises (McElreath, 2020).

As already noted, I preregistered a Bayesian growth curve analysis that incorporated information about infants' knowledge of the noun used during the highlight phase and their baseline preferences for the highlighted object (measured at the beginning of the trial, before stipulation took place). In the end, I dropped the anti-conservative growth curve model; instead, I devised several variations of a Bayesian linear model with the same structure but with PLH trial averages as the dependent variable—instead of gaze samples at multiple consecutive time points. The models assume that a test PLH data point on a given trial is a function of the experiment (1, 2, or 3), the trial type (Same Word or Different Word), and the block (Same Speaker or Different Speaker) on which it was produced, of the infant who produced it, of the infant's baseline PLH on that trial, and of the infant's knowledge of the label used during stipulation (as measured on Word Knowledge trials).

A model comparison based on the Widely Applicable Information Criterion (Watanabe, 2010) indicated that the best-fitting model, whose results I report in this section, is the one that groups condition into six levels: two for the Same Speaker block in Experiments 1 and 3 (one per trial type); two for Experiment 2 (one per trial type), and two for the Different Speaker block in Experiment 3 (one per trial type). The details of the modeling process can be found in Appendix A.

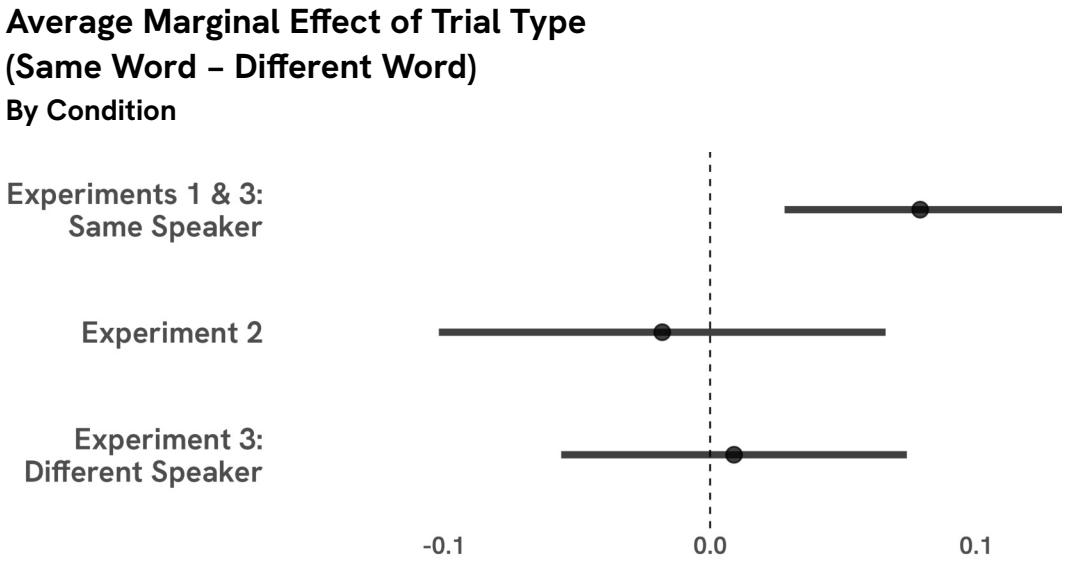

Figure 3.11. Posterior estimates for the average marginal effect of trial type (Same Word – Different Word) by condition. Points represent the means of the posterior distributions; horizontal lines depict the 89% credible interval around the mean. Dashed vertical lines mark the null value for ease of legibility.

Figure 3.11 plots the effect of trial type in each condition, averaged over all possible values for baseline preference and word knowledge. As predicted, trial type matters when the speaker delivering the test question is the same one who performed the stipulation and when the on-screen shapes look different (Experiment 1 and Experiment 3 Same Speaker block), $M_{difference}$ = 0.08, 89% credible interval [0.03, 0.13]. There is no effect of trial type either in Experiment 2, $M_{difference}$ = −0.02, 89% credible interval [−0.10, 0.07], or in the Different Speaker block of Experiment 3, $M_{difference}$ = 0.01, 89% credible interval [−0.06, 0.07].

The posteriors for the word knowledge and baseline coefficients for all six conditions are plotted in Figure 3.12. Word knowledge does not play a significant role in any condition (all 89% credible intervals include 0), whereas baseline preferences matter only in the conditions where infants were at chance. Baseline preferences moderately influence test PLHs in Experiment 1–Different Word trials, and strongly influence test PLHs in the Different Speaker block of Experiment 3, in both trial types. This indicates that infants revert to their baseline preferences when not knowing what to look at in response to the test question.

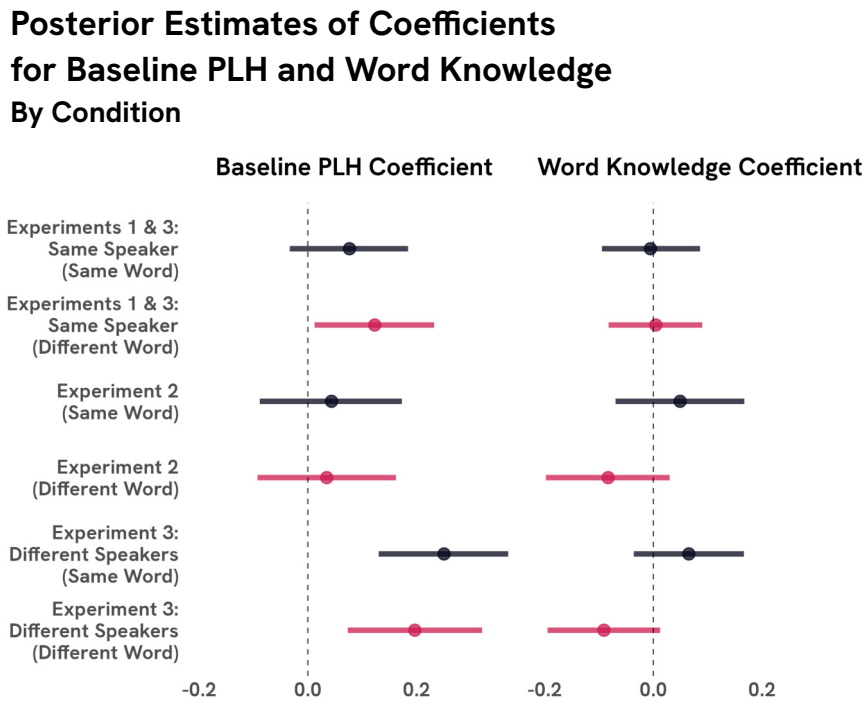

Figure 3.12. Posterior estimates for the Baseline PLH and Word Knowledge coefficients by condition. Points represent the means of the posterior distributions; line intervals depict the 89% credible interval around the mean. Dashed vertical lines mark the null value for ease of legibility.

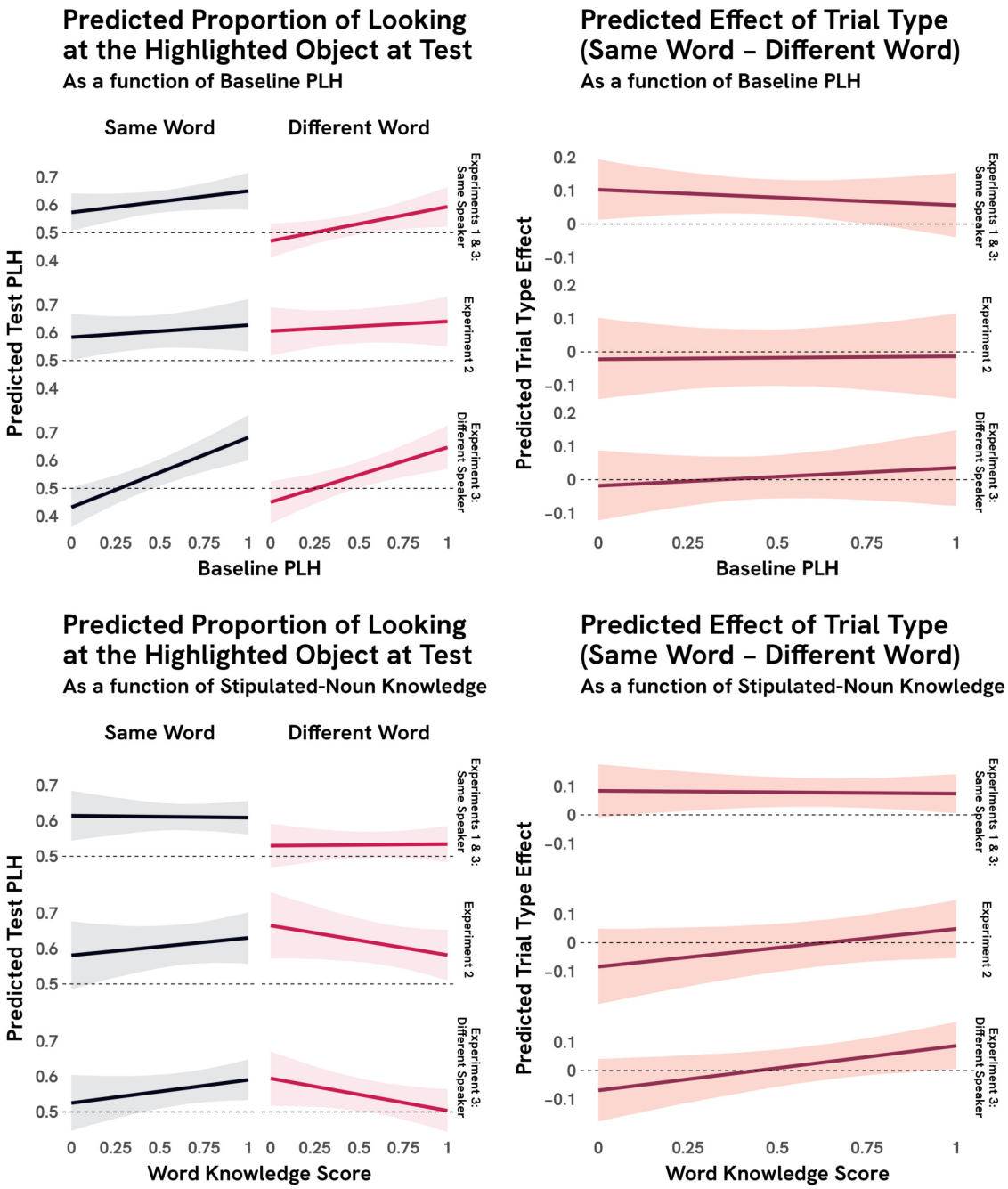

Figure 3.13. Predictions implied by the posterior distributions obtained for Experiments 1–3. Top: Predicted test PLH (left) and predicted effect of trial type (right) as a function of baseline PLH. Bottom: Predicted test PLH (left) and predicted effect of trial type (right) as a function of Word Knowledge score.

To better grasp the influence of these predictors, I simulated data from the posterior distributions for the entire range of values that baseline preference and word knowledge could take (i.e., the 0–1 interval). I then computed the predicted test PLHs and the predicted effects of trial type across that range. The results are plotted in Figure 3.13 and exhibit an elegant pattern. In the Same Speaker block of Experiments 1 and 3, the predicted above-chance test PLHs on Same Word trials and the trial type effect are robust to fluctuations in baseline preferences and word knowledge scores. It is only when the baseline preference for the highlighted object is very high that the effect of trial type is masked (Figure 3.13, top right). This differs from the Different Speaker block, where these factors strongly influence both measures. In addition, trial type has no effect in the Different Speaker block, regardless of baseline preferences and word knowledge (Figure 3.13, top and bottom right).

3.6.2. Discussion of Experiments 1–3

Experiments 1–3 show that infants have no problem assigning familiar nouns to visual objects that do not belong to the kinds denoted by the nouns. One possible exception is Experiment 2, which could be taken to show that infants reject the stipulation when there is no contrast between the two shapes. Alternatively, infants may have accepted the stipulation in Experiment 2 as well but generalized it to the other, same-looking symbol. I come back to this in the General Discussion.

The Different Speaker block in Experiment 3 shows that the assignment is restricted to the speaker who stipulated it. In addition, the effect holds regardless of whether the infant knows the word, further indicating that they do not interpret the experimental situation as a word-learning one. Finally, in all conditions in which infants looked at the highlighted object at chance levels, baseline preferences predicted their looking behavior at test better than in the conditions in which they looked at the highlighted object above chance.

Experiments 1–3 thus rule out several alternative explanations for infants' interpretation of the stipulation events in the current paradigm. However, there are still several accounts that the data cannot yet adjudicate between. Infants may still interpret the labeling events as revealing an underlying property of the speaker (incompetence or malevolence) or as introducing the speaker-specific stipulation that the shapes should be referred to by the labels applied to them

(referential pact). Both of these can explain why infants did not generalize the mapping to a new speaker. To test these accounts, Experiment 4 asked whether infants interpret the nouns under their ordinary meanings. That is, Experiment 4 investigated whether infants interpret "a dog" applied to a blue blob as inducing certain expectations about the subsequent behavior of the blue blob. If infants think the speaker is incompetent or malevolent, they should not generate any kind-related expectations. If infants believe the speaker is stipulating a referential pact (e.g., the local convention that blue blobs are to be called "dogs"), they should not have any such expectation either[6].

3.7. Experiment 4: Moving Symbols

One of the first distinctions that arise in infants' developing ontology is the distinction between agents and inert objects (e.g., Bonatti et al., 2002), and it is widely believed that representations of agents and representations of objects reside in separate cognitive mechanisms (for a review, see Carey, 2009). In short, agents are entities that act efficiently on the world to bring about goal changes (Gergely et al., 1995; Liu et al., 2017), while objects are mid-sized bodies that travel on continuous paths, preserve cohesion in motion, and cause changes in other objects only upon contact (for a review, see Spelke, 2022).

Experiment 4 asked whether infants recruit conceptual knowledge when setting up local assignments between symbol objects and discourse referents. To test this, infants were shown displays of the same shapes as in Experiments 1–3, except both shapes were highlighted. One of the shapes was labeled with a noun denoting an animate kind (e.g., "dog"), while the other was labeled with a noun denoting an inanimate kind (e.g., "shoe"). After the two stipulations, one of the shapes moved toward the other until reaching it. If infants distinguish the shapes based on the nouns heard during the stipulation, they should distinguish the trials in which animate referents move toward inanimate referents from the trials in which the opposite occurs.

If infants' conceptual system includes information about the animacy of entities in their environment and if infants recruit this information when interpreting external representations, they should have different expectations de-

[6] Experiment 4 also tests a variant of the **association** account, whereby the labeling events in Experiments 1–3 give rise to three-way associations between a label, an object, and a speaker.

pending on whether a symbol stands for an animate versus an inanimate entity. Previous work indicates (i) that infants younger than those tested here possess an early superordinate animal concept (Mandler & McDonough, 1993); (ii) that they expect animals, but not vehicles, to move on a straight path to their goals (Baker et al., 2014); and (iii) that they prefer to model movement actions on animals over vehicles (Rakison et al., 2007).

3.7.1. Methods

TRANSPARENCY AND OPENNESS

The hypotheses, methods, and data analysis for Experiment 4 were preregistered at the Open Science Framework (https://osf.io/jv79e).

PARTICIPANTS

The final sample consisted of 32 typically developing Hungarian-speaking[7] 14–16-month-old monolingual infants (M_{age} = 15 months 23 days, SD_{age} = 24.3 days). Results from a pilot with eight participants (first looks: Cohen's d = 1.35; total looking time: Cohen's d = .56) indicated that 28 participants would be enough to detect an effect of looking time with 80% power. An additional eight infants were tested and excluded due to fussiness (n = 6), technical error (n = 1), or maxing out on seven out of the eight trials (n = 1).

APPARATUS

Infants' gaze was collected using a Tobii T60XL with an integrated 23.8-inch-diagonal monitor (resolution: 1920 × 1080; refresh rate: 60 Hz). External speakers delivered the sound. A custom-made Python program building on PsychoPy 2021.1.3 (Peirce et al., 2019) was used to calibrate the infants, present the stimuli, and collect the eye data.

STIMULI

Two sets of visual stimuli were used: eight color photographs representing eight kinds of objects (**Figure 3.14**, top) that are familiar to Hungarian-speaking infants of this age (Parise & Csibra, 2012; Pomiechowska et al., 2021) and four pairs of geometric shapes (**Figure 3.14**, bottom). Each of the four photographs depicting animate entities (duck, bear, dog, and cat) had a horizontally flipped counterpart,

[7] The switch from an Austrian to a Hungarian sample was made for logistic reasons: it is easier to find monolingual infants in Budapest than in Vienna.

to have the eyes of the agent oriented toward the other object on the screen. As in Experiments 1–3, the objects' bounding boxes were matched in height and, whenever possible, in width; their display size was approximately 330 × 330 pixels. The stimuli also included an image of a pointing hand, displayed at 213 × 366 pixels, as in Experiments 1–3.

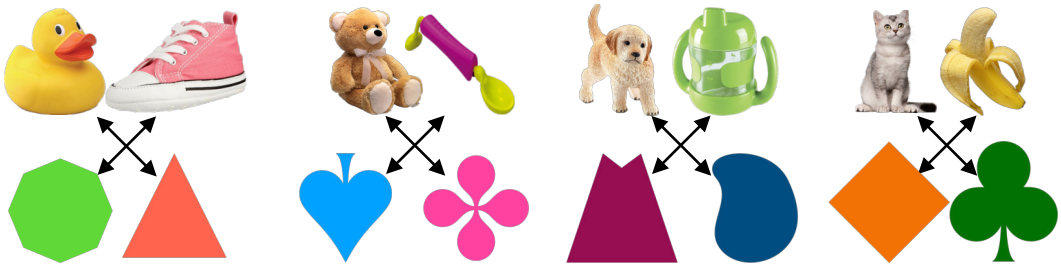

Figure 3.14. Visual stimuli used in Experiment 4. Top: eight photographs of objects belonging to kinds that are familiar to 15-month-old infants. Bottom: four pairs of symbol stimuli sampled from the ones in Experiment 1. The animate-inanimate pairs were fixed, as were the pairings between photo and symbol couplets. Which symbol stood for which entity within a given pair was counterbalanced across subjects.

Audio stimuli were the eight nouns corresponding to the familiar kinds in the photographs, embedded in different carrier phrases: "Hi baby! Look! An X! Here's an X! Wow, an X!", "Where is the X? X!". A female native speaker of Hungarian recorded the sound stimuli in infant-directed speech.

PROCEDURE

Infants were shown animated clips while seated on their caregivers' laps, as in Experiments 1–3. The experiment consisted of eight trials, split into four Training and four Experimental trials.

As in Experiments 1–3, Training trials (1–4) were meant to familiarize the infant with the general procedure and to give them evidence that the voice they hear is connected to what is happening on the screen. The overall structure of a trial was very close to that in Experiments 1–3 (Figure 3.15). A trial consisted of four parts: **Stipulation 1**, **Stipulation 2**, **Movement**, and **Test Measurement**. Each trial started with a blue curtain covering the entire display. An attention-getter appeared in the center of the screen and rotated until the infant oriented to it for 500 ms. The curtain then went up to reveal a static display of two object

photographs, one on the left and one on the right of the screen (e.g., a cat and a banana). After 2 seconds of silence, the first Stipulation part started and unfolded as in Experiments 1–3. An animated hand appeared above one of the two objects (e.g., the cat), pointing to it. The hand moved up and down while the infant was greeted ("Hello baby! Look!") to draw their attention to the object. The hand stopped above the object, and infants heard the word typically used to refer to it three times in different carrier phrases ("A cat! Here's a cat! Wow, a cat!"). After another 2 seconds of silence, the second Stipulation part started, and the procedure was repeated on the other object on the screen.

After another two-second break, the infants' attention was drawn to the screen by the speaker ("Look!"), after which one of the objects started moving toward the other object, on a straight path and at uniform velocity. On Congruent trials, the moving object belonged to the animate kind. On Incongruent trials, it belonged to the inanimate kind. After reaching the stationary object, a ding sound played, and the moving object wiggled for 666 ms (by rotating left and right around its vertical axis in two cycles). After the wiggling stopped, looking times were measured until infants looked away from the screen for 2 seconds without looking back to the screen or until 30 seconds passed. Each trial lasted between 30 seconds (without the test period) and a minute (with the test period).

Experimental trials were identical to Training trials except that the two object images were replaced by geometric shapes (**Figure 3.15**, bottom). Experiment 4 did not include the Word Knowledge phase in Experiments 1–3 because previous work already showed that Hungarian infants in this age range are familiar with the nouns used here (Parise & Csibra, 2012; Pomiechowska et al., 2021) and because the duration of the trials in Experiment 4 was up to twice as long as that of Experiments 1–3.

DESIGN

The experiment had a within-subjects design with one independent variable, **Trial type**, and two levels, **Congruent** and **Incongruent**. Each infant was administered the following trial alternation: ABBA-BAAB (Training–Experimental), with type of first trial (Congruent or Incongruent) counterbalanced across subjects. This alternation implies that an Incongruent trial in the Training part becomes a Congruent trial in the Experimental part. Several reasons drove this choice. First, because looking times often exhibit order effects, trial type alternated according to the ABBA pattern. Second, exposing infants to Congruent trials in the

Training phase was undesirable because an effect in the Experimental phase could have been attributed to the Incongruent trials being flipped across the two phases. Alternatively, Experiment 4 could have used two animate objects in the Training trials and two different animate objects in the Experimental trials. But in that case, it would have lost the power given by having four trials in the Experimental phase. In the **Results** section, I show that this counterbalancing choice does not affect the results.

The side of the object labeled with an animate noun and the side of the object labeled first were counterbalanced across subjects. The animate–inanimate object pairings and the visual symbol pairings were fixed (**Figure 3.14**), but the symbol–label pairings within each pair of symbols were counterbalanced (e.g., for half of the participants, the octagon was labeled as a duck, and the triangle as a shoe; for the remaining half, the mapping was reversed). The pair succession cycle was fixed across subjects (duck–shoe, bear–spoon, cat–banana, dog–sippy cup), but the identity of the pairing shown in the first trial was counterbalanced across subjects. Thus, across subjects, each pair appeared an equal number of times in each of the eight serial positions.

CODING

Infants' looking times were coded online by the experimenter who tested them, naïve to the research hypothesis. The experimenter pressed a key whenever the infant looked away from the screen (if the key was pressed for 2 seconds uninterruptedly, the trial ended). The key presses were recorded in the data file obtained for each infant at the end of the test session. Looking times were also coded offline based on the test session video recordings by myself. Inter-rater reliability was substantial (total looking time: Spearman's ρ = .96; first looks: Spearman's ρ = .89). I double-checked all the trials (naïve to trial type) in the Experimental phase for which the coding difference was higher than 2 seconds. Without these trials (which turned out to have been wrongly coded online), inter-rater reliability exceeded .95 for both measures. The analyses are based on the offline coding.

DATA EXCLUSION

The data exclusion criteria were preregistered. Experimental trials were excluded if infants attended the screen less than 40% of the time during the movement phase (n = 4: two Training and two Experimental trials). Two Experimental and

two Training trials were ended too early by the experimenter and also had to be excluded. Finally, two Training trials were compromised by a script error and were excluded as well. All infants who made it to the Experimental phase without fussing out at Training completed the experiment.

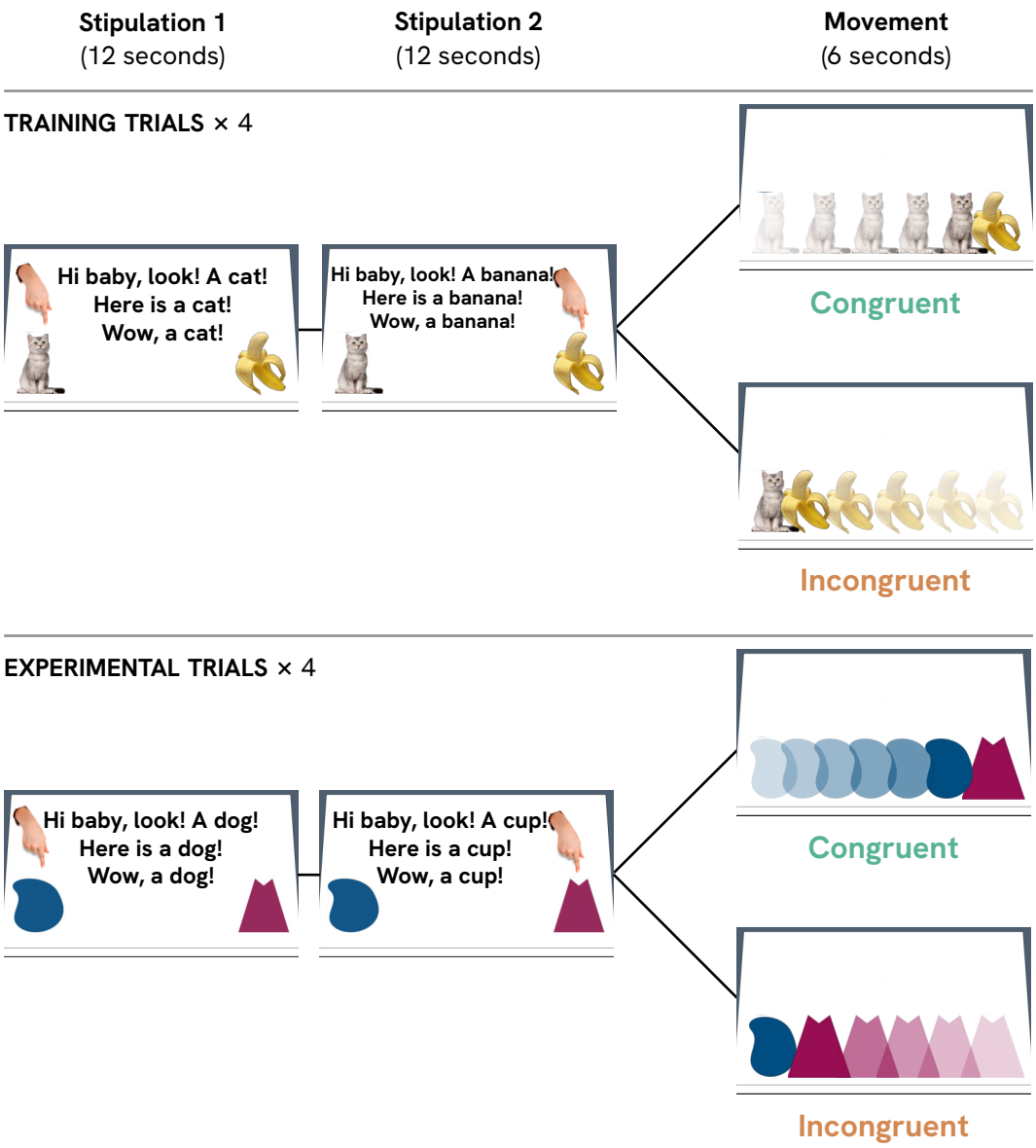

Figure 3.15. Trial sequence for each phase and trial type. In each trial, both objects on the screen were labeled with an animate-kind and an inanimate-kind denoting noun, respectively. On Congruent trials, the animate object moved toward the inanimate one until reaching it; on Incongruent trials, the opposite occurred. Once movement ceased, infants' looking time measurement began.

MEASURES

I preregistered two main analyses, both for the Experimental phase: paired two-tailed *t*-tests for total looking times and first looks (Manea et al., 2023; Newcombe et al., 1999; Yoon et al., 2008). Total looking time measures the amount infants spent looking anywhere on the screen from the first frame after movement stopped until they looked away for 2 seconds without looking back to the screen or until 30 seconds passed. First looks measure the amount infants spend uninterruptedly looking on-screen before disengaging the first time. Looking times were $\log_{10}$-transformed before the analysis (Csibra et al., 2016).

I predicted that infants would look longer to the screen on Congruent than on Incongruent trials. This predicted direction of the effect may seem counterintuitive, given that infants often look longer at incongruent stimuli (although not always: Hernik et al., 2014). However, the pilot showed the effect to hold in this direction rather robustly, in both the Training and Experimental phases and for both total and first looks. I come back to this reversal in the **Discussion**.

3.7.2. Results

LOOKING TIMES

Figure 3.16 plots total looking times and first looks by trial type and phase. As predicted, infants looked longer at the screen on Congruent trials than on Incongruent trials in the Experimental phase. However, this effect was significant only on first looks, $M_{\text{difference}}$ = 0.08 (1.8 seconds on the original scale), $t(31)$ = 2.53, p = .021, Cohen's d = 0.43, 95% CI [0.01, 0.15]. Total looking times exhibited a similar but not as strong a pattern, $M_{\text{difference}}$ = 0.05 (1.5 seconds on the original scale), $t(31)$ = 1.45, p = .157, Cohen's d = 0.26, 95% CI [−0.02, 0.13]. In the Training phase, infants did not exhibit an effect of trial type on either looking-time measure (first looks: $M_{\text{difference}}$ = 0.01, t(31) = 0.11, p = .914, Cohen's d = 0.02, 95% CI [−0.08, 0.09]; total looking time: $M_{\text{difference}}$ = 0.015, t(31) = 0.31, p = .76, Cohen's d = 0.05, 95% CI [−0.08, 0.11]).

EXPLORATORY: BAYESIAN MODEL (PREREGISTERED)

As an additional exploratory analysis, I preregistered a Bayesian generalized linear model for analyzing infants' looking times, identical to the one used in Manea et al. (2023). The model assumes that log-transformed and standardized first looks are sampled from a noisy normal distribution, whose mean is a linear function of the subject providing that data point, the phase and trial type from

which that data point came, the trial pair (1—for Trials 1–2 and 5–6; 2—for Trials 3–4 and 7–8) and the trial type order in that pair (Congruent-first versus Incongruent-first). The details of the model can be found in **Appendix A**; the fully reproducible code can be accessed at https://osf.io/x3naq/files/osfstorage. The model confirms that infants looked longer to the screen in Congruent than in Incongruent trials, M_{effect} = 0.25, 89% credible interval [0.02, 0.49].

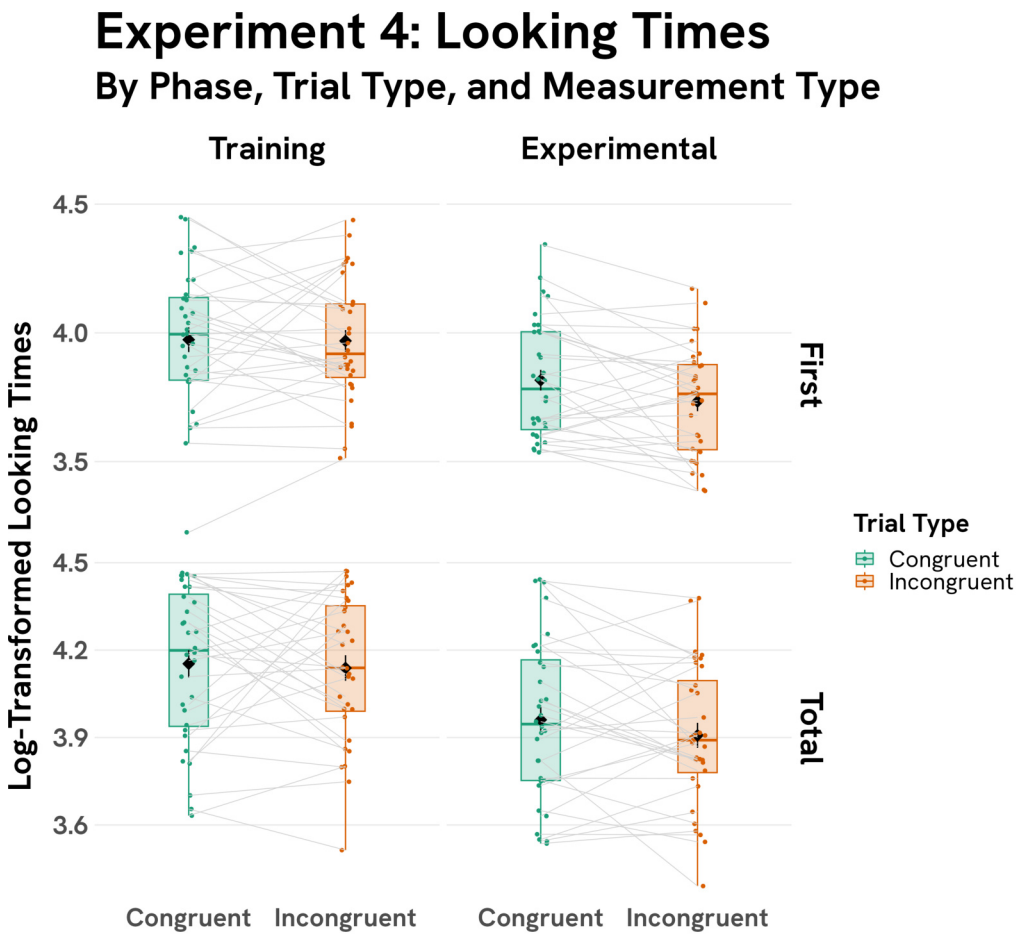

Figure 3.16. Results of Experiment 4, split by trial type and phase. The *y*-axis plots $\log_{10}$-transformed millisecond looking times. A score of 4 on the log-scale corresponds to a looking time of 10 seconds. Black circles and the lines connecting them represent individual averages as a function of trial type and phase; black diamonds depict group averages ±1 SEM; boxplots indicate the median and interquartile range.

In addition, I tested whether the choice to flip congruency for each pair from the Training to the Experimental phase might explain the results. Suppose there was an effect of trial type in the Training phase (e.g., Incongruent > Congruent). Suppose infants also remembered which object in a pair moved toward the other at Training. In that case, the effect may have carried over to the Experimental phase simply because of the movement reversal, which would have surprised infants equally in both trial types. Thus, a preference for one type of trial at Training and an additive effect of movement reversal might have caused a spurious effect in the Experimental phase.

To test this, I added the first looks obtained in the Training phase for all object pairs as a predictor to the Bayesian model above (e.g., looking time to the Training cat-banana trial was added as a predictor for the Experimental cat-banana trial). The posterior distribution of the effect in the Experimental phase continues to exclude 0 as a plausible value, M_{effect} = 0.26, 89% credible interval [0.03, 0.48]. By contrast, the posterior distribution of the coefficient of Training first looks does not, $M_{\text{training coefficient}}$ = −0.07, 89% credible interval [−0.22, 0.08].

EXPLORATORY: LOOKING TIMES BY PAIR (PREREGISTERED)

Splitting the data by phase and trial pair (1 versus 2) reveals an effect of trial type only in the second pair, in both phases: first looks in the Experimental phase (Congruent > Incongruent), $M_{\text{difference}}$ = 0.11, $t(29^{8})$ = 2.28, p = .03, Cohen's d = 0.42, 95% CI [0.01, 0.20]; total looking times in the Training phase (Congruent > Incongruent), $M_{\text{difference}}$ = 0.11, $t(29)$ = 2.09, p = .046, Cohen's d = 0.40, 95% CI [0.002, 0.22]. No other comparison was statistically significant (all ps > .15).

EXPLORATORY: PUPIL SIZE (PREREGISTERED)

I also analyzed infants' pupil sizes in the Experimental phase during the movement part of a trial to check for additional evidence that infants process the scene differently depending on the type of entity that moved in a self-propelled way (animate vs. inanimate). There was no effect of trial type, $t(31)$ = 0.54, p = .593, Cohen's d = 0.10, 95% CI [−0.03, 0.06].

---

[8] There are fewer degrees of freedom because not all infants provided data for both trials in the second pair.

EXPLORATORY: ADDITIONAL SUBJECTS

Seven additional babies were overbooked and were tested with counter-balanced experimental orders. To check the robustness of the first-looks effect in the Experimental phase, I reran the first-looks analysis on a dataset including these babies. The effect of trial type increased, $M_{\text{difference}}$ = 0.09, $t(29)$ = 3.10, p = .004, Cohen's d = 0.50, 95% CI [0.03, 0.16], indicating that the finding obtained with the main sample was not a spurious one.

3.7.3. Discussion

In the Experimental phase, infants looked longer at the screen without interruption when the animate-labeled geometric shape moved toward the inanimate-labeled geometric shape than when the opposite movement occurred. The results of Experiment 4 thus indicate that 15-month-old infants recruit their conceptual knowledge when faced with stipulation events, such that they distinguish symbols that stand for animate entities from symbols that stand for inanimate ones. The results also go a long way in ruling out several alternative explanations for the data in Experiment 1. Experiment 4 suggests that infants do not consider the speaker incompetent or malevolent and do not interpret the predicative expression as a referential pact. Nevertheless, two aspects of the findings in Experiment 4 are intriguing. I discuss them in turn.

THE DIRECTION OF THE EFFECT

First, there is the direction of the effect, which goes in the opposite direction from what is usually found in infant looking-time studies. While the direction of the effect was predicted, I admit that the prediction was based solely on the pilot, for which I also expected longer looking times on Incongruent trials. Nevertheless, both the pilot and the main experiment showed the effect in the same direction, suggesting that the directionality is robust. Why would it go this way? Unfortunately, I know of no comparable looking-time paradigm that I could draw on in answering this question, so I can do no better than speculate.

Attending a stimulus, for which looking time is a proxy, primarily reflects the cognitive utility of that stimulus (Kidd et al., 2012; Sperber & Wilson, 1995). In turn, while cognitive utility sometimes covaries with surprise (Sim & Xu, 2019), especially in creatures who must learn so much about their environment, it may also reflect other internal variables, such as the anticipation of forthcoming events (e.g., Hernik et al., 2014). In this case, infants may have expected that

more events would occur when agents moved toward inanimate objects (e.g., if the dog moved toward the spoon, it might do something with it, but not vice versa) and therefore looked longer on Congruent trials. Infants could also have expected that representations consistent with reality would convey more relevant information than inconsistent ones.

Alternatively, infants' preference for Congruent trials is related to stimulus complexity. Computational work on infants' looking behavior indicates that infants' probability of losing interest in a stimulus takes a U-shaped form as a function of stimulus complexity (Kidd et al., 2012). Thus, infants' probability of looking away is high for simple and complex stimuli and low for stimuli of intermediate complexity. In addition, the model predicts that infants will look longer at familiar/congruent stimuli when the stimuli are of low or high complexity and at novel/incongruent stimuli when the stimuli are of intermediate complexity.

I do not know of any metric that could measure the complexity of the stimuli in Experiment 4. However, it seems possible that infants showed a preference for Congruent trials due to the higher complexity of Incongruent trials in the Experimental phase. Specifically, keeping track of two STAND-FOR relations—and thus of four items—may have overtaxed infants' working memory. Unlike in the Training part, where visual cues sufficed for tracking the identities of the objects, infants could only rely on their memory to do this in the Experimental phase. Because the stipulation and the motion conflicted in Incongruent trials, infants may have been confused about which symbol stood for which referent, causing them to disengage faster.

The only study I know that investigated a question comparable to the one tested in Experiment 4 is Onishi et al.'s (2007) series of experiments, which tested whether 15-month-old infants detect violations in pretense actions. Onishi et al. exposed infants to a live actor pretending to pour a liquid from an empty jug into a cup. At test, infants looked longer when the actor pretended to drink from another cup (incongruent trials) than when she pretended to drink from the one she pretended to pour liquid into (congruent trials). I do not have a good explanation for this contrast. There are several differences between Onishi et al.'s (2007) study and Experiment 4 (e.g., the presence of a live experimenter, the experimenter's active role in the pretend world, the use of imaginary substances), but none that would explain the contrast in a principled way.

Regardless of which of these accounts turns out to be true, infants must have interpreted the noun phrases under their regular meanings and linked them to the visual shapes, as guided by the pointing hand. Otherwise, they would not have distinguished between the two trial types in the Experimental phase.

THE LACK OF AN EFFECT IN THE TRAINING PHASE

The second puzzling aspect concerns the Training phase. Although this part was meant to familiarize infants with the general procedure and although I did not preregister an analysis for it, the absence of an effect is worth discussing. Again, this is consistent with the pattern of results found in the Pilot, where the effect was more robust in the Experimental phase (first looks: Cohen's d = 1.35; total looking times: Cohen's d = 0.57) than in the Training phase (first looks: Cohen's d = 0.13; total looking times: Cohen's d = 0.56).

One possibility is that the lack of an effect stems from a practice effect. The paradigm did not include any pre-familiarization stage, so the task plunged infants into the Training trials. Because of this, it may have taken them a while to figure out what was going on in the animations, which could have masked the effect of trial type. Two pieces of evidence support this account. First, there was an effect on looking times in the second pair of trials in the Training phase (Trials 3 and 4) in the same direction as the one found in the Experimental phase. Second, even in the Experimental phase, the effect on looking times is only found in the second pair (Trials 7 and 8). This suggests that infants need time to habituate to the stimuli (to the overall pattern in the Training phase and to the novel geometric shapes in the Experimental phase) before differentiating between Congruent and Incongruent trials.

An experiment that switches the order of phases or one augmented with pre-familiarization trials with the stipulation part removed could test this explanation. However, neither suggestion is unproblematic. If the Experimental phase comes first (without any pre-familiarization stage), infants may take the speaker to be unconnected to the visual events on the screen or speak a different language. (This was why a Training phase prefaced all experiments reported here.) If that were the case, a null result would be impossible to interpret. If, on the other hand, a pre-familiarization stage using moving photos were included, one would either need to augment the set of visual stimuli (and 15-month-olds may not know many more animate entities than the ones tested here) or recycle the stimuli used during pre-familiarization in the Training phase, which may bias the

measurement (e.g., infants may look longer on Congruent trials because they are similar to the ones they saw in pre-familiarization).

The more interesting possibility for the lack of an effect in the Training part involves not the order of presentation but the difference between photographs and shapes. Except for the cat, the photographs used for the animate entities were not photographs of real animate beings but of toys. I was aware that this might be problematic, but two reasons weighed toward this decision. First, previous work has shown that Hungarian infants recognize the same stimuli used here by their labels (Parise & Csibra, 2012; Pomiechowska et al., 2021). Second, I assumed it was more likely for infants to have experienced rubber ducks than real ducks and (hopefully) teddy bears than real bears and thus to recognize them on-screen. Since infants encountered these entities only in their inert, symbolic form, they may not have learned yet that these entities are animate. However, if infants interpret them as symbols and if they assume that toys represent their referents iconically, the agency cues that these toys possess (e.g., eyes) may be enough for infants to infer that bears belong to an animate kind even though they learned the noun for bears in the context of teddy bears.

The result in the Experimental phase confirms that they did. This is surprising, given how much infants can learn from picture books without any contact with the entities the picture books depict (e.g., Simcock & DeLoache, 2006). Nevertheless, toy photographs are ambiguous (see **Chapter 4**), and infants could have interpreted the photographs as standing for toys. Because infants know that toys are inert objects, they might not have expected them to move either[9]. Unfortunately, any statistical analysis would have lost the within-subjects comparison because three out of four object pairs contained toys.

A further problem could have been the mechanics of movement, which consisted of animating the images such that they appeared to move in a straight line. This may have been odd in the photograph trials because this is not how these entities actually move. For instance, infants may have been dumbfounded by the fact that agents moved while their limbs were static or that agents approached the other object without reorienting their gaze to it. Shapes, on the

[9] Note, however, that this explanation requires positing one of two post-hoc assumption that need independent testing. As infants did show an effect in the Experimental phase, this explanation must assume either that visual cues trump linguistic stipulation when setting up STAND-FOR relations or that linguistic stipulation works differently depending on the nature of the symbol ("bear" means teddy bear when applied to a teddy bear but bear when applied to a shape).

other hand, may be better suited to convey motion in this minimal way precisely because they abstract away from these details.

So far, I have considered reasons for which infants could have found the movement of the animate-labeled entities unexpected. Alternatively, it could be that the movement of the inanimate-labeled entities was, for some reason, more expected in the Training phase than in the Experimental phase. While infants may know that only bananas, shoes, sippy cups, and spoons are inert objects, they have experienced the motion of all object types instantiated by the stimuli (e.g., when their caregivers bring these objects to the infants[10]). On trials with photographs, if infants' cognitive processes driven by conceptual knowledge (e.g., only animate beings self-propel) are active in tandem with other processes relying on infants' perceptual experience (which includes motion of both types of objects), the latter could mask the effect of trial type. By contrast, in trials with shapes, previous perceptual experience plays no role, which might allow the infants to focus on the discrepancy between what they see and their conceptual knowledge. This aligns with previous research findings indicating that it is sometimes easier for infants to interpret schematic stimuli than realistic ones (Cohen & Amsel, 1998; Oakes & Cohen, 1990).

Finally, it is possible that infants disengaged faster only on Experimental Incongruent trials because that was the only trial type in which they could not keep track of the on-screen object identities. As already noted, infants had several ways to track identity in the Training phase (label and photo), as opposed to the Experimental phase (label only). This ties back to the discussion of the directionality of the effect and can explain both puzzles at once. A version of this experiment, in which the moving object returns to its original position after reaching the stationary object, could test this by asking the infants where one of the objects is, as in Experiments 1–3. If infants were at chance on Incongruent trials with geometric shapes but not with photographs, this would show that infants cannot track the identities of the shapes if their behavior is incongruent with the labels they received during stipulation.

That said, the account for the lack of an effect in the Training phase that turns out to be correct bears no issue on the finding from the Experimental phase, which the pilot data, the main experiment, and the additionally tested subjects show to be robust. Infants look longer to animations in which animate

[10] I am indebted to Laura Schlingloff for this point.

discourse referents move toward inanimate ones than vice versa. In the absence of an alternative explanation for this result, I interpret the data as showing that infants interpret the spatiotemporal relations that discourse referents enter in a way that takes into account the identities of the referents along with infants' conceptual knowledge about the kinds to which the discourse referents belong.

3.8. General Discussion

Across five experiments that tested a large number of subjects (n = 160), I investigated whether 15-month-old infants understand that arbitrary geometric shapes can stand for familiar discourse referents. The results suggest that infants can set up STAND-FOR relations between a visual object and an instance of a familiar kind (Experiments 1, 3, and 4). These relations are local to the discourse, as they are not generalized across speakers (Experiment 3), but they are not merely referential pacts, as they recruit infants' conceptual knowledge in interpretation (Experiment 4).

IMPLICATIONS FOR EXISTING WORK

In retrospect, one might take the findings reported in this chapter as a foregone conclusion. After all, this is precisely what 15-month-olds will start doing only three months later when engaging in object substitution pretense themselves. In another sense, however, the fact that infants understood the task in a manner akin to pretend play even when the context was stripped down of many of the cues that accompany pretense is remarkable. This suggests that the cognitive mechanism underlying the representation of STAND-FOR relations is not only available but also easily activated.

The results have theoretical and methodological implications for the investigation of other cognitive phenomena in development. First, they seem to indicate (i) that object substitution pretense does not require the cues that have been proposed to help toddlers distinguish pretend from other kinds of behavior (e.g., smiling: Lillard & Witherington, 2004); and (ii) that object substitution pretense is not a special cognitive capacity (Harris & Kavanaugh, 1993; Leslie, 1987). Instead, infants (at least from 14–16 months onward) can set up local assignments between arbitrary objects and familiar discourse referents, and this capacity manifests itself under many guises, one of which is pretend play.

Second, the findings speak to how infants' and children's interpretation of labeling events has been characterized in the literature. Take the nature of reference, for instance. It is widely assumed that infants understand at some point that words refer to objects, and there are several proposals for how this comes about (e.g., Frank et al., 2009; Luchkina & Waxman, 2021). For instance, according to a recent theoretical proposal (Luchkina & Waxman, 2021), reference is a three-way link between a word (e.g., "a dog"), a mental representation (e.g., of a dog), and a mind-external object (e.g., the dog in question). The findings presented here suggest that the story is bound to be more complex.

On the one hand, infants can interpret predicative expressions as establishing a relation between a mind-external object (e.g., the blue blob on the screen) and a distal discourse referent belonging to the kind denoted by the noun (e.g., a dog). Instead of a three-way link between a word, a mental representation, and a mind-external object, there is a two-way relation between a perceptually available object (the blue blob) and a distal discourse referent (the dog) whose existence in the world is not at issue. In these cases, the external object is the one that refers to something else rather than the one that is being referred to. Moreover, infants can interpret nouns (e.g., "A dog!") as referring not to external objects but to properties that infants have concepts for[11] (e.g., the property of being a dog). And they use these concepts to generate descriptions under which the referents are brought. Modeling reference as a three-way link between words, objects, and mental representations fails to capture these niceties.

Third, there are two experimental paradigms for which these findings are directly relevant: mislabeling events (e.g., Csink et al., 2021; Dautriche et al., 2021; Koenig & Woodward, 2010) and referential pacts (e.g., Matthews et al., 2006; Matthews et al., 2010). I illustrate this point with mislabeling studies, but the same argument applies to referential pacts. In experiments investigating mislabeling, infants or children are exposed to an adult speaker who consistently mislabels everyday objects, and researchers measure infants' and children's subsequent inferences about the speaker (e.g., whether they refrain from learning new words from unreliable speakers). In the present experiments, mislabeling occurred as well, yet infants did not make any inference about speaker reliabili-

[11] It is of course possible that 15-month-olds thought the shapes stood for actual objects in the world, but this seems unlikely given that they succeed at connecting representations to particular states of affairs much later (e.g., DeLoache, 1987).

ty. Else, they would have rejected the mappings altogether and would not have shown a preference for trials in which animate entities move in Experiment 4.

Given the present findings, mislabeling events (and referential pacts) may often be interpreted as stipulating STAND-FOR relations. This would explain why children are less likely to learn a new word in these contexts. They know they are not in a word-learning situation, so they will restrict the mappings to the current discourse without inferring anything about speaker reliability. I also think infants and children may sometimes interpret the mislabeling events not as linguistic or ontological shortcomings on the speaker's behalf but as pragmatic ones—by relying on inferences about the felicity of symbol–referent links. A pilot experiment that motivated the use of geometric shapes in the present study suggested that infants do not accept that a shoe photo can stand for a dog if a dog photo is next to it. If infants are sensitive to the pragmatics of the stipulation or if they think that iconic and linguistic cues should match, infants might not accept such cross-mappings. In these cases, mislabeling will be interpreted—and rejected—as mis-stipulation.

DIRECTIONS FOR FUTURE WORK

There are also several questions that Experiments 1–4 leave open and that could be addressed in future research. The first concerns infants' behavior in Experiment 2, where infants looked at the highlighted object even when the question involved a new noun. I can think of two possibilities for why this may have been the case. First, infants may have rejected the stipulations because the shapes looked the same. In this case, the observed above-chance levels in both trial types of Experiment 2 would have been driven by the asymmetric highlighting only. This would mean that infants in this age range need distinctive visual features when assigning object symbols to discourse referents and that location is not prioritized as an individuation criterion. One way to test this possibility is to run a version of Experiment 2 in which the second, same-looking shape is not displayed on the screen during the stipulation phase.

Second, infants may have accepted the stipulations but generalized them because stipulations operate over symbol kinds instead of tokens: if one blue blob stands for a dog, another blue blob will stand for another dog, all else being equal. Kind-level stipulations would be in line with other proposals in the literature (Walton, 1990), are supported by experiments in which toddlers generalize stipulations at the kind level (Harris & Kavanaugh, 1993), and are often used

in legends (e.g., ◯ = lions). However, the possibility that the indexing system used for tracking symbols prioritizes visual features over location becomes again relevant. Instead of generalizing the mapping to symbol kinds, infants may have forgotten which of the two same-looking objects was the one entering the STAND-FOR relation because location faded away as the trial unfolded. One way to assess the different weights infants place on location and visual features would be to run a version of Experiment 1 in which the different-looking shapes are surreptitiously swapped before the test question.

The second open question involves the formal properties of STAND-FOR relations. As things stand, all that can be concluded is (i) that STAND-FOR relations are **functions**, as infants map the highlighted symbol to one discourse referent only (otherwise, infants would have looked at it on Different Word trials, too); and (ii) that STAND-FOR relations are **injective**, as infants map only the highlighted symbol to the discourse referent (otherwise, infants would have been at chance on Same Word trials, too). However, based on the data, I cannot discern whether STAND-FOR relations are **total** (i.e., whether each object in a visual scene must be a symbol) and whether they are **surjective** (i.e., whether each referent in the universe of discourse is assigned an object symbol). While infants were consistently at chance in Different Word trials, when the display consisted of two different objects, it is unclear how to interpret this result. Since infants behaved the same way in the Training phase across all experiments, this can only mean that pointing and labeling have an additional effect on infants' behavior at test. This effect is not a simple attentional mechanism—since it is speaker-dependent—but it introduced measurement noise and may have masked a mutual exclusivity inference. Therefore, whether infants expect STAND-FOR relations to be one-to-one functions remains unknown. Future work could address this shortcoming by having both objects pointed to before the test (while still labeling only one) and by leaving a longer break between the stipulation and the test question.

The third question concerns the interpretation of predicative expressions. The results indicate that 15-month-old infants can interpret predicative expressions as introducing STAND-FOR relations. Still, it is unclear whether this is the default interpretation or whether infants used contextual cues Experiments 1–4 to figure out that the IS-A interpretation could not have been correct (e.g., the blue blob is clearly not a dog, so it must stand for one). Future work could address this through a version of Experiment 3 in which the familiar nouns are replaced by unfamiliar ones (e.g., "Look! A blicket!"). If infants generalize the unfamiliar

nouns to a different speaker (e.g., Buresh & Woodward, 2007), this will establish that they can access both interpretations and select the one that fits best in the context.

Finally, the lack of an effect in the Training phase in Experiment 4 raises potential questions about the ease with which infants set up STAND-FOR relations based on the visual features of the symbols. While the lack of an effect with photos is probably best explained by the fact that these trials were presented first (as suggested by the effect on looking times in the second pair of the Training trials), the rich perceptual features and/or high amounts of iconicity may have hindered infants from drawing the appropriate inferences. This would be in line with findings suggesting that high similarity may actually defeat the purpose of symbols (DeLoache & Sharon, 2005). Infants may have been distracted by the fact that the photographs depicted real objects (and did not care about anything else) or because they interpreted the toys as toys—and toys do not move. Future work could address this by reversing the Training–Experimental block order, by exposing infants to pre-familiarization events that are similar to the test trials, or by using photographs of real animals instead of toys. In addition, the results of Experiment 4 suggest that infants may struggle with iconicity more than with language when setting up STAND-FOR relations. Future experiments could stipulate the identity of the discourse referents by other types of iconic features, such as movement or acoustic ones (e.g., barking). Language would then be only used at test to check whether infants mapped the shapes to the referents in a way consistent with the stipulations.

3.9. Conclusion

The experiments in this chapter support three claims. First, infants understand that visual objects can be used as symbols for other things. Second, infants know that the relations between the symbols and the referents they stand for are local to the speaker that stipulated them. This implies that they understand that the symbols are not the referents. Third, infants interpret the referents and the predicates applied to them based on conceptual knowledge. I take these findings as evidence for a cognitive mechanism dedicated to the representation and interpretation of STAND-FOR relations that can be activated without rich contextual cues and that emerges early and reliably in human ontogeny.

Chapter 4. Adults Interpret Images as Symbols: The Case of Automatic Size Measurement

4.1. Introduction

While many cognitive psychologists agree that pictures are representations of objects and scenes, they rarely consider the possibility that this fact contributes to adults' behavior in their experiments. Many times, this is not relevant, but sometimes it can be. While studying how infants understand screen-based depictions of events (**Chapter 2**), I realized that this aspect of the experimental situation is under-appreciated and under-researched even outside developmental research.

In this chapter, I focus on two properties of symbols. First, unlike ordinary objects, symbols require a referent that they can stand for. That is, after all, what a symbol is for: to carry information about another entity. This implies that one must interpret external symbols to determine what they currently stand for. Second, the interpretation assigned to symbols is generally achieved only with respect to the communicative context in which they are used. Consider a blue and red proportional bar graph without a legend and labels—one could not tell what the graph stands for only by looking at it. The interpretation of the graph is only possible in response to a perceived context, like an explanation, a legend, or a previously established expectation about its communicative content. This implies that the nature of the referent that a given symbol stands for cannot always be determined based on the perceptible properties of the symbol. While there is little doubt that ad-hoc, arbitrary symbols (e.g., blue and red bars) require more than just visual information for interpretation, it is often assumed that iconic symbols, such as images of objects, have a more direct link to the objects they stand for by virtue of their iconicity. However, even iconic symbols do not always elicit an appropriate mapping in a context-independent way. While an image of a green olive may be interpreted by default as standing for a green olive, the very same image can stand for olives in general (in a grocery store), for olive oil (on a bottle), or for an olive tree plantation (on a map).

Such symbolic communication beyond natural language is ubiquitous, effortless, and quick in human adults' everyday lives. For instance, movies, graphic novels, or video games are often fast-paced and still understood instantaneously, with rich meaning attributed to them beyond what is visually encoded (e.g., Hochberg, 1986). While these art pieces are created and consumed as communicative media, understanding their content rarely requires reasoning over particular interlocutors (e.g., movie directors or game designers).

In short, human adults often find themselves in communicative situations in which they have to (i) figure out what the symbols in front of them stand for; and (ii) use these assignments to interpret the messages that the symbols help convey. Nonetheless, many experiments in cognitive psychology present participants with pictures or animations on a screen, glossing over the possibility that participants might interpret these stimuli as symbols that are part of a communicative context (but see, e.g., Politzer, 2004; Snow & Culham, 2021). However, the very fact that participants are ostensively shown something (by the experimenters) may prompt them to interpret such stimuli as part of a communicative act (Sperber & Wilson, 1995). In addition, screens themselves are widely used as representational devices outside the psychology lab, which may be sufficient to trigger a communicative interpretation of the situation (Ittelson, 1996; but see Millikan, 2017). If symbolic interpretations were triggered upon encountering experimental stimuli, this would carry both methodological and theoretical implications. Methodologically, it might prompt researchers to control for unintended effects of communicative inferences that their stimuli might induce. Theoretically, it would provide evidence that external symbols gain a communicative interpretation rapidly and automatically.

Are experimental stimuli presented on a screen interpreted as communicative symbols? The case study I will focus on in this chapter is the familiar-size Stroop effect reported by Konkle & Oliva (2012). Konkle and Oliva (2012) had participants judge which of two images was displayed smaller or larger on the screen. They found that participants slowed down and made more errors on trials in which the size difference direction between the two images was opposite to the real-world size difference direction of the depicted objects. For instance, participants responded slower when presented with a large image of a palm leaf and a small image of an elephant compared to a display of a small image of a palm leaf and a large image of an elephant. The fact that elephants are larger in the world than palm leaves interfered with judgments of image sizes, even

though participants did not need to interpret the image contents for the task. This suggests that the process of encoding image contents is automatic. However, this conclusion leaves open several possibilities regarding the nature of this encoding process.

Here I consider two accounts that might underlie the familiar-size Stroop effect. The first possibility is that the interfering size measurement is an outcome of the perceptual processes that identify the category or the features of the objects depicted by the images. If this is the case, automatic size computation reflects the previous experience of encountering such features and objects and/or computations internal to the visual system that use featural information as cues for object size. There is work suggesting that object features, rather than object categories, may drive the effect on the automatic size measurements that give rise to the size Stroop effect. Long and Konkle (2017) ran the familiar-size Stroop task using distorted images of objects (called "texforms"), which preserved only mid-level featural information (e.g., curvature). Even though the basic kinds to which these objects belonged were no longer recognizable, a Stroop effect was still present, implying that the mid-level features carry sufficient information about the size of objects. While these findings show that accurate basic-level recognition is not necessary for the Stroop effect to occur, they do not entirely rule out a category-based explanation. It remains possible that participants inadvertently attempted to categorize the images at the basic level based on the mid-level features they were presented with. Even if these categorization attempts did not correctly identify the basic level kind of the underlying objects, the categories identified by the participants could still have been in the right ballpark in terms of size (e.g., they could have been more likely to guess "building" or "statue" when shown a texform of a vending machine than when shown a texform of a perfume bottle).

Thus, based on the available evidence, neither object features nor object category can be conclusively ruled out as the causal factor. For the current purposes, though, these options are equivalent because both assume that the Stroop effect is driven by the mismatch between what is perceived on the screen and the perceptually similar individuals in the world (Konkle & Oliva, 2012; Long & Konkle, 2017). I will thus group these explanations (categories and features) under a single general account, which I will refer to as the **object recognition** account.

The second option is that the familiar-size Stroop effect arises because of communicative inferences derived from the images. Under this account, one can construe each trial as a mini-discourse consisting of a question (e.g., "Which one is larger on the screen?") with two possible answers (the two images/response buttons). If participants interpret the images on the screen as symbols, they may inadvertently encode what entities these images might be conveying information about. Under this account, the incongruency comes from the mismatch between the images' relative sizes and the interpreted referents' relative sizes (e.g., a horse image conventionally refers to a horse).

This account implies that the real-world size of the on-screen object matters less than the real-world size of the referent that the on-screen object currently stands for (e.g., an image of a toy horse may activate the HORSE concept just as well as an image of an actual horse). Moreover, the interpretation assigned to the images should be a function not only of the image features but also of the context in which the image is embedded (e.g., it might be influenced by other symbols on the screen). Under this account, there is no one-to-one mapping between pictures of objects and corresponding representations of size. For example, an image of a toy horse will sometimes be taken to stand for a horse, sometimes for a toy, depending on what other image accompanies it. I refer to this as the **symbol interpretation** account.

I designed three experiments to evaluate the relative likelihoods of the hypotheses outlined above. Experiment 1 is a replication of the familiar-size Stroop effect (Konkle & Oliva, 2012, Experiment 1a) with new stimuli (**Figure 4.1**, left and middle column). Experiment 2 introduces miniature objects, such as toys, which are ideally suited to tease the two hypotheses apart because they are small in the real world but also typically used to stand for entities that are large in the real world (**Figure 4.1**, middle and right column). If the familiar-size Stroop effect is driven by object recognition, a small image of a toy horse next to a large image of a watermelon should not slow down image size judgments, as toy horses are much smaller than watermelons. However, if the toy horse (and, consequently, the image depicting it) is taken to stand for an actual horse, the opposite pattern should obtain. Participants' judgments will be slower when a small image of a toy horse is presented next to a large image of a watermelon, even though this configuration preserves the real-world size difference between the objects depicted on the screen.

Note that the **symbol interpretation** account does not predict that an image of a toy will necessarily be interpreted as standing for its non-toy counterpart. Instead, it predicts that some pictures—due to what they stand for in the context—give rise to a contrast between the real-world size of the depicted object and the perceived object size if participants interpret the depicted object to stand for a different referent. This prediction does not apply, a priori, to toys. Indeed, an image of a toy horse is ambiguous between a toy-horse interpretation and a horse interpretation. If participants opt for the toy interpretations when seeing images of toys, Experiment 2 will not be able to adjudicate between the **object recognition** and the **symbol interpretation** accounts. But if they opt for the toy referents interpretation, Experiment 2 would undermine the **object recognition** account, as this would show a dissociation between the object and the size measurement.

Finally, Experiment 3 asks whether the relationship between an object and its size measurement would be modulated by the identity of the second object. I investigate this question by comparing images of toy objects to images of the larger objects that the same toys typically represent (Figure 4.1, left and right columns). Should the very same image (e.g., a toy horse) be interpreted as a large object in Experiment 2 but as a small toy in Experiment 3 (due to the explicit within-category contrast), this would suggest that participants interpret the two pictures presented on a screen in an integrated rather than piecemeal fashion. This pattern would provide evidence in favor of the **symbol interpretation** account, as external symbols should be interpreted as constituent parts of the scene they are embedded in. On the other hand, the **object recognition** account—in its current formulation—takes perception to output size measurements of individual object images, for which the identity of the second object on the screen should be irrelevant. I return to this issue in the General Discussion, where I discuss several ways the **object recognition** account could be expanded to accommodate such contextual effects in light of the data I present.

Figure 4.1. Sample object images used across the Experiments 1–3. Experiment 1 consisted of comparisons between images in the first two columns (large object versus mid-sized object). Experiment 2 consisted of comparisons between the last two columns (mid-sized object versus small symbol of large object). Experiment 3 consisted of comparisons between the first and the last column (large object versus small symbol of large object).

4.2. Experiment 1: Replication

4.2.1. Methods

Experiment 1 was closely modeled on Experiment 1a of Konkle and Oliva (2012). Participants were presented with displays consisting of two different-sized images of real-world objects. Their task was to judge which of the two images was larger or smaller on the screen. In **Congruent** trials, the larger image depicted the object that is larger also in the real world; in **Incongruent** trials, the larger image depicted the smaller object (Figure 4.2, top row).

TRANSPARENCY AND OPENNESS

I report how I determined the sample size, data exclusions, manipulations, and measures of the study. The design of the study and the analyses were not preregistered, as all Experiments closely followed the design and analyses of Experiment 1a in Konkle and Oliva (2012). All stimuli, anonymized data, analysis code, and research materials are available on the Open Science Framework repository of the project, accessible at https://osf.io/q2yzc/. Data for all experiments were analyzed using R 4.2.2 (R Core Team, 2022), and the packages *ggplot* 3.4.2 (Wickham, 2016) and *effectsize* 0.6.0 (Ben-Shachar et al., 2020).

PARTICIPANTS

The sample consisted of 50 English-speaking participants ($\text{range}_{\text{age}}$ = 19–70 years, M_{age} = 29.9 years, SD_{age} = 12 years) recruited via the Testable Minds platform from all over the world. The sample size was chosen based on a pilot with 12 subjects to detect an effect of trial type with 99.9% power at significance level α = .05 (pilot Cohen's d = 0.73). All participants gave informed consent before completing the experiment.

STIMULI

For Experiments 1–3, I gathered 36 triplets of objects, each of which contained a large object X, a small toy object X, and a different object Y, whose size lay between that of the real object X and that of the toy object X (Figure 4.1). In addition, for each of the three triplet pairs (X–Y in Experiment 1, toy X–Y in Experiment 2, X–toy X in Experiment 3), the absolute difference between the aspect ratios of the bounding boxes was at most 0.25, based on (2012, Experiment 1a), but with a slightly wider margin because the constraint had to hold across triplets rather than pairs. The triplet items were not matched in terms of filled

areas (proportions of filled space in the corresponding bounding boxes), but I controlled for these differences statistically when analyzing the data (see Appendix B). The 108 object images were then resized to create two different-sized versions for each image. The large versions were resized such that the diagonal of the object's bounding box was approximately 1,000 pixels; the small versions were created by scaling the large images down by a factor of 0.6. In Congruent pairs, the size difference between the images was in the same direction as the real-world size difference of the depicted objects; in Incongruent pairs, the size difference between the images was in the opposite direction to the real-world size difference of the depicted objects (Figure 4.2).

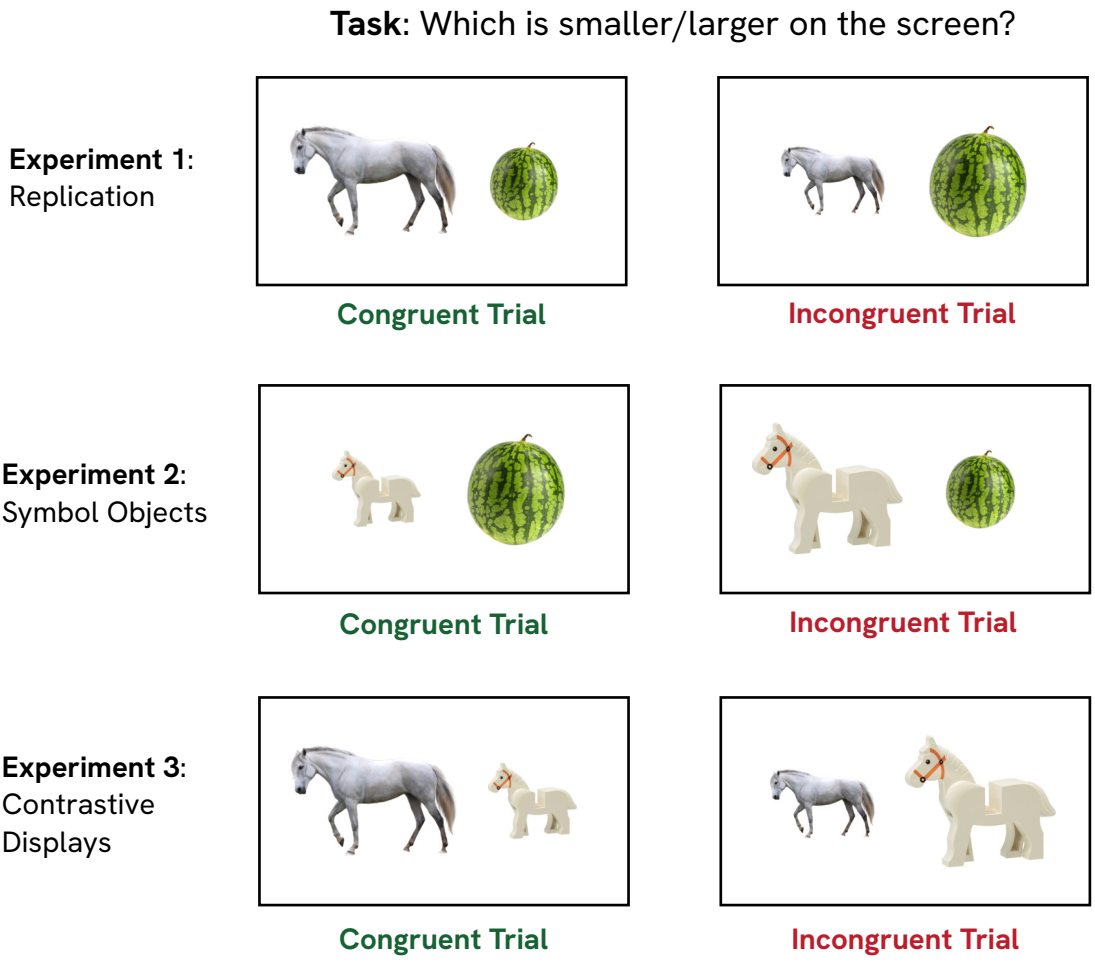

Figure 4.2. Schematic design of Experiments 1–3. Left: the size difference between the images is in the same direction as the real-world size difference of the objects (Congruent trials). Right: the size difference between images is in the opposite direction to the real-world size difference (Incongruent trials). In Experiment 2, congruency is defined based on the actual size of the objects depicted in the images.

PROCEDURE

Participants were told that they would see two images on each trial and were instructed to press the **F**-key on their keyboard to select the image on the left or the **J**-key for the one on the right. They were also instructed that in one block of trials, they would have to judge which of the two images was smaller on the screen, while in another block of trials, they would have to judge which of the two images was larger on the screen. Participants underwent two short practice blocks (eight trials in total), in which they had to select which of two colored circles was smaller/larger on the screen, to get familiarized with the task and the two response keys. The test phase began once participants answered correctly to four consecutive practice trials in both blocks.

Each of the two test blocks (**Larger** vs. **Smaller**, order randomized across participants) consisted of 144 trials (36 Pairs × 2 Trial Types [**Congruent** vs. **Incongruent**] × 2 Sides [**Left** vs. **Right** of the screen]). Thus, the entire experiment consisted of 288 trials. Each trial started with a fixation cross for 700 ms, followed by the image comparison display. Correct responses were immediately followed by the next trial. In contrast, incorrect responses received error feedback ("Oops, this is incorrect! Remember, choose the one which is **smaller**/**larger** on the screen.") and by a 5-second interval before the next trial began. The order of trials was randomized for each participant. Once participants finished the first block, they were congratulated and told which task they would have to solve in the remaining block (Smaller or Larger, depending on the first block).

4.2.2. Results

REACTION TIMES

As in Konkle and Oliva (2012, Experiment 1), I excluded trials with incorrect answers and trials in which reaction times were shorter than 200 ms or longer than 1500 ms from the analysis. After these exclusions, each participant produced 273.3 valid trials, on average, out of a possible 288. To investigate reaction times, I obtained a Stroop effect for each participant by subtracting the average reaction times on Congruent trials from those on Incongruent trials. A positive Stroop score would mean that participants take longer to answer on Incongruent trials; a negative Stroop score would imply that participants take longer to respond on Congruent trials. To consult the results by task (Larger vs. Smaller), see Appendix B.

The overall effect of real-world size congruency replicated the original finding (Figure 4.3A). It took participants longer to make a visual size judgment on two images when the image sizes were incongruent with the real-life sizes of the objects depicted in the images ($M_{Congruent}$ = 625.8 ms, $M_{Incongruent}$ = 650.0 ms; $t(49) = 5.24$, $p < .001$, Cohen's $d = 0.74$, 95% CI [0.43, 1.06]).

ERROR RATES

Following Konkle and Oliva (2012), I also compared error rates across trial types within each condition (Figure 4.3B). While participants were, on average, 98% accurate, they were more likely to err in Incongruent trials than in Congruent trials, $t(49) = 5.31$, $p < .001$, Cohen's $d = 0.75$, 95% CI [0.44, 1.07].

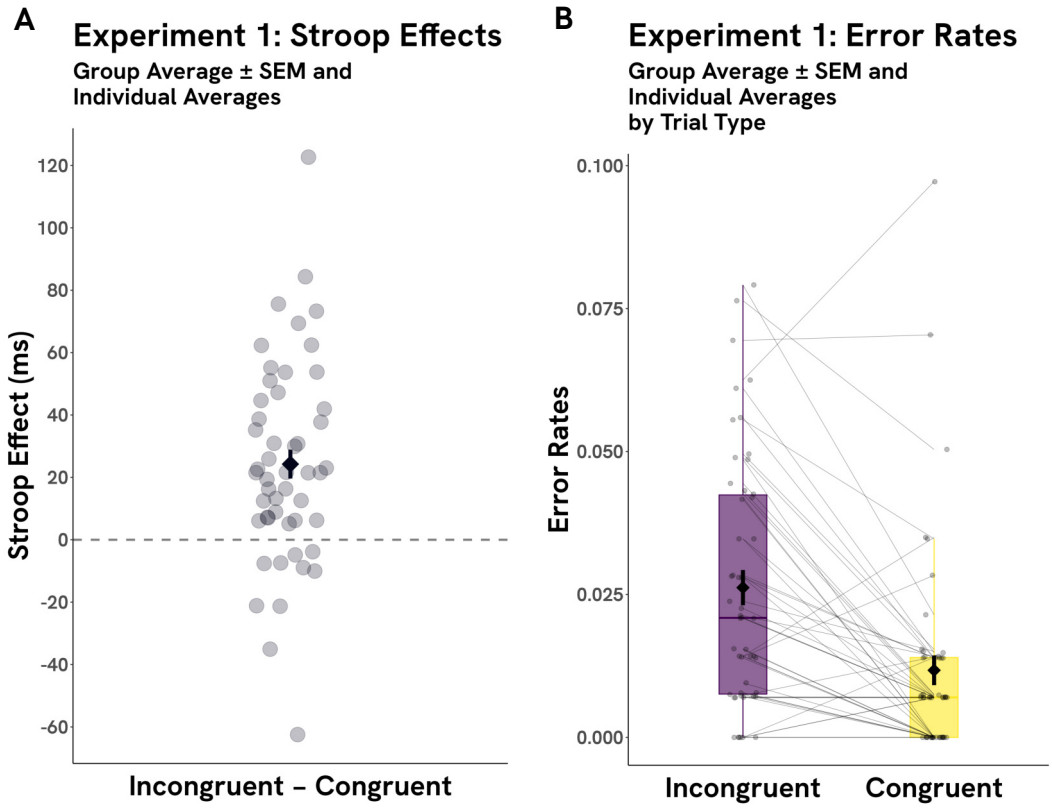

Figure 4.3. Results of Experiment 1. **(A)** Stroop effects. Transparent circles represent within-subject Stroop effects (Incongruent – Congruent reaction times); black diamond shows average Stroop effect ±1 SEM. **(B)** Error rates. Transparent circles and the lines connecting them represent individual error rates as a function of trial type; opaque diamonds depict group averages ±1 SEM; boxplots indicate the median and interquartile range.

4.2.3. Discussion

Experiment 1 replicated the size Stroop effect reported by Konkle and Oliva (2012): (i) participants were slower to make a visual size judgment in Incongruent trials, i.e., when the size relation of two on-screen images did not align with the size relation of the depicted objects; and (ii) participants were less accurate in Incongruent than in Congruent trials. As noted in the Introduction, however, the results are ambiguous as to the nature of the process that gives rise to this effect. Is it simply the case that the category and/or the perceptual features of the objects in the images are associated with a previous encoding of such features and objects? Or do participants compute what these images stand for? I addressed this question in Experiment 2.

4.3. Experiment 2: Symbol Objects

To determine whether the familiar-size Stroop effect is driven by **object recognition** or **symbol interpretation**, I replaced the large objects in Experiment 1 with miniature versions of the same objects (**Figure 4.1**, right column). When participants compare images of watermelons to images of toy horses, which of the two size differences will they take longer to judge? Since toy objects are small, the visual **object recognition** account predicts that participants will take longer if the toy/miniature objects are depicted as larger than the mid-sized objects from Experiment 1. By contrast, only the **symbol interpretation** hypothesis can account for the possibility that participants will find those trials easier in which the toy/miniature objects are depicted as larger than the same mid-sized objects.

4.3.1. Methods

The methods were identical to Experiment 1 except that the large objects were replaced by small symbol objects that represent them (e.g., horse → toy horse). Thus, in Experiment 2, participants had to compare displays of mid-sized objects versus small objects that typically stand for large objects (**Figure 4.2**, middle row). Importantly, I define congruency based on the actual size of the objects depicted. For instance, a large image of a toy horse next to a small image of a watermelon is an Incongruent trial, as toy horses are typically smaller than watermelons. This choice is, of course, arbitrary. After all, this experiment aimed to

determine which of the two trial types would be incongruent for participants. Still, it is important to keep in mind for interpreting the results.

PARTICIPANTS

The sample consisted of 50 participants (range$_{age}$ = 18–67 years, M_{age} = 31.9 years, SD_{age} = 11.7 years) recruited via the Testable Minds platform. The sample size was chosen based on a new pilot with 12 subjects to detect an effect of trial type with 99.9% power at significance level α = .05 (pilot Cohen's d = 0.74). All participants gave informed consent before completing the experiment.

4.3.2. Results

REACTION TIMES

As in Experiment 1, trials with incorrect answers and trials with reaction times shorter than 200 ms or longer than 1500 ms were removed from the analysis. Based on these criteria, participants provided, on average, 280.2 valid trials. I obtained a Stroop effect for each participant by subtracting the average reaction times on Congruent trials from those on Incongruent trials. Because congruency is defined based on the actual size of the objects depicted by the images, a positive Stroop score would indicate that slower responses were produced when miniature object images were large (e.g., large toy horse versus small watermelon). In contrast, a negative score would indicate slower responses when miniature object images were small (e.g., small toy horse versus large watermelon).

Unlike in Experiment 1, there was an overall negative Stroop effect (**Figure 4.4A**): participants' reaction times were higher on Congruent trials, in which the visual size of the images on the screen matched the sizes of the objects that were depicted on-screen ($M_{Congruent}$ = 568.5 ms, $M_{Incongruent}$ = 561.1 ms; $t(49) = -2.46$, $p = .017$, Cohen's d = 0.35, 95% CI [0.06, 0.64]).

ERROR RATES

Consistent with the reaction times results, participants were more likely to make a mistake on Congruent than on Incongruent trials (**Figure 4.4B**), $t(49) = -4.28$, $p < .001$, Cohen's d = 0.61, 95% CI [0.30, 0.91].

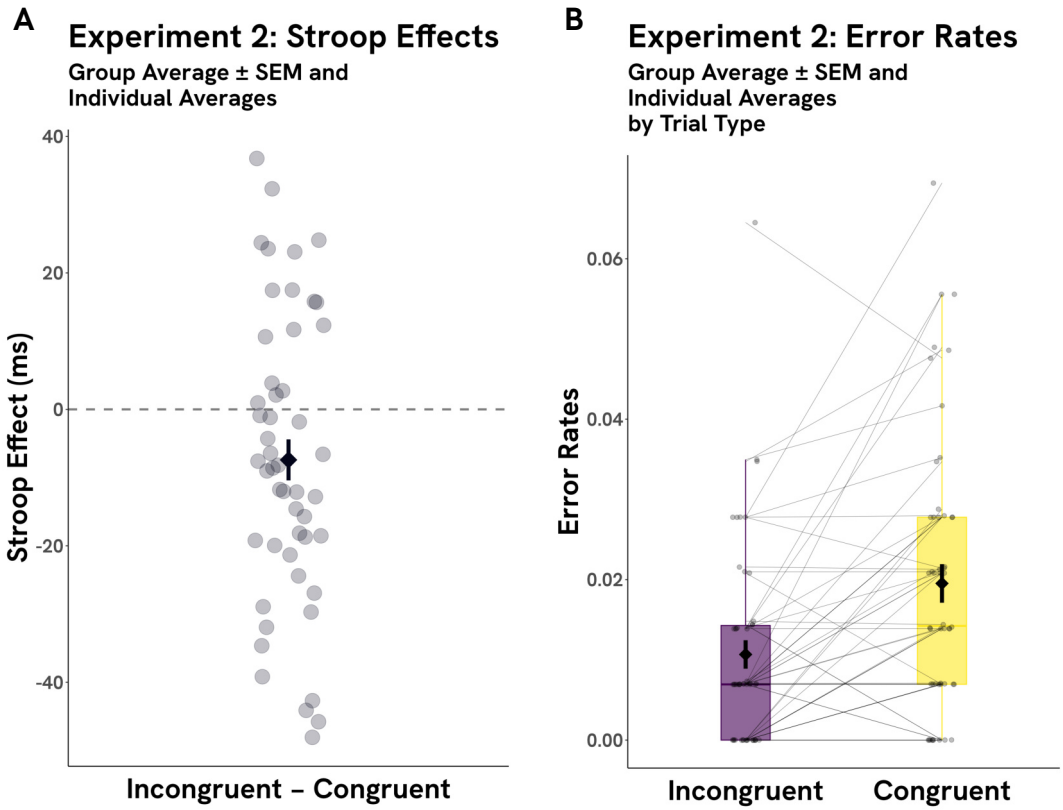

Figure 4.4. Results of Experiment 2. **(A)** Stroop effects. **(B)** Error rates.

BIMODALITY

The smaller Stroop effect in Experiment 2, combined with the observation that one-third of the participants seem to have exhibited the opposite effect, could be driven by an underlying bimodal distribution. The bimodality would arise because some participants would take the size measurements of toys, whereas others would take the size measurements of the large objects represented by the toys. This was not the case. First, this smaller effect size compared to Experiment 1 is only apparent in the reaction times but not in the error rates. If error rates and reaction times resulted from the same incongruence, the effect size differences should go hand in hand, but they do not. Second, the statistical analysis for multimodality on reaction time difference scores does not reject the null hypothesis (Hartigan's dip test for bimodality, D = .04, p = .944).

4.3.3. Experiments 1 and 2: Contrast

REACTION TIMES ACROSS EXPERIMENTS 1 AND 2

To compare the results of the first two experiments, I aggregated the two datasets and analyzed reaction times by trial type (Congruent vs. Incongruent, within-subjects) and experiment (1 vs. 2, between-subjects). A 2 × 2 mixed ANOVA revealed a main effect of experiment, $F(1, 98) = 8.87$, $p = .004$, $\eta_p^2 = .083$; a main effect of trial type, $F(1, 98) = 9.33$, $p = .003$, $\eta_p^2 = .09$; and a Trial Type × Experiment interaction, $F(1, 98) = 32.93$, $p < .001$, $\eta_p^2 = .25$.

ITEM-BASED COMPARISON ACROSS EXPERIMENTS 1 AND 2

As an additional exploratory measure, I grouped the data by image pair (e.g., train–fountain, bear–drumset) and calculated the item-wise correlation of Stroop effects across the two conditions (Figure 4.5). There was a strong negative correlation, $r(34) = -.65$, $p < .001$, indicating that Stroop effects tended to be driven by the same pairs across conditions. If, for instance, there was a processing advantage for Congruent truck–hairdryer trials in Experiment 1, there was a similar but opposite effect in Experiment 2 despite the many differences between trucks and toy trucks. (The correlation is negative because congruency is defined at the level of the depicted objects.)

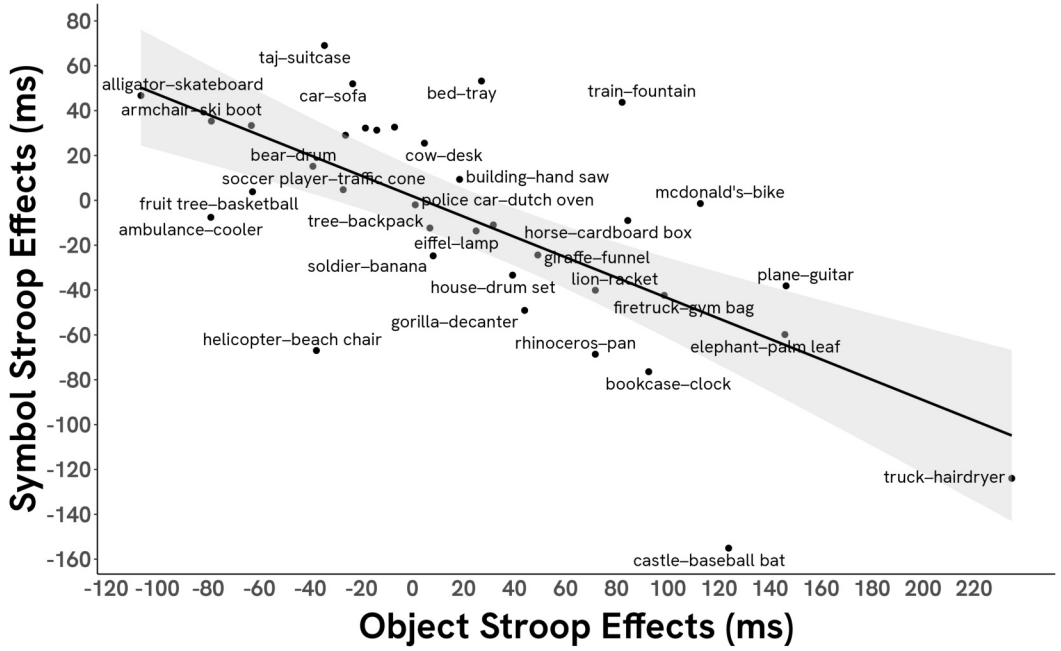

Figure 4.5. Correlation between Stroop effects in Experiments 1 and 2 across items (*x*-axis: Experiment 1; *y*-axis: Experiment 2). The gray-shaded area represents the 95% confidence interval around the regression line.

4.3.4. Discussion

Experiment 2 produced a different pattern of results from that of Experiment 1. Participants were slower and more error-prone on Congruent trials even though the difference in size between the two on-screen images went in the same direction as the real-world difference. Moreover, the Stroop effects by stimuli pair in the two conditions were strongly correlated, indicating that size judgments were similarly slowed down when the large object (e.g., a zebra) was depicted in a small image, irrespective of whether it was directly represented by a member of the kind (an image of a real zebra) or indirectly by an object that is often used to refer to it (an image of a toy zebra). Taken together, the results of Experiments 1 and 2 suggest that the Stroop size effect is not driven primarily by object recognition (e.g., toy zebra → small or toy zebra features → small) but by the inferred referent of the image (toy zebra → zebra → large).

Compared to Experiment 1, more participants in Experiment 2 exhibited a positive Stroop effect in reaction times (but not error rates). The participants who produced a positive Stroop effect may have interpreted the toy objects as toys (rather than as the objects they were toys of). If so, this would provide further evidence for the **symbol interpretation** account. Under this account, communicative inferences on the visual input are responsible for the size measurements underlying the Stroop effect. Due to the ambiguity of toy images, some participants could have had different assumptions about what the images communicate. On the **object recognition** account, these results would require auxiliary hypotheses. How could the same visual input drive different effects across trials and across participants? Are toy objects visually bistable between toy and non-toy objects? Or do people have different visual systems to such an extreme degree? Neither of these explanations seems promising. In short, whether the true distribution of the Stroop effect in Experiment 2 was bimodal remains unclear. But if it were, it would support the **symbol interpretation** account better than its alternative.

Two concerns remain, however. First, the features attended to in the visual processing of the images may be orthogonal to the toy–real distinction. If this were the case, images of toys and images of real things would end up in identical outputs (e.g., both a zebra image and a toy zebra image output ZEBRA). That is, participants in Experiment 2 might have, in some sense, mistaken the toys for the objects the toys stood for. Second, it is possible to modify the **object recognition** account to accommodate the results even if participants did not mistake

the toys for real objects. If the size Stroop effect is driven by object categories, one can postulate that participants always retrieve the conceptual content conventionally associated with the object in the image (e.g., a toy zebra always activates ZEBRA). If it is driven by object features, one can argue that toy objects share the relevant mid-level visual features with real-object counterparts. Both these modifications preserve the core idea of the **object recognition** account, namely that the primary input to the size measurement of an image is the set of its perceptible features. The **symbol interpretation** account, by contrast, assumes that participants assign an interpretation to the symbols presented to them in relation to a discourse context. Under this account, the context (e.g., the other image on the screen) can shift the interpretation of the images. I tested this prediction in Experiment 3 while also controlling for the possibility that participants in Experiment 2 mistook the toys for the objects the toys represented.

4.4. Experiment 3: Contrastive Displays

By pairing the 36 large-object images from Experiment 1 with their corresponding miniature versions from Experiment 2, Experiment 3 tested whether participants are sensitive to the context in which an image is embedded when judging its relative size. If participants inflexibly assign a ZEBRA interpretation to both a toy zebra image and a zebra image because that is the commonly associated conceptual content of both images, the Stroop effect should disappear. If, on the other hand, participants are sensitive to the communicative context, they should consider both images when assigning an interpretation. In Experiment 3, participants were faced with a direct contrast between large objects and their miniature versions. Because of this, under the **symbol interpretation** account, they should opt for a different interpretation of the toy object images compared to Experiment 2. Toys should now stand for the corresponding concepts of toys and not for the concepts that the toys usually stand for because the paired images already stand for those concepts. This contrastive reading predicts a processing advantage for Congruent trials and thus a size Stroop effect.

4.4.1. Methods

Experiment 3 was identical to Experiments 1 and 2, except for the stimuli pairs, which now consisted of pairs of objects and their corresponding miniature versions selected from Experiments 1 and 2, respectively (**Figure 4.2**, bottom row).

PARTICIPANTS

The sample consisted of 50 participants (range$_{age}$ = 20–60 years, M_{age} = 31.9 years, SD_{age} = 10.4 years) recruited via the Testable Minds platform. The sample size was chosen based on a pilot with 12 subjects to detect an effect of trial type with 99.9% power at significance level α = .05 (pilot Cohen's d = 0.86). All participants gave informed consent before completing the experiment.

4.4.2. Results

REACTION TIMES

As in Experiments 1 and 2, trials with incorrect answers and trials that lasted shorter than 200 ms or longer than 1500 ms were excluded. Each participant provided, on average, 276.9 valid trials out of a maximum of 288. There was an advantage for Congruent trials (Figure 4.6A). Reaction times on Incongruent trials were longer than those on Congruent trials ($M_{Congruent}$ = 592.3 ms, $M_{Incongruent}$ = 616.8 ms; $t(50)$ = 7.58, p < .001, Cohen's d = 1, 95% CI [0.73, 1.43]). The Stroop effect replicates the original Konkle and Oliva (2012) finding, as well as its replication in Experiment 1 above.

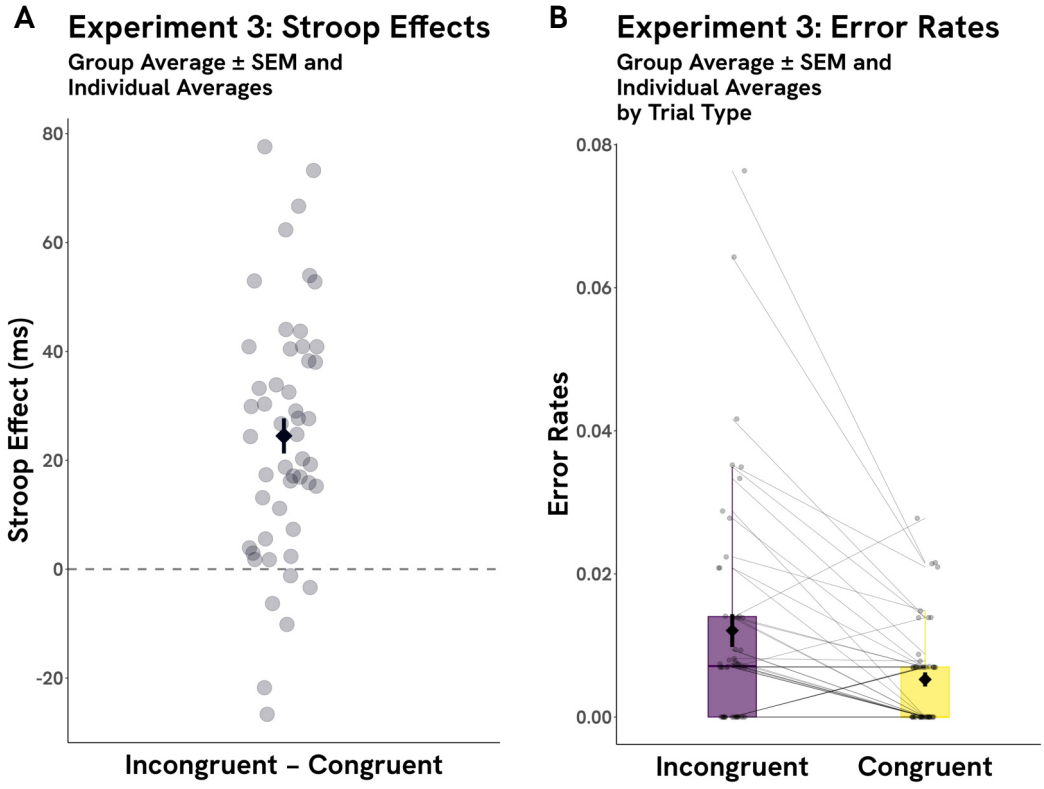

Figure 4.6. Results of Experiment 3. **(A)** Stroop effects. **(B)** Error rates.

ERROR RATES

As in Experiments 1 and 2, participants were highly accurate in solving the task (on average, 99% correct responses), but they were more likely to make a mistake on Incongruent trials than on Congruent trials, $t(50) = 3.75$, $p < .001$, Cohen's $d = 0.53$, 95% CI [0.23, 0.83], as in Experiment 1 (**Figure 4.6B**).

4.4.3. Discussion

Experiment 3 rules out two potential explanations for the findings of Experiment 2. First, the results of Experiment 2 were not due to participants' mistaking the toys for the objects they typically represent. Had they done so, they would have assigned the same interpretation to both images, which would have led to a null result in Experiment 3. On the contrary, participants took longer and made more mistakes on Incongruent trials, i.e., when an object was displayed as smaller than its miniature counterpart. Second, and more importantly, the results show that people's interpretation of images is not inflexible but changes according to context. By placing the toys next to the objects they usually represent, a different interpretation was elicited in Experiment 3 (e.g., a toy zebra stands for a toy) compared to Experiment 2 (e.g., a toy zebra stands for a zebra). These results rule out the modified **object recognition** accounts. Participants could not have assigned a conceptual description to an image dictated solely by the conventional use of the image, nor could they have derived a size measurement from the visual features in the image: neither of these accounts are compatible with a shift in interpretation due to a change in context. Taken together, Experiments 2 and 3 suggest that people interpret images presented to them as symbols. While the visual properties of a symbol constrain the domain of entities and concepts it can represent, participants flexibly use the context to converge on an interpretation when multiple reasonable candidates are present (as in the case of toys).

4.5. General Discussion

The present study explored how participants represent visual stimuli in experimental contexts. Do participants encode the object images to reflect the real-life entities that are depicted or interpret these images as symbols? And if the latter, how do they decide what a picture stands for? The case study for investigating these two questions was the familiar-size Stroop effect (Konkle & Oliva, 2012), which led to progress on both issues with simple manipulations.

In Experiments 1 and 2, I asked whether visual stimuli are recognized as the real-world objects they depict or if they are represented as symbols—objects that stand for something else. Having successfully replicated the original size Stroop effect in Experiment 1, I changed the stimuli in Experiment 2 such that the larger objects in Experiment 1 were swapped with toys of the same category. If automatic size measurements are based on the size of real-world entities, toy objects should be construed as small objects relative to the objects they were paired with. But if a picture of a toy object carries the same symbolic content as an image of its non-toy counterpart, participants should still treat it as larger than an actually bigger object. The size Stroop effect followed the symbolic content rather than the real-life size of the depicted objects: participants represented pictures of toy zebras on a par with real zebras when it came to automatic calculations of object size. This fits neatly only with the proposal that participants conceive of object presentations as symbolic.

Having found evidence for symbolic encoding in Experiments 1 and 2, I turned to the second question: what processes are responsible for creating the symbolic connection between picture and content? On one view, symbols, just like real objects, are recognized based on their perceptible properties. In line with this view, iconic conventions often mediate communication using external symbols. Pictures of zebras conventionally symbolize zebras and not, for instance, horses or houses. In fact, one would arguably call any picture a zebra picture if it could be recognized by others as representing a zebra, even if that picture is cartoonish or barely resembles entities belonging to the subgenus *Hippotigris*. If you know how zebras are conventionally depicted, you might automatically encode any zebra depiction, toy or otherwise, as standing for a zebra. On the alternative view, the connection between symbol objects and their content is not just a recognition process but an interpretive one, whereby participants decide what an object stands for in some context. A stick figure at a crosswalk signals that it is ok to walk, but on a door at the airport, it signals the location of the restroom. No recognition process could reliably output the appropriate content because the context—in this case, the physical environment—needs interpretation. This view suggests that there is no context-independent way of identifying whether an image of a toy zebra stands for a toy zebra, a real zebra, or something else. The results in Experiment 3 were consistent with the **interpretation** account: participants encoded a toy object as **smaller** than its non-toy counterpart. This is the opposite of Experiment 2, where the contrast

between object categories drove the effect rather than ontological status (toy versus real).

Just as in Konkle and Oliva (2012), the processes that automatically generated the irrelevant size measurements of the depicted objects are rapid, spontaneous, and lacking control—properties that have been argued to be necessary, if not sufficient, signatures of visual processing (Hafri & Firestone, 2021; Scholl & Gao, 2013). Nevertheless, I have reasons to suspect that visual processes cannot explain the size Stroop effect and, especially, the contextual effect in Experiments 2 and 3. Why does the automatic size measurement of toy objects depend on the other object on the screen?

One could defend the visual origin of this effect by arguing that visual features of images of toy objects are mapped both to toy and non-toy versions of the same object category, such that upon seeing an image of a toy zebra, the visual system would either output a toy zebra (and its size) or a non-toy zebra (and its size). However, this mapping alone would not predict when participants should encode a toy object one way or another. As such, it fails to explain the systematic context dependence the findings exhibit. If one of the two content types is chosen randomly, there should have been no Stroop effect in Experiment 2. In half of the trials, the toy objects would have been perceived as the smaller object, and in the other half, they would have been perceived as the bigger object, resulting in no size difference on average. In addition, there size effect in Experiment 3 should have been smaller: the toy objects would have been perceived as small only in half of the trials. However, the size Stroop effect was even larger in Experiment 3 than in Experiment 1, ruling this option out.

A second visual account of the present results could be that toy objects—counterintuitively—look larger than the mid-sized objects they were presented next to in Experiment 2. By this account, the toy objects might have created a visual illusion: they shared enough features with their large non-toy variants that the size measurement they activated was closer to these objects than to their de facto (small) sizes. Under the additional assumption that this illusion was only partial, one might also be able to explain how Experiment 3 worked: perhaps the toys looked larger than the mid-sized objects (in Experiment 2) but smaller than the large objects (in Experiment 3). I dispense with this account because it is highly stipulative. Intuitively, toy objects do not look large; they look like toys. In addition, there is suggestive empirical evidence against it. If toy objects looked

similar to the large objects in terms of size, Experiment 3 should have produced the smallest effect size of all. It did not.

A further way to incorporate contextual dependence into visual processes is to assume that toy objects **look** more toylike when the non-toy objects are presented next to them, as in Experiment 3. This could be similar to well-known perceptual contrast effects, such as the modulation of color perception by the brightness of the background (simultaneous contrast effect: e.g., Kinney, 1965). However, such a contrast effect should be more than just the relation of toy and non-toy features neighboring each other. That would also predict a straightforward contrast effect between toys and non-toys in Experiment 2, where there was none. To generate the appropriate contrast effect, vision should apply a rule along the lines of "for any toy object X, if and only if there is another object Y that represents the same category as X but is not a toy, encode X as a toy". A rule of this sort would radically differ from the visual contrast effects discussed in the literature for at least two reasons. First, to create such contrast effects, the relevant constraints would have to encode symbolic properties such as TOY VERSION and REPRESENTING THE SAME CATEGORY AS—properties that fall outside paradigmatic perceptual processes. Second, this potential constraint would be a post-hoc stipulation that does not follow from any general account of perceptual processing. As such, it has little to no predictive power.

Yet another way to root the contextual contrast effect in visual processes is to suggest that it reflects a more general process of optimal visual inference under uncertainty (e.g., Weiss et al., 2002). This option would concede that the process that creates object descriptions is interpretive in some sense but would still assume that this interpretation is created within the confines of the visual system. Depending on the visual context, it might be more optimal to encode an object as a toy zebra rather than zebra. But why would it be optimal to encode pictures of toys as toys in Experiment 3 but as non-toys in Experiment 2? To make this work, one would have to posit that the likelihood that two neighboring objects that, in principle, belong to the same category are actually from the same category is low. Could the presence of a zebra decrease the likelihood of encountering another zebra and increase the relative likelihood of encountering a toy one? If anything, the opposite seems more plausible: if a zebra is around, the probability of encountering another (non-toy) zebra increases.

A different type of rational process would be needed to generate the correct predictions—one that asks, "Why am I presented with these pictures?" in-

stead of “What is most likely to be out there?”. If the stimulus is understood as part of a communicative act, there is good reason to interpret the contrast between two items as a matter of identifying the communicative message and not as a matter of identifying object categories. One might assume that being presented with side-by-side images of two objects that conventionally stand for the same category is not the outcome of a random sampling of tokens that happen to belong to the same object category but a deliberate contrast. And if the difference between the images is made on purpose, then this distinction should play a role in how one interprets what they stand for. This is a rational inference, just not one for which the visual system could be straightforwardly responsible.

I do not mean that vision plays no role in the above inferential processes and in the size Stroop effect in general. Every account of the effect must involve vision, as there could be no size Stroop effect without a visual representation of the stimuli. But the interfering size measurements may originate not from perceptual processes directly but from a communicative interpretation of their outputs instead. An excellent analogy to the role of vision in this study is the role of the auditory systems in understanding the meaning of a spoken sentence. In both cases, perceptual processes must create an encoding of the input that is amenable for an interpretation by other processes, but they themselves do not provide the interpretation.

If the size Stroop effect stems from symbol interpretation and reflects fast and automatic processing, it follows that, like visual processes, communicative interpretation of visual stimuli can also be fast and automatic. If so, the signature features of perceptual processing can no longer be taken for granted, and it will become necessary to consider how to best tease apart the mechanisms that reason over communication via symbols from other processes. Using pictures or other visual stimuli in experimentation is ubiquitous, from vision science to social psychology. The present findings have methodological implications for such studies. Participants engage in a communicative interpretation of the stimuli even when it comes to rapid automatic decisions. Therefore, in any experiment where participants encounter visual stimuli, their behavior might reflect their interpretation of the stimuli as external communicative symbols rather than mere recognition of the entities depicted on the screen.

Consider a simple animation involving two geometric shapes moving in a contingent way on the screen (Heider & Simmel, 1944). People interpret such an animation in agentive terms and parse the on-screen interaction as a chasing

event. This interpretation has been argued to be due to the self-propelled motion exhibited by the geometric shapes (e.g., Scholl & Tremoulet, 2000). But if viewers treat the animation as a representation and its constitutive parts as symbols to be interpreted, finding that, say, self-propelled motion is a cue to agency is ambiguous between purely a perceptual interpretation (people perceive agency when confronted with self-propelled motion) and a communicative interpretation (self-propelled motion is a good way of conveying agency). While both interpretations imply a strong link between self-propulsion and agency ascription, teasing apart the relative contribution of these candidate processes requires careful experimental controls.

4.6. Conclusion

The experiments reported in this chapter provide evidence that participants encode pictures of objects as having symbolic and context-dependent content, indicating that the familiar-size Stroop effect is driven by communicative inferences rather than just visual recognition. I have argued that, when presented with images on a screen, humans do not simply encode their features or category but automatically try to figure out what the visual objects in front of them currently stand for. Moreover, this interpretive process, which depends on perception but does not originate in perception, exhibits signature properties of vision: it happens quickly, automatically, and without direct relevance to the task at hand.

Chapter 5. Coda

5.1. Overall Summary

Humans are arguably the only species that uses symbols not only internally, to represent the world (e.g., Dehaene et al., 2022) but also externally, to communicate information to conspecifics (e.g., Csibra & Shamsudheen, 2015). Inside the mind, symbols are primitives that denote objects (in a broad sense) and that combine with other symbols to form sentences in a language of thought (Fodor, 1975; Piantadosi et al., 2016). Outside the mind, on top of the units of natural languages (words and signs), humans often use visually available objects as symbols for conversationally relevant referents in a variety of communicative devices (Clark, 2016): a circle for a ball in an on-screen animation; a doll for an agent in a puppet show; a banana for a phone in children's pretend play.

In Chapter 1, I put forth a cognitive architecture that creates local assignments between symbols and discourse referents by taking a visual object as input (e.g., a ◯) and assigning it to a mentally indexed referent (e.g., a ball). While the assignment is in place, the perceptually available symbol and the mentally represented referent are in a STAND-FOR relation, which allows the actions performed on or by the symbol object to be interpreted as information about the referent. For instance, if an on-screen circle is established as a symbol of a ball, the horizontal movement of the circle can be interpreted as the ball rolling.

In Chapter 2, I investigated whether 19-month-olds understand animated events as representational by testing whether they think on-screen events are happening in the here and now or decouple them from the immediate spatial environment. I found that infants do not expect on-screen animated falling objects to end up outside the confines of the screen, indicating that they did not take animated events to be continuous with the surrounding environment. At the same time, infants accepted that the same animated environment could be depicted on multiple screens, suggesting that they have already figured out the medium–content independence of animation—a signature feature of symbolic representation.

In Chapter 3, I asked whether 15-month-olds can set up representational relations between arbitrary objects and familiar discourse referents. I found evidence that infants create assignments between geometric shapes (e.g., a blue blob) and familiar category tokens (e.g., a dog) when the shapes are labeled as such (e.g., "Look! A dog!" while pointing to the blob). Infants restricted the assignments to the local communicative episode, as they did not generalize them to a new speaker. At the same time, the identity of the discourse referent mattered. Infants distinguished between symbols for animate versus inanimate referents and preferred it when the predicates ascribed to the referents were consistent with their identity. Thus, STAND-FOR relations are available early in human ontogeny. In line with the theoretical account in Chapter 1, the assignments are local and recruit infants' conceptual knowledge about the discourse referents when the assignments are created.

In Chapter 4, I presented an analogous result in human adults, who are also often shown symbolic stimuli in tasks intended to measure unrelated cognitive processes (e.g., Heider & Simmel, 1944). For instance, previous work using photographs of objects showed that adults automatically estimate the sizes of objects (Konkle & Oliva, 2012). What remained unclear, however, was whether adults engaging in automatic size computation recruit object recognition processes (e.g., a photo of a horse is recognized as a real horse) or symbol interpretation processes (e.g., a photo of a horse is interpreted as a symbol of a horse). To investigate this, I ran a Stroop task with photos of toy objects (i.e., small objects that often represent large objects) and tested the direction of the automatic size computations. Against the object recognition account, adults' reaction times indicated that they interpreted small toy objects as standing for the large entities they were toys of. This finding did not arise because people mistook the toys for the real entities the toys represented. When the toys were paired with the entities they typically represent (e.g., a horse vs. a toy horse), adults were sensitive to the contrast and interpreted the toy photos as small objects. Adults thus interpret even realistic photos of objects as symbols, and they do so in an automatic, task-independent, but context-sensitive manner.

The common thread surfacing from all these experiments is that symbols may well constitute a class of stimuli to which human cognition is uniquely attuned. In the next section, I turn to the methodological implications of these findings in the context of the wide use of symbolic stimuli in experimental research.

5.2. Methodological Implications

A few years ago, in the summer of 2020, a heated exchange took place on the Cognitive Development Society's mail forum (cogdevsoc@lists.cogdevsoc.org). At the heart of the exchange was the observation that developmental research runs experiments with animations and puppet shows as stimuli but draws conclusions as if such stimuli were interchangeable with the real-world stimuli that researchers would primarily be interested in (Packer & Moreno-Dulcey, 2022). The exchange turned into a special issue in *Cognitive Development* (Paulus et al., 2022), meant to assess this observation from multiple perspectives and find solutions for how to deal with it.

In the target article of the special issue, Packer and Moreno-Dulcey (2022) focused on social cognition research. They argued that the external validity of studies using puppets to tap into participants' conception of agents is under threat. They pointed out that the stimuli used in the lab differ from the situations participants encounter in the wild: shapes on the screen, wooden circles, and puppets. Based on this observation, they concluded that using such stimuli introduces an experimental confound because they ask children to enter a pretend game with the experimenter. And since researchers measure only children's behavior toward mentally inert entities, there is nothing to justify the inference that infants and toddlers attribute goals, beliefs, and moral dispositions to actual agents. While I agree that using symbolic stimuli introduces an additional representational layer, that this should not be taken for granted, and that this raises interesting theoretical questions, I doubt that this methodological choice threatens social cognition research.

Taking up Packer and Moreno-Dulcey's challenge, Kominsky et al. (2022) wrote a thoughtful reply on the unavoidable tradeoff between internal and external validity, emphasizing that finding the right balance is even more challenging for developmental research than for research with adults. Not only does simplicity allow for better experimental control over the measure of interest, but it might also be necessary to tap into infants' conceptual workings because of infants' inherent processing limits. In short, simple stimuli in infancy research are useful because they reduce potential confounds and the risk of masking an underlying ability[1].

[1] This may have occurred in **Experiment 4** of Chapter 3, where there was an effect with simple shapes but not with realistic photographs.

Kominsky et al. cite evidence from the social cognition literature indicating that infants respond to schematic stimuli as they do to live ones in many cases (e.g., preference for schematic faces, puppet imitation, gaze following, and fairness studies). This suggests that they can extract the content embedded in the representational stimuli exactly as the experimenters intend them to. In addition, another paper written in response to the target article reported no effect of stimulus type (puppet versus real person) in a meta-analysis of Theory of Mind tasks between 2 and 6 years of age (Yu & Wellman, 2022).

While I fully agree with this line of reasoning, I would like to point out that the main question that sparked this debate is actually twofold. On the one hand, the question asks whether simplified stimuli tap into what they are meant to measure. From a practical perspective, an affirmative answer to this question is enough to address validity concerns, and I think that Kominsky et al.'s response successfully alleviates most of these worries. If these stimuli convey the construct of interest to infants, research is back on safe epistemic grounds—on condition that researchers test this assumption by comparing schematic stimuli to ecologically valid ones or by validating them with older participants. But there is a second aspect to the question which also requires attention: how exactly do these stimuli convey the construct of interest in infancy and beyond? While not of immediate interest for drawing conclusions about infants' psychological capacities in a specific domain of investigation, I believe the **how** question raises valid theoretical points and calls attention to further cognitive capacities that remain hidden if the focus stays on the **whether** question.

The standard answer to the how question exploits some version of the proper–actual domain distinction (Millikan, 1984; Sperber, 1994): cognitive mechanisms designed to handle a specific class of stimuli in the real world (e.g., the faces of conspecifics) will respond to stimuli outside the class as long as those stimuli satisfy the input conditions of the respective mechanism (e.g., schematic drawings, puppets with eyes)[2]. In the case of agents, intentions, goals, and mental states will be attributed from infancy onward whenever animacy cues are present (Scholl & Tremoulet, 2000; Spelke & Kinzler, 2007). From this perspective, researchers need only strike the right input conditions when designing the stimuli to tap into the cognitive subsystem of interest.

[2] The possibility that children, who were the target age of Packer and Dulcey-Moreno's (2022) critique, mistake puppets for real agents flies in the face of evidence that children's ontological categories are accurate (Asaba et al., 2022; Lillard, 2001, 2022).

There is, however, an alternative to the standard view. Kominsky and his colleagues acknowledge that animations and live puppet shows are depictions of agents and their interactions. This, at least, is how they are generated by the experimenters: physical entities (in this case, puppets) are used to create scenes through which other entities (in this case, fictional agents) and the relations between them are represented. The possibility they do not consider is that this is also how they are interpreted, i.e., as representations. From this perspective, researchers need only provide sufficient iconic or linguistic evidence for the conceptual content of the representation to be extracted by the participants[3]. The wealth of evidence reviewed in **Section 1.2** and the experiments presented in **Chapter 3** show that at least from 15 months onward, infants excel in setting up local relations between symbols and discourse referents—precisely what these tasks require. In addition, Asaba et al. (2022) present evidence that children interpret puppets as social agents only if the experimenter depicts them as such. If the experimenter treats the puppet as an object, children do not attribute any social features to it either.

If animations and puppet shows are interpreted on the same par with pretend play, the claim that infants set up representational relations for such stimuli (at least from some point in development onward) becomes a serious competitor for the standard view, which cannot make sense of pretense at all. For object substitution pretense, it would be difficult to argue that a wooden block satisfies the input conditions of anything. For one, children can accommodate multiple pretend identities of the same object across different contexts and speakers (Wyman et al., 2009).

Other position articles did focus on the how question but defended theses that I think are incomplete. For instance, Lillard (2022) and Wellman and Yu (2022) attack Packer and Dulcey-Moreno's premise that symbolic stimuli introduce a frame of pretense. Both articles argue that children in these studies do not pretend puppets are agents but take puppets as stand-ins for agents (Lillard, 2022; Wellman & Yu, 2022). However, the disagreement between the two camps is illusory because the distinction between pretend play and depictions is vacuous. Object substitution pretense is one manifestation of a general capacity for interpreting external representations, not a separate system. Lillard's (2022) and

[3] These requirements are precisely the ones people must meet when designing depictive representations more broadly: infographics, pedagogical animations, assembly instructions, or scientific figures.

Wellman and Yu's (2022) main argument for a distinction between object substitution pretense and depictions in these papers comes from the fact that toddlers smile only when engaged in pretend play but not when solving a Theory of Mind task in which puppets depict agents. While this difference may distinguish object substitution pretense from other types of representational activities, it cannot be taken to mean that the two are driven by different mental operations, at least not without a theory linking emotional expression to the relevant cognitive subsystem. But if both pretend play and depictions require the capacity to link an object to a discourse referent, they should be treated in tandem.

Rakoczy's position article (2022) comes closest to the view defended here. He rightly points out that the use of puppets in developmental psychology is not exotic but part of a widespread phenomenon that humans take part in all the time, both in the lab (e.g., adult Theory of Mind experiments with computer-generated avatars: Samson et al., 2010) and outside the lab (e.g., when adults engage with pictures, replicas, or movies). He also rightly points out that "even if what we see is their thinking and acting in simulative or off-line mode, quarantined from their serious actions, we see *how* they reason, *which concepts* they use, *which types of inferences* they draw and consider valid" (p. 3).

However, I depart from Rakoczy (2022) in his conception of pretend play and related phenomena. While he acknowledges that pretend play is on a par with pictures, replicas, and movies, he argues that symbol–referent relations are part of a scoped mental representation system that is not linked to the interpreter's beliefs. I pointed out that there is no need to quarantine STAND-FOR relations because quarantining is implicit in the architecture I have advocated, so the hypothesis that symbols are their referents is never entertained.

Finally, Rakoczy (2022) notes an asymmetry between positive and null findings, which went unnoticed in Packer and Dulcey-Moreno's target article. If social cognition studies involve an additional representation layer, positive findings are convincing evidence that the system under study is robust. On the other hand, negative findings are more difficult to interpret because it is unclear whether participants' difficulty comes from a conceptual deficit, from an executive deficit, or from experimenters' failure to convey the construct of interest. However, even positive findings may have intriguing implications of their own. Suppose there exist tasks that are easier to solve if they are mentally represented in a language-of-thought format (even if they are supposed to measure an unrelated domain, such as social cognition). Suppose also—for the sake of argument—

that I am right and that symbolic stimuli are automatically translated into such a format internally. If these assumptions are valid, symbolic stimuli may make the tasks easier for children precisely because they are symbols. As such, they would require only translation from the physical space of objects to the language-of-thought format that discourse referents are already represented in—instead of creating a language-of-thought representation from scratch. Without such guidance, infants and children may be stuck with other strategies that may be unhelpful and therefore fail to solve the tasks. This is a highly speculative proposal, but I believe it is worth further investigation. If it turns out to be true, these tasks will have low external validity, but for a very different reason from the one highlighted in Packer and Dulcey-Moreno (2022).

I would also like to draw attention to several domains of investigation which received little attention in the Theory of Puppets discussion, which mainly dealt with social cognition. As already stated, using puppets and other representational stimuli does not threaten the validity of the conclusions researchers draw based on them. As long as the target cognitive system is activated, it matters less **how** it is activated than **that** it is activated. But ignoring the use of symbolic stimuli may be unwise in inferences about object cognition from experiments that actually use symbols (e.g., Perner & Leahy, 2015). In such cases, Packer and Dulcey-Moreno's point stands strong.

In Chapter 1, I noted several similarities between **mental file** theory and the STAND-FOR architecture I argued for. Here, I would like to go one step further and speculate that **mental file** theory, which has been put forth as an account of object cognition, may, in fact, be an account of symbol interpretation. According to **mental file** theory, a representation of a particular object—a mental file—is headed by a label that captures the perspective under which the object is represented (e.g., "Cicero", "Tully", "the one and only Roman orator"). Properties about the object are written and stored on the file. As far as I know, however, the developmental data that support the **mental file** theory have been gathered exclusively from tasks involving symbols. In these tasks, an experimenter—or a puppet voiced by the experimenter—applies verbal descriptions to props. The descriptions provide the labels for the mental files that children use to represent the props (Doherty & Perner, 1998; Perner et al., 2015; Perner et al., 2011; Wimmer & Perner, 1983). But without a version that does not use symbols, it is impossible to tell whether mental files underlie object cognition or whether the mental file format handles symbol interpretation only.

A second research area in which the use of symbolic stimuli may influence participants' responses comes from studies that use crossovers between autonomous representational media and the surrounding environment (Kinzler et al., 2007; Lucca et al., 2018; Ma & Lillard, 2006). For instance, in a recent study on fairness (Lucca et al., 2018), 13- and 17-month-old infants saw videos on different monitors of two people distributing goods to third parties in distinct ways. Then, the on-screen people turned toward the infant (i.e., toward the camera) while holding an object in their hands and dropped it. The objects seemed to fall into real tubes projecting downward from the screen and extending into trays on the floor. The question of interest was whether infants would prefer to take a toy from one person over another based on the fairness exhibited by the two people before the object-dropping event. Infants preferentially approached the tray under the fair-person monitor, from which researchers concluded that infants prefer to interact with the person who had distributed resources fairly.

Did infants think that the videotaped events happened in the here and now? If yes, this means that infants of this age are confused about the nature of screens and believe that what happens on-screen also continues outside it. I presented evidence against this possibility in **Chapter 2**. If, on the other hand, on-screen events are perceived as separate from the surrounding environment, why did infants move toward one of the two monitors? They could not have thought that what they would find in the trays was the same object the videotaped person dropped. Nor could they have thought they would interact with that person, since nobody was there. Something else must have driven their preferences beyond the desire to accept resources with a partner or to affiliate with them. The assumption that screen stimuli can be used alongside live stimuli as interchangeable can thus be problematic, especially when the conclusions depend on infants perceiving the two as continuous.

Yet another line of research in which the use of symbolic stimuli may interfere with the intended construct are labeling and mislabeling studies (e.g., Dautriche et al., 2021; Dautriche et al., 2022; Koenig & Echols, 2003; Koenig & Woodward, 2010). In a typical study, infants or toddlers are presented with live or on-screen objects sequentially (mis)labeled by a live experimenter or by a disembodied voice from a speaker. In virtually all of these studies, labeling an object with a noun phrase is assumed to be interpreted as a referential act that picks out the perceptually available object. Mislabeling, on the other hand, is assumed to be interpreted as a failed referential act. This failure should lead in-

fants and toddlers to infer that the mislabeling speaker is an unreliable source. However, both labeling and mislabeling events might be interpreted not as referential actions but as explicit stipulations that provide the conceptual identity of the discourse referents that the symbols stand for. In **Chapter 3**, I presented a series of experiments providing evidence supporting this hypothesis.

Finally, symbolic stimuli are also prevalent in psychological research on adult visual perception (e.g., Hafri & Firestone, 2021; Hafri et al., 2023; Konkle & Oliva, 2012). In these cases, participants are often shown 2D images on the screen (by an experimenter) under the same tacit assumption that governs the use of puppets in developmental research—that depiction interpretation is equivalent to visual perception. In **Chapter 4**, I presented a series of Stroop experiments indicating that the interchangeability assumption may not be warranted in adult empirical work either. If adults interpret even realistic photographs as symbols, they likely interpret other types of stimuli as symbolic too. For instance, Hafri et al. (2023) presented adults with short animated clips depicting symmetrical or asymmetrical events (e.g., collision vs. launching). After each clip, they were asked to choose between a symmetrical and asymmetrical predicate with related meanings (e.g., negotiate vs. propose). Because participants chose matching predicates based on symmetry, the authors concluded that language and perception have both access to an abstract concept of symmetry. But suppose participants did not interpret the experimental stimuli only through visual processes but as symbolic representations driven by STAND-FOR relations. In that case, Hafri et al.'s task can no longer be taken to establish an equivalence between perception and language. Instead, participants would have selected the predicate that best conveyed the symbolic content of the animation. The finding that adults understood symmetry as the relevant dimension for matching is not trivial under any account, but the fact that such stimuli are interpreted as depictions undermines the conclusion that symmetry is encoded in vision.

In sum, the recent debate over the nature of the stimuli used in developmental research is highly relevant from both a methodological and theoretical perspective. While the strong claim that conclusions drawn from lab research are invalidated simply by the use of schematic depictions (the whether question) is probably misguided, ignoring the symbolic nature of stimuli is not wise either. This is especially problematic in the case of findings that lump together ordinary objects and symbols, as the theoretical models that the findings are used to bolster may inadvertently have a different scope from the intended one.

5.3. Theoretical Outlook

To sum up, this dissertation presented evidence from a variety of tasks, testing both infants and adults, that symbols and representations are special in human cognition. Infants decouple animated events from the surrounding environment but accept the same depiction across multiple screens. Infants can also interpret labeling events applied to neutral shapes as stipulating STAND-FOR relations between the shapes and the discourse referents. Adults automatically interpret photographs of objects as symbols of (possibly other) objects and shift their interpretation depending on what other symbols are displayed on the screen. Taken together, these lines of evidence point to a cognitive mechanism dedicated to the representation and interpretation of symbolic objects.

At the same time, the architecture in **Chapter 1** needs to be revised in at least two ways in light of the empirical findings in Chapters 2–4. First, I focused on location as the primary criterion by which symbols are indexed, but two lines of evidence suggest that visual features might trump location, all else equal. When location was pitted against visual features in **Experiment 4: Aquarium** of Chapter 2, infants prioritized the visual background of the animation rather than the screen on which it was presented. **Experiment 2: Identical Symbols** in Chapter 3 also suggests that infants find symbol–discourse-referent assignments more difficult when other objects in the scene are visually indistinguishable from the target symbol. This may have happened because they assumed the stipulation holds at the level of symbol kinds (e.g., blue blobs stand for ducks), but it could also have been because distinctive visual features are needed to keep the symbols apart. Thus, while spatiotemporal tracking is required locally for translating the movements of symbols into predicates of the corresponding discourse referents, visual features may matter more for individuation and recognition in the long term. Future work should better specify how location and features interact in the internal representation of individual symbols.

Second, in **Chapter 1**, I speculated that there are at least four ways in which the conceptual identity of the discourse referent is established: (i) visual features—how the symbol looks like; (ii) behavior—how the symbol acts; (iii) linguistic stipulation; and (iv) convention. In **Chapter 4**, I presented evidence that the semantic interpretation of a symbol is sensitive to the other symbols that are part of a representation. Adults interpreted a photo of a toy horse as a horse when displayed next to a neutral object but as a toy horse when displayed

next to a horse. This may seem trivial in retrospect. After all, a circle is interpreted as a sun when surrounded by cloud-shaped objects but as a wheel when attached to a car-shaped object. However, the finding that context plays a role even with realistic photos, where the richness of visual cues mitigates the need to disambiguate, is not at all trivial. Thus, the three paths to figuring out what a symbol stands for—visual features, behavior, and linguistic stipulation—must be augmented by an additional path—the other constituent symbols of the representation. This opens the door for linking STAND-FOR relations to the pragmatics of human communication writ large (see Tieu et al., 2019, for a first step in this direction).

There also remain many open theoretical questions on which the current work is silent. I will mention one which I think is promising to pursue, but there are, to be sure, many more. The theoretical and experimental work presented here focused on the semantics of individual symbols. But symbols are rarely used in isolation. More often, they interact with other entities of the same type in lawful ways to form larger representational units. In other words, symbols have not only a semantics but a syntax too. Syntax arose tangentially in discussing how symbols can be manipulated to convey propositions—by moving the symbols in physical space, which corresponds to applying predicates to the discourse referents the symbols stand for. But while this involves a notion of structure, predicates and propositions still fall under the scope of semantics. I think it would be worth finding out whether humans have intuitions of ungrammaticality in the domain of visual symbols, barring syntactic combinations (spatiotemporal arrangements) that should be legal by semantic standards. One starting point for investigating this question could be sign languages, in which physical space plays a crucial role too (e.g., Schlenker, 2022).

Another promising avenue into the relations between syntax and semantics would be to look at the internal syntactic representations of the symbols themselves and check whether they systematically interface with semantic interpretation. Recent research indicates that human adults represent a variety of stimuli by means of syntactic structures in a language of thought (visual and auditory temporal sequences: Amalric et al., 2017; Planton et al., 2021; simple geometric shapes: Sablé-Meyer et al., 2022; quadrilateral figures: Sablé-Meyer et al., 2021) In particular, Sablé-Meyer and colleagues (2022) found evidence that adults represent simple shapes as generative programs in a language of thought. The programs consist of sequences of commands, which mirror the movement

of a pen on paper. They are compositionally built from a small set of primitive functions (e.g., trace, turn, move, repeat) and parameters (e.g., duration and speed of drawing, number of repetitions). While any given shape can be drawn by an infinity of programs, adults seem to infer the shortest program that generates it, as shown by their encoding and choice times in a match-to-sample task.

One outstanding question about the language of thought proposed by Sablé-Meyer et al. (2022) concerns its semantics. After all, simple shapes are often used as symbols in human communication, so adults should be able to easily assign meanings to them. But the fact that shapes have meaning does not imply that the language of thought used to represent them has one as well. Thus, on the one hand, the programs that adults seem to infer when shown these shapes might be purely formal. The programs are only employed to represent the shapes in a compact format, perhaps for memory efficiency reasons. On the other hand, if these shapes are interpreted as symbols, adults should use their form, as well as their arrangement in relation to other shapes, to retrieve their meaning. In this case, the syntax of the inferred program should inform what the drawing stands for.

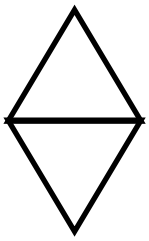

Figure 5.1. A pyramid and its reflection, or a diamond crossed by a rod?

Consider the structurally ambiguous drawing in Figure 5.1 (after Van Sommers, 1984) and suppose you witnessed its creation: one triangle followed by another triangle (as opposed to a rhombus followed by a horizontal line). If made to choose between the two titles in the caption to Figure 5.1, you would probably go for the **pyramid and its reflection** as a better description of the drawing than the **diamond crossed by a rod**. This choice cannot rely only on the end state of the drawing, so it must be that syntax determines which interpretation is more plausible. Once the syntax of the inferred program is overt, the meanings that rely on other parsing trees are out. Therefore, the syntactic structure of visual symbols might provide important constraints on candidate meanings, just as it does in natural language (Chomsky, 1957; Montague, 1970).

Ultimately, the end goal of the research program initiated in this dissertation is a cognitive theory that can explain the semantics of external symbols, their syntax, and the interplay between the two in internal representation and interpretation. The theory should also specify the activation conditions of the posited cognitive system, its format of internal representation (e.g., language of thought), and its connection to natural language. Until then, the best I can hope for is that the first steps taken in this dissertation have placed this research project on firm footing and successfully motivated its further pursuit.

References

Adair, H. V., & Carruthers, P. (2022). Pretend play: More imitative than imaginative. *Mind & Language*, *38*(2), 464–479. [link]

Amalric, M., Wang, L., Pica, P., Figueira, S., Sigman, M., & Dehaene, S. (2017). The language of geometry: Fast comprehension of geometrical primitives and rules in human adults and preschoolers. *PLoS Computational Biology*, *13*(1), e1005273. [link]

Andrasi, K., Schvajda, R., & Kiraly, I. (2022). Young children expect pretend object identities to be known only by their partners in joint pretence. *British Journal of Developmental Psychology*, *40*(3), 398–409. [link]

Asaba, M., Li, X., Yow, W. Q., & Gweon, H. (2022). Children selectively demonstrate their competence to a puppet when others depict it as an agent. *Cognitive Development*, *62*. [link]

Baer, C., & Friedman, O. (2016). Children's generic interpretation of pretense. *Journal of Experimental Child Psychology*, *150*, 99–111. [link]

Baker, R. K., Pettigrew, T. L., & Poulin-Dubois, D. (2014). Infants' ability to associate motion paths with object kinds. *Infant Behavior and Development*, *37*(1), 119–129. [link]

Baldwin, D. A., & Markman, E. M. (1989). Establishing word-object relations: A first step. *Child Development*, *60*(2), 381–398. [link]

Bastos, A. P. M., Wood, P. M., & Taylor, A. H. (2021). Are parrots naive realists? Kea behave as if the real and virtual worlds are continuous. *Biology Letters*, *17*(9), 20210298. [link]

Ben-Shachar, M., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, *5*(56), 2815. [link]

Bonatti, L., Frot, E., Zangl, R., & Mehler, J. (2002). The human first hypothesis: Identification of conspecifics and individuation of objects in the young infant. *Cognitive Psychology*, *44*(4), 388–426. [link]

Bosco, F. M., Friedman, O., & Leslie, A. M. (2006). Recognition of pretend and real actions in play by 1- and 2-year-olds: Early success and why they fail. *Cognitive Development*, *21*(1), 3–10. [link]

Bourchier, A., & Davis, A. (2002). Children's understanding of the pretence-reality distinction: a review of current theory and evidence. *Developmental Science*, *5*(4), 397–413. [link]

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482–1493. [link]

Brody, G. (2020). *Indexing objects in vision and communication* [Doctoral dissertation, Central European University]. CEU Cognitive Science Department Repository. [link]

Buresh, J. S., & Woodward, A. L. (2007). Infants track action goals within and across agents. *Cognition*, *104*(2), 287–314. [link]

Carey, S. (2009). *The origin of concepts*. Oxford University Press. [link]

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, *76*. [link]

Chomsky, N. (1957). *Syntactic structures*. De Gruyter Mouton. [link]

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press. [link]

Clark, H. H. (2016). Depicting as a method of communication. *Psychological Review*, *123*(3), 324–347. [link]

Cohen, L. B., & Amsel, G. (1998). Precursors to infants' perception of the causality of a simple event. *Infant Behavior and Development*, *21*(4), 713–731. [link]

Cosmides, L., & Tooby, J. (2000). Consider the source: Adaptations for decoupling and metarepresentation. In D. Sperber (Ed.), *Metarepresentations: A multidisciplinary perspective* (pp. 53–115). Oxford University Press.

Crane, T. (2003). *The mechanical mind: A philosophical introduction to minds, machines and mental representation*. Routledge.

Csibra, G. (2010). Recognizing communicative intentions in infancy. *Mind & Language*, *25*(2), 141–168. [link]

Csibra, G., Hernik, M., Mascaro, O., Tatone, D., & Lengyel, M. (2016). Statistical treatment of looking-time data. *Developmental Psychology*, *52*(4), 521–536. [link]

Csibra, G., & Shamsudheen, R. (2015). Nonverbal generics: Human infants interpret objects as symbols of object kinds. *Annual Review of Psychology*, *66*(1), 689–710. [link]

Csink, V., Mareschal, D., & Gliga, T. (2021). Does surprise enhance infant memory? Assessing the impact of the encoding context on subsequent object recognition. *Infancy*, *26*(2), 303–318. [link]

Dautriche, I., Goupil, L., Smith, K., & Rabagliati, H. (2021). Knowing how you know: Toddlers reevaluate words learned from an unreliable speaker. *Open Mind: Discoveries in Cognitive Science*, *5*, 1–19. [link]

Dautriche, I., Goupil, L., Smith, K., & Rabagliati, H. (2022). Two-year-olds' eye movements reflect confidence in their understanding of words. *Psychological Science*, *33*(11), 1842–1856. [link]

Deacon, T. (1997). *The symbolic species: The co-evolution of language and the brain*. W.W. Norton & Co.

Dehaene, S. (2009). *Reading in the brain: The new science of how we read*. Penguin Books.

Dehaene, S., Al Roumi, F., Lakretz, Y., Planton, S., & Sable-Meyer, M. (2022). Symbols and mental programs: A hypothesis about human singularity. *Trends in Cognitive Science*, *26*(9), 751–766. [link]

DeLoache, J. S. (1987). Rapid change in the symbolic functioning of very young children. *Science*, *238*(4833), 1556–1557. [link]

DeLoache, J. S. (1991). Symbolic functioning in very young children: Understanding of pictures and models. *Child Development*, *62*(4). [link]

DeLoache, J. S. (2004). Becoming symbol-minded. *Trends in Cognitive Sciences*, *8*(2), 66–70. [link]

DeLoache, J. S., & Burns, N. M. (1994). Early understanding of the representational function of pictures. *Cognition*, *52*(2), 83–110. [link]

DeLoache, J. S., Pierroutsakos, S. L., Uttal, D. H., Rosengren, K. S., & Gottlieb, A. (1998). Grasping the nature of pictures. *Psychological Science*, *9*(3), 205–210. [link]

DeLoache, J. S., & Sharon, T. (2005). Symbols and similarity: You can get too much of a good thing. *Journal of Cognition and Development*, *6*(1), 33–49. [link]

DeLoache, J. S., Strauss, M. S., & Maynard, J. (1979). Picture perception in infancy. *Infant Behavior and Development*, *2*, 77–89. [link]

Doherty, M., & Perner, J. (1998). Metalinguistic awareness and theory of mind: Just two words for the same thing? *Cognitive Development*, *13*(3), 279–305. [link]

Fodor, J. A. (1975). *The language of thought*. Harvard University Press.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–585. [link]

Friedman, O. (2013). How do children represent pretend play? In M. Taylor (Ed.), *The Oxford handbook of the development of imagination* (pp. 186–195). Oxford University Press. [link]

Friedman, O., & Leslie, A. M. (2007). The conceptual underpinnings of pretense: Pretending is not 'behaving-as-if'. *Cognition*, *105*(1), 103–124. [link]

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. MIT Press. [link]

Gelman, A. (2018, March 15). You need 16 times the sample size to estimate an interaction than to estimate a main effect. *Statistical Modeling, Causal Inference, and Social Science*. [link]

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165–193. [link]

Goldin-Meadow, S. (2005). Symbolic communication without a language model: The starting point for language learning. In L. Namy (Ed.), *Symbol use and symbolic representation: Developmental and comparative perspectives* (pp. 101–122). Lawrence Erlbaum. [link]

Golomb, C., & Galasso, L. (1995). Make believe and reality: Explorations of the imaginary realm. *Developmental Psychology*, *31*(5), 800–810. [link]

Gomila, R. (2021). Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis. *Journal of Experimental Psychology: General*, *150*(4), 700–709. [link]

Goodman, N. (1968). *Languages of art: An approach to a theory of symbols*. Bobbs-Merrill.

Greenberg, G. (2013). Beyond resemblance. *The Philosophical Review*, *122*(2), 215–287. [link]

Grimm, H., & Doil, H. (2019). *ELFRA. Elternfragebögen für die Früherkennung von Risikokindern.* [Parent Questionnaire for Early Detection of Children at Risk]. Göttingen: Hogrefe.

Hafri, A., & Firestone, C. (2021). The perception of relations. *Trends in Cognitive Sciences*, *25*(6), 475–492. [link]

Hafri, A., Gleitman, L. R., Landau, B., & Trueswell, J. C. (2023). Where word and world meet: Language and vision share an abstract representation of symmetry. *Journal of Experimental Psychology: General*, *152*(2), 509–527. [link]

Halberda, J. (2003). The development of a word-learning strategy. *Cognition*, *87*(1), B23–34. [link]

Harris, P. L. (2000). *The work of the imagination*. Wiley-Blackwell.

Harris, P. L. (2021). Early constraints on the imagination: The realism of young children. *Child Development*, *92*(2), 466–483. [link]

Harris, P. L., & Kavanaugh, R. D. (1993). Young children's understanding of pretense. *Monographs of the Society for Research in Child Development*, *58*(1), v–92. [link]

Harris, P. L., Kavanaugh, R. D., & Dowson, L. (1997). The depiction of imaginary transformations: Early comprehension of a symbolic function. *Cognitive Development*, *12*(1), 1–19. [link]

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, *57*, 243–259. [link]

Heim, I. (1982). *The semantics of definite and indefinite noun phrases* [Unpublished doctoral dissertation]. University of Massachusetts Amherst ProQuest Dissertations Publishing. [link]

Hernik, M., Fearon, P., & Csibra, G. (2014). Action anticipation in human infants reveals assumptions about anteroposterior body-structure and action. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1781), 20133205. [link]

Hochberg, J. (1986). Representation of motion and space in video and cinematic displays. In L. K. K.R. Boff, & J.P. Thomas (Ed.), *Handbook of perception and human performance* (Vol. 1, pp. 22.21–22.64). Wiley-Interscience.

Hochberg, J., & Brooks, V. (1962). Pictorial recognition as an unlearned ability: A study of one child's performance. *The American Journal of Psychology*, *75*(4), 624–628. [link]

Hopkins, E. J., Dore, R. A., & Lillard, A. S. (2015). Do children learn from pretense? *Journal of Experimental Child Psychology*, *130*, 1–18. [link]

Huang, Y., & Snedeker, J. (2020). Evidence from the visual world paradigm raises questions about unaccusativity and growth curve analyses. *Cognition*, *200*, 104251. [link]

Ittelson, W. H. (1996). Visual perception of markings. *Psychonomic Bulletin and Review*, *3*(2), 171–187. [link]

Jaswal, V. K., & Markman, E. M. (2007). Looks aren't everything: 24-month-olds' willingness to accept unexpected labels. *Journal of Cognition and Development*, *8*(1), 93–111. [link]

Kabdebon, C., & Dehaene-Lambertz, G. (2019). Symbolic labeling in 5-month-old human infants. *Proceedings of the National Academy of Sciences*, *116*(12), 5805–5810. [link]

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, *24*(2), 175–219. [link]

Kaminski, J., & Nitzschner, M. (2013). Do dogs get the point? A review of dog–human communication ability. *Learning and Motivation*, *44*(4), 294–302. [link]

Kamp, H., & Reyle, U. (1993). *From discourse to logic*. Springer Dordrecht. [link]

Karttunen, L. (1976). Discourse referents. In J. D. McCawley (Ed.), *Notes from the linguistic underground* (pp. 363–385). Brill. [link]

Kavanaugh, R. D., & Harris, P. L. (1994). Imagining the outcome of pretend transformations: Assessing the competence of normal children and children with autism. *Developmental Psychology*, *30*(6), 847–854. [link]

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLoS ONE*, *7*(5), e36399. [link]

Kinney, J. (1965). Effect of exposure duration on induced color. *Journal of the Optical Society of America*, *55*, 731–736. [link]

Kinzler, K. D., Dupoux, E., & Spelke, E. S. (2007). The native language of social cognition. *Proceedings of the National Academy of Sciences*, *104*(30), 12577–12580. [link]

Koenig, M. A., & Echols, C. H. (2003). Infants' understanding of false labeling events: The referential roles of words and the speakers who use them. *Cognition*, *87*(3), 179–208. [link]

Koenig, M. A., & Woodward, A. L. (2010). Sensitivity of 24-month-olds to the prior inaccuracy of the source: Possible mechanisms. *Developmental Psychology*, *46*(4), 815–826. [link]

Kominsky, J. F., Lucca, K., Thomas, A. J., Frank, M. C., & Hamlin, J. K. (2022). Simplicity and validity in infant research. *Cognitive Development*, *63*. [link]

Konkle, T., & Oliva, A. (2011). Canonical visual size for real-world objects. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 23–37. [link]

Konkle, T., & Oliva, A. (2012). A familiar-size Stroop effect: Real-world size is an automatic property of object representation. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(3), 561–569. [link]

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press. [link]

Kuhlmeier, V., Wynn, K., & Bloom, P. (2003). Attribution of dispositional states by 12-month-olds. *Psychological Science*, *14*(5), 402–408. [link]

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*, *3*(3), 299–321. [link]

Leahy, B. P., & Carey, S. E. (2020). The acquisition of modal concepts. *Trends in Cognitive Science*, *24*(1), 65–78. [link]

Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind.". *Psychological Review*, *94*(4), 412–426. [link]

Leslie, A. M., & Happé, F. (1989). Autism and ostensive communication: The relevance of metarepresentation. *Development and Psychopathology*, *1*(3), 205–212. [link]

Lillard, A. (2001). Pretend play as Twin Earth: A social-cognitive analysis. *Developmental Review*, *21*(4), 495–531. [link]

Lillard, A., Pinkham, A. M., & Smith, E. (2011). Pretend play and cognitive development. In U. Goswami (Ed.), *The Wiley-Blackwell handbook of childhood cognitive development* (2nd ed., pp. 285–311). Wiley-Blackwell. [link]

Lillard, A. S. (2022). Pretending at hand: How children perceive and process puppets. *Cognitive Development*, *63*. [link]

Lillard, A. S., & Witherington, D. C. (2004). Mothers' behavior modifications during pretense and their possible signal value for toddlers. *Developmental Psychology*, *40*(1), 95–113. [link]

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041. [link]

Long, B., & Konkle, T. (2017). A familiar-size Stroop effect in the absence of basic-level recognition. *Cognition*, *168*, 234–242. [link]

Lucca, K., Pospisil, J., & Sommerville, J. A. (2018). Fairness informs social decision making in infancy. *PLoS ONE*, *13*(2), Article e0192848. [link]

Luchkina, E., & Waxman, S. (2021). Acquiring verbal reference: The interplay of cognitive, linguistic, and general learning capacities. *Infant Behavior and Development*, *65*, 101624. [link]

Ma, L., & Lillard, A. S. (2006). Where is the real cheese? Young children's ability to discriminate between real and pretend acts. *Child Development*, *77*(6), 1762–1777. [link]

Mandler, J. M., & McDonough, L. (1993). Concept formation in infancy. *Cognitive Development*, *8*(3), 291–318. [link]

Manea, V., Kampis, D., Grosse Wiesmann, C., Revencu, B., & Southgate, V. (2023). An initial but receding altercentric bias in preverbal infants' memory. *Proceedings of the Royal Society B: Biological Sciences*, *290*(2000), 20230738. [link]

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77–80. [link]

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121–157. [link]

Mascaro, O., & Sperber, D. (2009). The moral, epistemic, and mindreading components of children's vigilance towards deception. *Cognition*, *112*(3), 367–380. [link]

Matthews, D., Lieven, E., Theakston, A., & Tomasello, M. (2006). The effect of perceptual availability and prior discourse on young children's use of referring expressions. *Applied Psycholinguistics*, *27*(3), 403–422. [link]

Matthews, D., Lieven, E., & Tomasello, M. (2010). What's in a manner of speaking? Children's sensitivity to partner-specific referential precedents. *Developmental Psychology*, *46*(4), 749–760. [link]

Mazzocco, M. M. (1997). Children's interpretations of homonyms: A developmental study. *Journal of Child Language*, *24*(2), 441–467. [link]

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). Chapman and Hall/CRC. [link]

Millikan, R. G. (1984). *Language, thought, and other biological categories*. MIT Press.

Millikan, R. G. (2017). *Beyond concepts: Unicepts, language and natural Information*. Oxford University Press. [link]

Montague, R. (1970). Universal grammar. *Theoria*, *36*(3), 373–398. [link]

Müller, C. (2013). Gestural modes of representation as techniques of depiction. In C. Müller, A. Cienki, L. S., D. McNeill, & J. Bressem (Eds.), *Body – language – communication: an international handbook on multimodality in human interaction* (pp. 1687–1701). De Gruyter Mouton. [link]

Mussavifard, N. (2023). *The pedagogical origin of human communication* [Doctoral dissertation, Central European University]. CEU Cognitive Science Department Repository. [link]

Newcombe, N., Huttenlocher, J., & Learmonth, A. (1999). Infants' coding of location in continuous space. *Infant Behavior and Development*, *22*(4), 483–510. [link]

Nichols, S., & Stich, S. (2000). A cognitive theory of pretense. *Cognition*, *74*(2), 115–147. [link]

Oakes, L. M., & Cohen, L. B. (1990). Infant perception of a causal event. *Cognitive Development*, *5*(2), 193–207. [link]

Oatley, K., & Yuill, N. (1985). Perception of personal and interpersonal action in a cartoon film. *British Journal of Social Psychology*, *24*(2), 115–124. [link]

Offside (association football). (2021, March 24). In *Wikipedia*. [link]

Onishi, K. H., Baillargeon, R., & Leslie, A. M. (2007). 15-month-old infants detect violations in pretend scenarios. *Acta Psychologica*, *124*(1), 106–128. [link]

Opfer, J. E., & Gelman, S. A. (2011). Development of the animate-inanimate distinction. In U. Goswami (Ed.), *The Wiley-Blackwell handbook of childhood cognitive development* (2nd ed., pp. 213–238). [link]

Packer, M. J., & Moreno-Dulcey, F. A. (2022). Theory of puppets?: A critique of the use of puppets as stimulus materials in psychological research with young children. *Cognitive Development*, *61*. [link]

Parise, E., & Csibra, G. (2012). Electrophysiological evidence for the understanding of maternal speech by 9-month-old infants. *Psychological Science*, *23*(7), 728–733. [link]

Paulus, M., Caporaso, J., & Marcovitch, S. (2022). The use of puppets in developmental science: Possibilities, limitations, and prospects [Special issue]. *Cognitive Development*.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Hochenberger, R., Sogo, H., Kastman, E., & Lindelov, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. [link]

Perner, J. (1991). *Understanding the representational mind*. MIT Press. [link]

Perner, J., Huemer, M., & Leahy, B. (2015). Mental files and belief: A cognitive theory of how children represent belief and its intensionality. *Cognition*, *145*, 77–88. [link]

Perner, J., & Leahy, B. (2015). Mental files in development: Dual naming, false belief, identity and intensionality. *Review of Philosophy and Psychology*, *7*(2), 491–508. [link]

Perner, J., Mauer, M. C., & Hildenbrand, M. (2011). Identity: Key to children's understanding of belief. *Science*, *333*(6041), 474–477. [link]

Piaget, J. (1962). *Play, dreams, and imitations in childhood* (G. Gattegno & F. M. Hodgson, Trans.). Norton. (Original work published 1945)

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, *123*(4), 392–424. [link]

Pierroutsakos, S. L., & Troseth, G. L. (2003). Video verité: Infants' manual investigation of objects on video. *Infant Behavior and Development*, *26*(2), 183–199. [link]

Pietroski, P. M. (2018). *Conjoining meanings: Semantics without truth values*. Oxford University Press. [link]

Planton, S., van Kerkoerle, T., Abbih, L., Maheu, M., Meyniel, F., Sigman, M., Wang, L., Figueira, S., Romano, S., & Dehaene, S. (2021). A theory of memory for binary sequences: Evidence for a mental compression algorithm in humans. *PLoS Computational Biology*, *17*(1), e1008598. [link]

Politzer, G. (2004). Reasoning, judgement and pragmatics. In I. A. Noveck & D. Sperber (Eds.), *Experimental pragmatics* (pp. 94–115). Palgrave. [link]

Pomiechowska, B., Brody, G., Csibra, G., & Gliga, T. (2021). Twelve-month-olds disambiguate new words using mutual-exclusivity inferences. *Cognition*, *213*, 104691. [link]

Powell, L. J., & Spelke, E. S. (2013). Preverbal infants expect members of social groups to act alike. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(41), E3965–E3972. [link]

Preissler, M. A., & Bloom, P. (2007). Two-year-olds appreciate the dual nature of pictures. *Psychological Science*, *18*(1), 1–2. [link]

Pylyshyn, Z. W. (1989). The role of location indexes in spatial perception: a sketch of the FINST spatial-index model. *Cognition*, *32*(1), 65–97. [link]

Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition*, *80*(1–2), 127–158. [link]

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spatial Vision*, *3*(3), 179–197. [link]

Quilty-Dunn, J. (2020). Polysemy and thought: Toward a generative theory of concepts. *Mind & Language*, *36*(1), 158–185. [link]

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. [link]

Rakison, D. H., Cicchino, J. B., & Hahn, E. R. (2007). Infants' knowledge of the path that animals take to reach a goal. *British Journal of Developmental Psychology*, *25*(3), 461–470. [link]

Rakoczy, H. (2022). Puppet studies present clear and distinct windows into the child's mind. *Cognitive Development*, *61*. [link]

Rakoczy, H., Tomasello, M., & Striano, T. (2005). How children turn objects into symbols: A cultural learning account. In L. Namy (Ed.), *Symbol use and symbolic representation: Developmental and comparative perspectives* (pp. 69–100). Lawrence Erlbaum. [link]

Rakoczy, H., Tomasello, M., & Striano, T. (2006). The role of experience and discourse in children's developing understanding of pretend play actions. *British Journal of Developmental Psychology*, *24*(2), 305–335. [link]

Recanati, F. (2012). *Mental files*. Oxford University Press. [link]

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439. [link]

Royka, A., Chen, A., Aboody, R., Huanca, T., & Jara-Ettinger, J. (2022). People infer communicative action through an expectation for efficient communication. *Nature Communications*, *13*(1), 4160. [link]

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson Education. [link]

Sablé-Meyer, M., Ellis, K., Tenenbaum, J., & Dehaene, S. (2022). A language of thought for the mental representation of geometric shapes. *Cognitive Psychology*, *139*, 101527. [link]

Sablé-Meyer, M., Fagot, J., Caparos, S., van Kerkoerle, T., Amalric, M., & Dehaene, S. (2021). Sensitivity to geometric shape regularity in humans and baboons: A putative signature of human singularity. *Proceedings of the National Academy of Sciences*, *118*(16). [link]

Sagan, C. (1973). *The cosmic connection: An extraterrestrial perspective*. Cambridge University Press.

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: Evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(5), 1255–1266. [link]

Schlenker, P. (2022). *What it all means: Semantics for (almost) everything*. MIT Press. [link]

Scholl, B. J., & Gao, T. (2013). Perceiving animacy and intentionality: Visual processing or higher-level judgment? In M. D. Rutherford & V. Kuhlmeier (Eds.), *Social perception: Detection and interpretation of animacy, agency, and intention* (pp. 197–229). MIT Press. [link]

Scholl, B. J., & Leslie, A. M. (1999). Explaining the infant's object concept: Beyond the perception/cognition dichotomy. In E. Lepore & Z. W. Pylyshyn (Eds.), *What is cognitive science?* (pp. 26–73). Wiley-Blackwell.

Scholl, B. J., & Pylyshyn, Z. W. (1999). Tracking multiple items through occlusion: clues to visual objecthood. *Cognitive Psychology*, *38*(2), 259–290. [link]

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*(8), 299–309. [link]

Scott-Phillips, T. (2014). *Speaking our minds: Why human communication is different, and how language evolved to make it special*. Red Globe Press.

Sim, Z. L., & Xu, F. (2019). Another look at looking time: Surprise as rational statistical inference. *Topics in Cognitive Science*, *11*(1), 154–163. [link]

Simcock, G., & DeLoache, J. (2006). Get the picture? The effects of iconicity on toddlers' reenactment from picture books. *Developmental Psychology*, *42*(6), 1352–1357. [link]

Snow, J. C., & Culham, J. C. (2021). The treachery of images: How realism influences brain and behavior. *Trends in Cognitive Science*, *25*(6), 506-519. [link]

So, W. C., Kita, S., & Goldin-Meadow, S. (2009). Using the hands to identify who does what to whom: Gesture and speech go hand-in-hand. *Cognitive Science*, *33*(1), 115. [link]

Spelke, E. S. (1990). Principles of object perception. *Cognitive Science*, *14*(1), 29–56. [link]

Spelke, E. S. (2022). *What babies know: Core knowledge and composition* (Vol. 1). Oxford University Press. [link]

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96. [link]

Sperber, D. (1994). The modularity of thought and the epidemiology of representations. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 39-67). Cambridge University Press. [link]

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Blackwell Publishing.

Striano, T., Tomasello, M., & Rochat, P. (2001). Social and object support for early symbolic play. *Developmental Science*, *4*(4), 442–455. [link]

Suddendorf, T. (2003). Early representational insight: Twenty-four-month-olds can use a photo to find an object in the world. *Child Development*, *74*(3), 896–904. [link]

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science*, *18*(7), 580–586. [link]

Sutherland, S. L., & Friedman, O. (2012). Preschoolers acquire general knowledge by sharing in pretense. *Child Development*, *83*(3), 1064–1071. [link]

Tatone, D., Geraci, A., & Csibra, G. (2015). Giving and taking: Representational building blocks of active resource-transfer events in human infants. *Cognition*, *137*, 47–62. [link]

Tauzin, T., & Gergely, G. (2018). Communicative mind-reading in preverbal infants. *Scientific Reports*, *8*(1), Article 9534. [link]

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285. [link]

Tieu, L., Schlenker, P., & Chemla, E. (2019). Linguistic inferences without words. *Proceedings of the National Academy of Sciences*, *116*(20), 9796–9801. [link]

Tomasello, M., Call, J., & Gluckman, A. (1997). Comprehension of novel communicative signs by apes and human children. *Child Development*, *68*(6). [link]

Tomasello, M., Striano, T., & Rochat, P. (1999). Do young children use objects as symbols? *British Journal of Developmental Psychology*, *17*(4), 563–584. [link]

Troseth, G. L., & DeLoache, J. S. (1998). The medium can obscure the message: Young children's understanding of video. *Child Development*, *69*(4), 950–965. [link]

Van Sommers, P. (1984). *Drawing and cognition: Descriptive and experimental studies of graphic production processes*. Cambridge University Press. [link]

Vygotsky, L. (2016). Play and its role in the mental development of the child (N. Veresov & M. Barrs, Trans.). *International Research in Early Childhood Education*, *7*(2), 3–25. (Original work published 1966) [link]

Walton, K. (1990). *Mimesis as make-believe: On the foundations of the representational arts*. Harvard University Press.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and Widely Applicable Information Criterion in singular learning theory. *Journal of Machine Learning Research*, *11*(12), 3571–3594. [link]

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology*, *29*(3), 257–302. [link]

Weisberg, D. S. (2015). Pretend play. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(3), 249–261. [link]

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, *5*(6), 598–604. [link]

Wellman, H. M., & Yu, C.-L. (2022). Theory of puppets or theory of mind? Misunderstanding how children construe puppets in psychological research: A commentary on Packer and Moreno-Dulcey (2022). *Cognitive Development*, *63*. [link]

Werchan, D. M., Collins, A. G., Frank, M. J., & Amso, D. (2015). 8-month-old infants spontaneously learn and generalize hierarchical rules. *Psychological Science*, *26*(6), 805–815. [link]

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer. [link]

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128. [link]

Woolley, J. D., & Phelps, K. E. (1994). Young children's practical reasoning about imagination. *British Journal of Developmental Psychology*, *12*(1), 53–67. [link]

Wyman, E., Rakoczy, H., & Tomasello, M. (2009). Young children understand multiple pretend identities in their object play. *British Journal of Developmental Psychology*, *27*(2), 385–404. [link]

Yoon, J. M., Johnson, M. H., & Csibra, G. (2008). Communication-induced memory biases in preverbal infants. *Proceedings of the National Academy of Sciences*, *105*(36), 13690–13695. [link]

Yu, C.-L., & Wellman, H. M. (2022). Young Children Treat Puppets and Dolls Like Real Persons in Theory of Mind Research: A meta-analysis of false-belief understanding across ages and countries. *Cognitive Development*, *63*. [link]

Yu, X., & Lau, E. (2023). The binding problem 2.0: Beyond perceptual features. *Cognitive Science*, *47*(2), e13244. [link]

Appendix A. Supplemental Materials to Chapter 3

A1. Counterbalancing in Experiments 1–3

Table A1 shows a counterbalancing order for one participant in Experiment 1. Trial type and the side of the highlighted object were counterbalanced within subjects. The first trial type (Same Word vs. Different Word) and the highlighted side in the first trial (Right vs. Left) were counterbalanced across subjects.

TRIAL NUMBER	PHASE	TRIAL TYPE	VISUAL STIMULI	LABEL STIMULI (HIGHLIGHT/TEST)	HIGHLIGHTED/ CORRECT SIDE
1	Training	Same Word	ball–bottle	bottle/bottle	Right
2	Training	Different Word	duck–book	duck/book	Left
3	Training	Same Word	car–dog	car/car	Left
4	Training	Different Word	banana–cat	banana/cat	Right
5	Experimental	Same Word	diamond–asterisk	bird/bird	Right
6	Experimental	Different Word	triangle–octagon	bed/shoe	Left
7	Experimental	Different Word	clover–spades	ball/dog	Right
8	Experimental	Same Word	star–plus	spoon/spoon	Left
9	Word Knowledge	NA	bottle–bird	–/bird	Right
10	Word Knowledge	NA	bed–duck	–/bed	Left
11	Word Knowledge	NA	spoon–cat	–/spoon	Left
12	Word Knowledge	NA	car–ball	–/ball	Right
13	Word Knowledge	NA	dog–banana	–/dog	Left
14	Word Knowledge	NA	book–shoe	–/shoe	Right

Table A1. Example of a randomization order for one participant in Experiment 1. No label used during Training phase is repeated during the Experimental phase; all labels used during the Experimental phase are tested at the end. The last column specifies the side of the highlighted object for the Training and Experimental phases, and the side of the correct answer for the Word Knowledge phase.

Experiment 2 had an identical counterbalancing design. The only difference from Experiment 1 was that infants saw identical-looking objects in Experimental trials. Experiment 3A and Experiment 3B were identical to Experiment 1, except for introducing a new speaker and trial block (Trials 15–18). For details, see the **Methods** section of Experiment 3A in Chapter 3.

A2. Bayesian Model for Experiments 1–3

MODEL SPECIFICATION

To better estimate infants' behavior in Experiments 1–3, I ran several variants of models that included all the data. The models were identical in structure to the preregistered one, except that instead of modeling looks to the highlighted object at each time point with a growth curve analysis, I modeled the proportion of looking at the highlighted object with trial type, individual subject, baseline proportion of looking at the highlighted object, and word knowledge scores as predictors.

The **complete model** assumes that the average proportion of looking at the target object (PLT) over the entire test period in each trial is sampled with noise from a normal distribution with mean μ standard deviation σ. The assumption that proportions are drawn from a normal distribution might seem odd. However, there are at least three reasons which support this decision. First, the assumption does not imply that proportions are truly normally distributed (McElreath, 2020). Second, linear regression is as accurate as logistic regression even when the outcome is on a binary scale, let alone proportions (Gomila, 2021). Third, a beta regression was inappropriate since PLHs were 0 or 1 on many trials, and beta distributions are defined only on the open interval (0, 1).

Then, for each PLH observation, the model estimates the true mean μ_i as a function of condition, the infant who produced that observation, their word knowledge for the stipulated label (separate coefficients for each condition), and their baseline preference for the highlighted object (separate coefficients for each condition). There are eight conditions (two in Experiment 1: Same Word and Different Word; two in Experiment 2; and four in Experiment 3 because trial type is crossed with speaker identity), each receiving its own estimate. Both word knowledge scores and baseline preferences were mean-centered to increase coefficient interpretability. To reduce overfitting, the intercepts for individual subjects were estimated with pooling (i.e., the model itself estimates the stan-

dard deviation). All priors and hyperpriors were chosen to be neutral about the direction of any effect and to predict plausible values, since proportions can only be between 0 and 1 (see the **Prior predictive check** subsection below). The specification of the **complete model** is as follows:

Likelihood function $\text{PLH}_i \sim \mathcal{N}(\mu_i, \sigma)$

Link function

$$\mu_i = \beta_{\text{cond}}[\text{condition}] + \beta_{\text{ID}}[\text{ID}] + \beta_{\text{wk}}[\text{condition}] \cdot \text{wk} + \beta_{\text{baseline}}[\text{condition}] \cdot \text{bPLH}$$

Priors

$\beta_{\text{ID}}[\text{ID}] \sim (0, \sigma_{\text{ID}})$ $\beta_{\text{cond}}[\text{condition}] \sim \mathcal{N}(0, 0.1)$

$\beta_{\text{wk}}[\text{condition}] \sim \mathcal{N}(0, 0.1)$ $\beta_{\text{baseline}}[\text{condition}] \sim \mathcal{N}(0, 0.1)$

Hyperpriors

$\sigma \sim \text{Exponential}(8)$ $\sigma_{\text{ID}} \sim \text{Exponential}(8)$

The remaining models are variations on the **complete model** but are otherwise identical. The **baseline model** removes Word Knowledge as a predictor; the **word knowledge model** removes Baseline PLH as a predictor; the **simple model** removes both Word Knowledge and Baseline PLH as predictors; the **no condition model** removes condition as a predictor and estimates a single grand mean; the **grouped model** groups the Same Speaker block in Experiment 3 with (the Same Speaker block in) Experiment 1. These models, as well as the models presented in the following sections, were run with four chains and 2,000 iterations each to obtain 4,000 posterior samples for each model[1].

PRIOR PREDICTIVE CHECK

To check that the model priors are sensible, I simulated data for 128 subjects from the priors of the complete model. **Figure A1** plots the means of the predicted PLHs at test and the 89% credible intervals around it. The model is not biased to find an effect in any condition, and its priors restrict its search space to plausible values only (between 0 and 1).

[1] The first half of the posterior samples is discarded because early samples are not as close to the bulk of the target distribution as later samples.

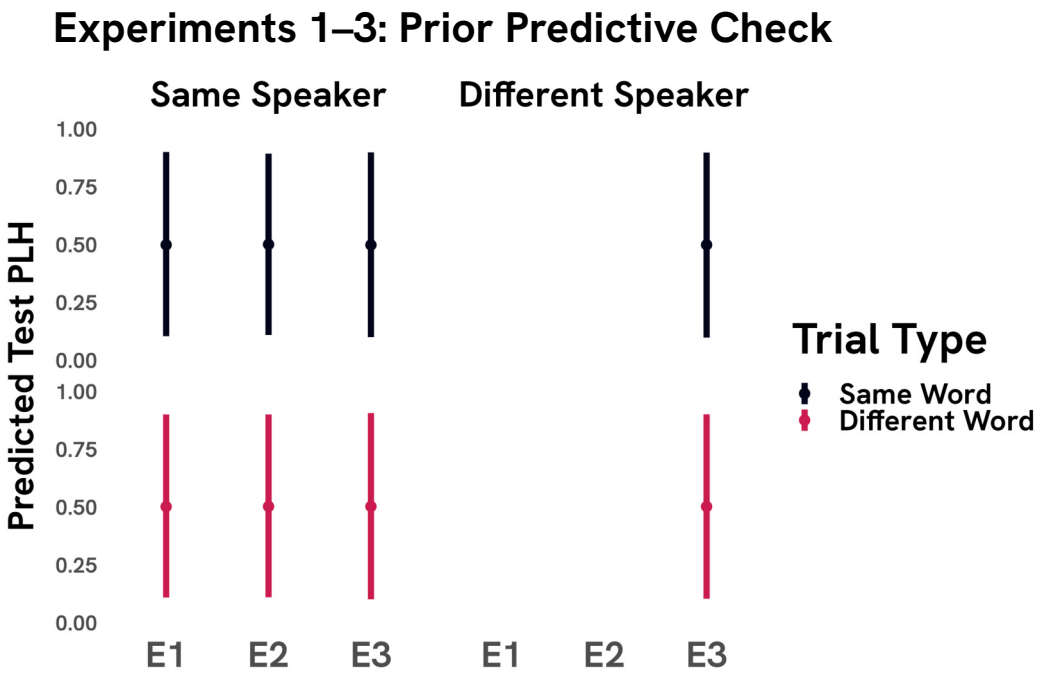

Figure A1. PLH means (circles) and 89% credible intervals (vertical lines) simulated from the priors of the Bayesian model for all experiments and conditions.

MODEL COMPARISON

Table A2 shows the ranking of the models based on the widely applicable information criterion (McElreath, 2020; Watanabe, 2010). Because the **grouped model** receives the highest weight by this criterion, I report its results in the main text. The **simple**, **no condition**, and **word knowledge** (no-baseline) models performed poorly on the dataset. This indicates that condition, baseline, and word knowledge are significant predictors of the PLH dependent variable.

MODEL	WAIC	WEIGHT
Grouped	292.4	0.40
Baseline	292.8	0.32
Complete	293.4	0.23
No condition	296.7	0.05
Word knowledge	322.0	<0.001
Simple	323.4	<0.001

Table A2. Model comparison results, sorted based on the Widely Applicable Information Criterion. Higher weight corresponds to better fit.

A3. Bayesian Vocabulary Model for Experiments 1–3

As preregistered, I ran a Bayesian model to impute the missing data in the Word Knowledge phase (*n* = 18 out of 768 trials: 2.3%). The model assumes that the proportion of looking at the target object (PLT), averaged over the entire test period, is sampled with noise from a normal distribution with mean μ standard deviation σ. Then, for each observation PLT_i, the model estimates the true mean μ_i as a function of the infant who provided data on that trial and the label on which they were tested. In addition, baseline correction is implicitly performed by including the baseline proportion of looks in the regression. The coefficients for individual subjects and words are estimated with pooling (i.e., the model itself estimates the standard deviation).

The hyperpriors on the standard deviations come from an exponential distribution family (which conditions the values to be nonnegative); the parameters were chosen such that the model predicts plausible values for the dependent variable, which should be constrained to be between 0 and 1. Finally, a regularizing prior is used for the influence of baseline preferences, β_{baseline}, so as not to overfit the data. This is the full specification of the model:

Likelihood function $\text{PLT}_i \sim \mathcal{N}(\mu_i, \sigma)$

Link function $\mu_i = \beta_{\text{ID}}[\text{ID}] + \beta_{\text{word}}[\text{testLabel}] + \beta_{\text{baseline}} \cdot \text{bPLH}$

Priors $\beta_{\text{baseline}} \sim (0, 0.1)$ $\beta_{\text{ID}}[\text{ID}] \sim (0, \sigma_{\text{ID}})$
$\beta_{\text{word}}[\text{testLabel}] \sim (0, \sigma_{\text{word}})$

Hyperpriors

$\sigma \sim \text{Exponential}(8)$ $\sigma_{\text{ID}} \sim \text{Exponential}(8)$ $\sigma_{\text{word}} \sim \text{Exponential}(8)$

The posteriors obtained included individual intercepts for both individual subjects and words. This made it straightforward to impute the missing data. Vocabulary scores were computed from the posteriors by means of the linear link function, then averaged to get a score between 0 and 1. For instance, if subject 15 did not provide a valid Word Knowledge trial for “dog”, this would be imputed as $\beta_{\text{ID}}[15] + \beta_{\text{word}}[\text{"dog"}] + \beta_{\text{baseline}} \cdot \text{bPLH}$, where bPLH was either the actual score of the subject on the “dog” trial (when the infant provided valid data at baseline) or imputed as neutral preference (when the entire trial was missing: bPLH = 0.5). These scores replaced infants’ missing data in the dataset.

A4. Bayesian Model for Experiment 4: Moving Symbols

The model used to analyze infants' first looks in Chapter 3 is identical to the one I used to model infants' looking times in Manea et al. (2023). First-look measurements were log-transformed and standardized before the analysis. The model assumed that each data point is sampled from an underlying normal distribution with mean μ and standard deviation σ. The mean μ is modeled as a linear function of several predictors: the infant who produced the measurement, the condition–outcome combination from which that measurement came, the pair of trials, and the outcome order in that pair.

The model assumes that individual subjects' intercepts are sampled from a normal distribution with a unique mean for each condition (4 levels: Training–Congruent, Training–Incongruent, Experimental–Congruent, Experimental–Incongruent) and with a standard deviation that the model estimates. This prevents overfitting by allowing information to flow across subjects when the target distribution is approximated.

I centered the parameters of interest (those representing the effect of condition) on 0 to avoid biasing the model toward finding effects. As in Manea et al. (2023), I assumed the effect size was unlikely to be larger than 0.8; the model, therefore, uses a standard deviation of 0.35 for the parameters corresponding to the four conditions. Since differences are relevant here, a standard deviation of 0.35 for each distribution implies a standard deviation of 0.5 in the distribution of their difference[2]. This means that 89% of the prior distribution on trial type differences will be between −0.8 and 0.8.

In addition, the model incorporates trial pair (1 vs. 2) and order within each pair (Congruent-first vs. Incongruent-first), which are not of immediate interest but could be used to query the additive influence of these variables on infants' looking times[3]. Finally, I use exponential distributions with means of 0.2 for subjects' variability and measurement noise. Prior predictive checks (Figure A2) confirm that the model makes sensible predictions and is not biased toward finding an effect before seeing the data.

---

[2] $\sigma_{d_1-d_2} = \sqrt{\sigma_{d_1}^2 + \sigma_{d_1}^2} = \sqrt{0.35^2 + 0.35^2} \approx 0.495$

[3] The model estimates only the equivalent of main effects of trial pair and order, not their interaction with trial type.

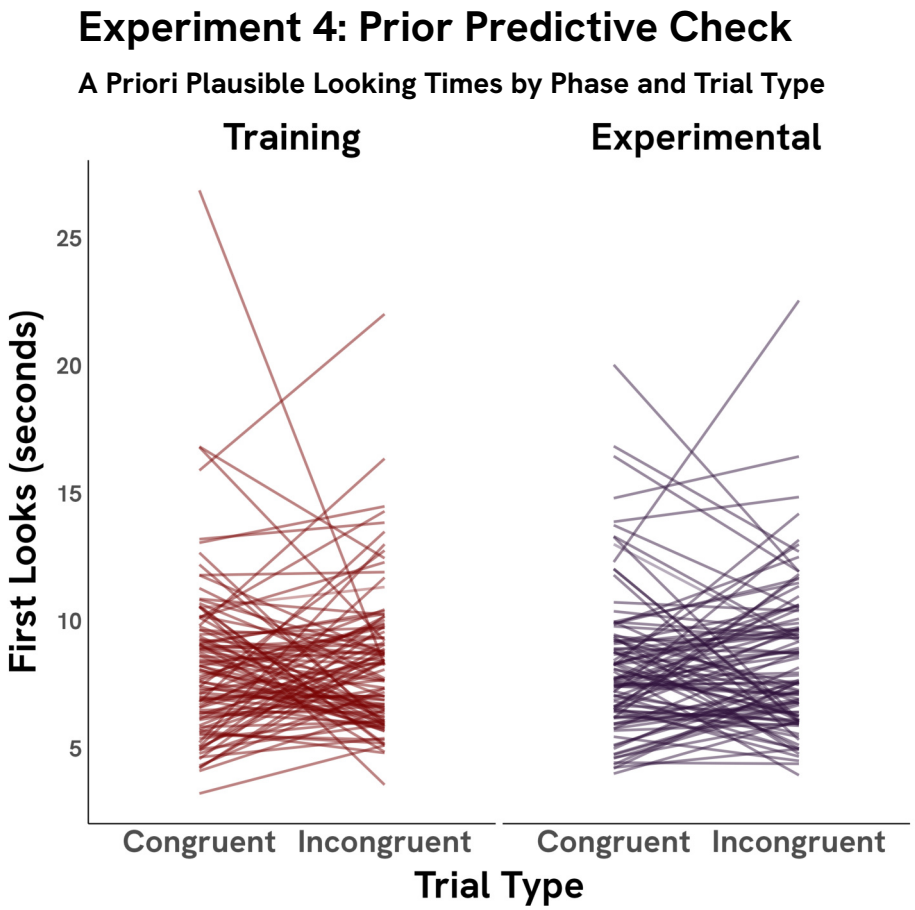

Figure A2. Average looking times simulated from the model priors by phase and trial type (*n* = 100 samples).

The full specification of the model is as follows:

Likelihood function	standardized $\log_{10}(\text{first looks})_i \sim \mathcal{N}(\mu_i, \sigma)$	
Link function	$\mu_i = \beta_{\text{ID}} + \beta_{\text{trialType}} + \beta_{\text{pair}} + \beta_{\text{order}}$	
Priors	$\beta_{\text{ID}} \sim \mathcal{N}(0, \sigma_{\text{ID}})$	$\beta_{\text{condition}} \sim \mathcal{N}(0, 0.35)$
	$\beta_{\text{pair}} \sim \mathcal{N}(0, 0.35)$	$\beta_{\text{order}} \sim \mathcal{N}(0, 0.35)$
Hyperpriors	$\sigma \sim \text{Exponential}(5)$	$\sigma_{\text{ID}} \sim \text{Exponential}(5)$

The additional model controlling for looking times during Training included the parameter $\beta_{\text{training}} \sim \mathcal{N}(0, 0.5)$. For each trial in the Experimental phase, the parameter multiplied the first-looks measurement obtained for the same pair at Training: $\mu_i = \beta_{\text{ID}} + \ldots + \beta_{\text{order}} + \beta_{\text{training}} \cdot \text{first-looks}_{\text{training}}$.

Figure A3 plots the posterior distribution of effects of trial type in the Experimental phase on looking times for the two models. In both models, the 89% HDI excludes the null value.

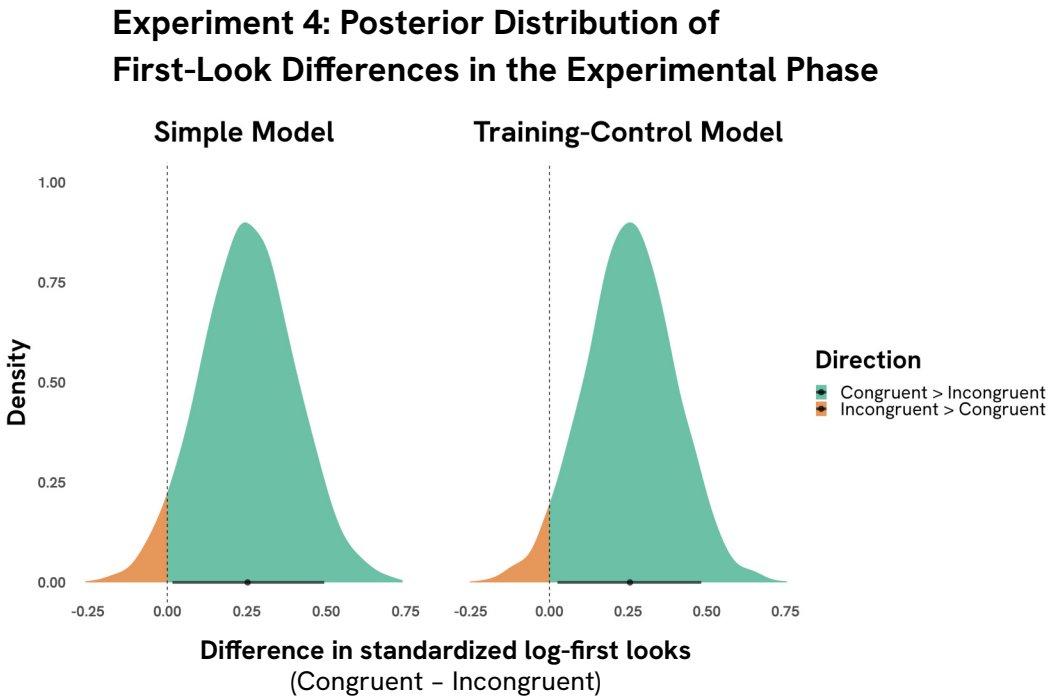

Figure A3. Posterior distribution of looking time differences between the two trial types in the Experimental phase. Black circles represent the mean of the posterior distribution; black horizontal lines give the 89% credible interval around the mean. Dashed vertical lines mark the null value. Left: the output of the simple model. Right: the output of the model controlling for looking times during Training.

A5. The Effect of Age in Experiments 1–4

In Experiments 1–3, I tested a wider age range than is usually tested in infant studies (14 months 0 days; 16 months 30 days). Figure A4 plots the effects of condition in each of Experiments 1–4 by age (Same Word vs. Different Word in Experiments 1–3, Congruent vs. Incongruent in Experiment 4). The lack of any significant influence of age on effects (all *p*s > .142) suggests that the cognitive capacity tapped into by Experiments 1–4 does not undergo any important development within the three months in the age range tested here.

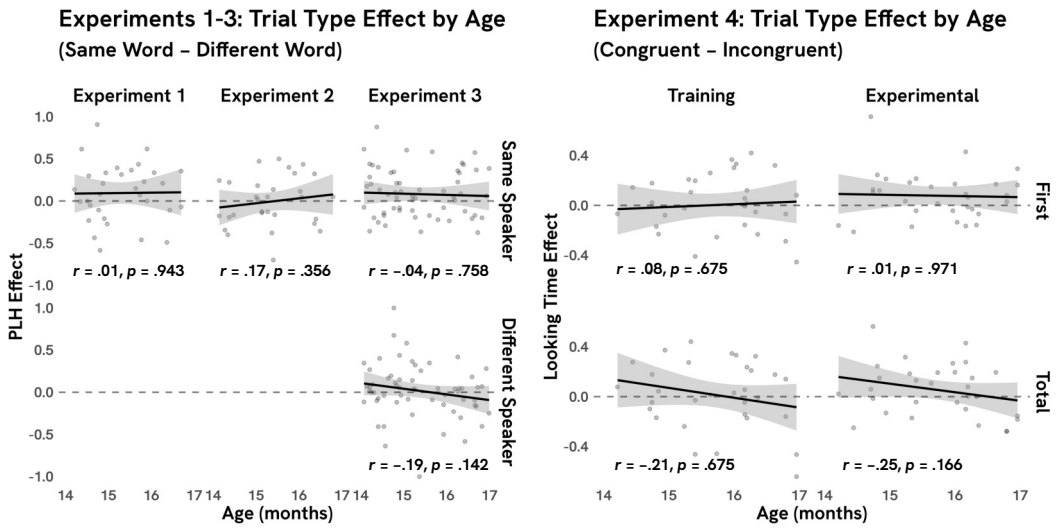

Figure A4. Trial type effects regressed against age in Experiments 1–4. Dots represent individual subject effects; the gray-shaded area represents the 95% confidence interval around the regression line.

Appendix B. Supplemental Materials to Chapter 4

B1. Trial Type × Task Interaction

REACTION TIMES

An overview of Stroop effects by task and experiment is displayed in Figure B1. In all three experiments, the repeated-measures ANOVA reveals a main effect of trial type (Experiment 1: $F(1, 49) = 28.69$, $p < .001$, Experiment 2: $F(1, 49) = 6.18$, $p = .016$; Experiment 3: $F(1, 49) = 60.90$, $p < .001$), a main effect of task (Experiment 1: $F(1, 49) = 26.82$, $p < .001$, Experiment 2: $F(1, 49) = 6.99$, $p = .011$; Experiment 3: $F(1, 49) = 23.99$, $p < .001$), and, except for Experiment 2, a Trial Type × Task interaction (Experiment 1: $F(1, 49) = 29.12$, $p < .001$, Experiment 2: $F(1, 49) = 3.65$, $p = .062$; Experiment 3: $F(1, 49) = 16.33$, $p = .002$). The main effect of task arises because, on average, participants are quicker to solve the Larger task than the Smaller task in all three experiments. The interaction occurs because the Stroop effect is stronger in the Smaller task in all three experiments.

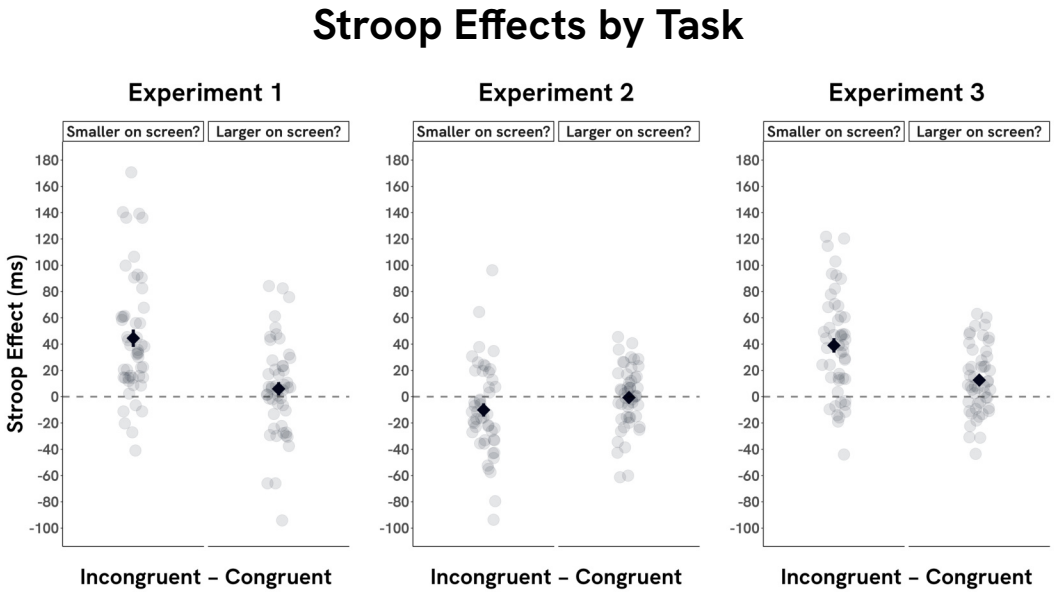

Figure B1. Stroop Effects in Experiments 1–3 by task. Transparent circles depict individual Stroop effects (Incongruent – Congruent reaction times); black diamonds show group average Stroop effect ±1 SEM.

In Experiment 1, if reaction times are separated by task, the Stroop effect turns out to have been driven mainly by the Smaller task, $t(49) = 6.72$, $p < .001$, Cohen's $d = 0.95$, 95% CI [0.62, 1.29], as there was little difference between Incongruent and Congruent trials in the Larger task, $t(49) = 1.17$, $p = .246$, Cohen's $d = 0.17$, 95% CI [−0.12, 0.45]. This partly replicates the original finding, as Konkle and Oliva (2012) also found a stronger effect in the Smaller task.

In Experiment 2, the effect was also primarily driven by the Smaller task, $t(49) = -2.52$, $p = .015$, Cohen's $d = 0.36$, 95% CI [0.07, 0.65]; in the Larger task, trial type had no effect on participants' reaction times, $t(49) = -0.19$, $p = .851$, Cohen's $d = 0.03$, 95% CI [−0.25, 0.31].

In Experiment 3, there was a larger Incongruent–Congruent difference in the Smaller task, $t(50) = 7.09$, $p < .001$, Cohen's $d = 0.96$, 95% CI [0.67, 1.35]. However, unlike in Experiments 1 and 2, there was also a Stroop effect in the Larger block, $t(50) = 3.51$, $p < .001$, Cohen's $d = 0.50$, 95% CI [0.20, 0.80].

ERROR RATES

An overview of participants' error rates by task and trial type in all three experiments is displayed in **Figure B2**. In Experiment 1, a repeated-measures ANOVA reveals a main effect of trial type, $F(1, 49) = 28.30$, $p < .001$, no main effect of task, $F(1, 49) = 2.05$, $p = .158$, and a Trial Type × Task interaction, $F(1, 49) = 6.46$, $p = .014$. Participants made many more errors on Incongruent trials in the Smaller task, $t(49) = 5.48$, $p < .001$, while this difference was less pronounced in the Larger task, $t(49) = 2.11$, $p = .040$.

In Experiment 2, a repeated-measures ANOVA reveals a main effect of trial type, $F(1, 49) = 17.93$, $p < .001$, a main effect of task, $F(1, 49) = 18.08$, $p < .001$, and a Trial Type × Task interaction, $F(1, 49) = 14.03$, $p < .001$. When the error rates are split by task, the same pattern is observed on the Smaller task as in Experiment 1, except in the opposite direction, $t(49) = -5.22$, $p < .001$, and no significant effect in the Larger task even though the difference was in the same direction as that of the Smaller task, $t(49) = -1.22$, $p = .229$. In Experiment 3, a repeated-measures ANOVA indicates a main effect of trial type, $F(1, 49) = 13.92$, $p < .001$, a main effect of task, $F(1, 49) = 5.83$, $p = .02$, and no Trial Type × Task interaction, $F(1, 49) = .01$, $p = .922$.

Splitting the error rates by task reverses the pattern of Experiments 1 and 2 reverses. The Incongruent–Congruent difference in error rates is narrower on the Smaller task, $t(49) = 2.06$, $p = .045$, than on the Larger task, $t(49) = -3.58$, $p = .001$.

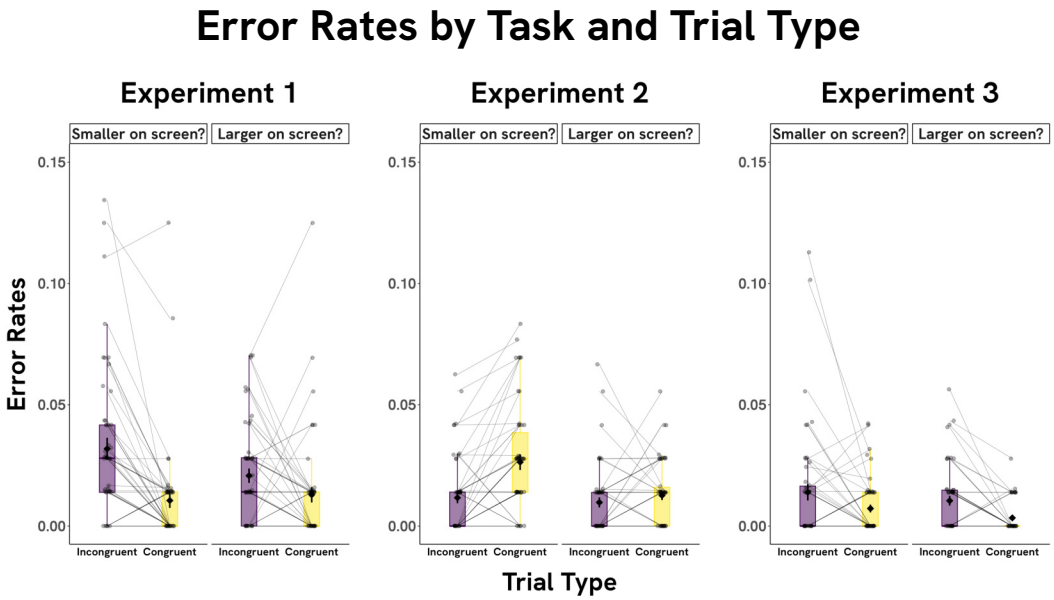

Figure B2. Error rates in Experiments 1–3 by task and trial type. Transparent circles and the lines connecting them represent individual error rates as a function of trial type; opaque diamonds depict group averages ±1 SEM; boxplots indicate the median and interquartile range.

B2. Trial Type × Animacy Interaction

In each of the three experiments, 15 of the 36 pairs contained one image of an animate entity (e.g., camel–monitor; toy zebra–watermelon—in Experiment 2, I coded toy animals as animate). Testing for a Trial Type × Animacy interaction reveals that animacy does not significantly interact with trial type in terms of reaction times in any of the three experiments (Experiment 1: $F(1, 49) = 1.81$, $p = .185$; Experiment 2: $F(1, 49) = .41$, $p = .527$; Experiment 3: $F(1, 49) = 1.02$, $p = .318$).

B3. Item Effects

In each experiment, for each pair, and for each participant, I obtained a Stroop effect by averaging over task (Smaller vs. Larger) and side of presentation (Left vs. Right). **Figure B3** plots the pair Stroop effects by experiment, in decreasing order.

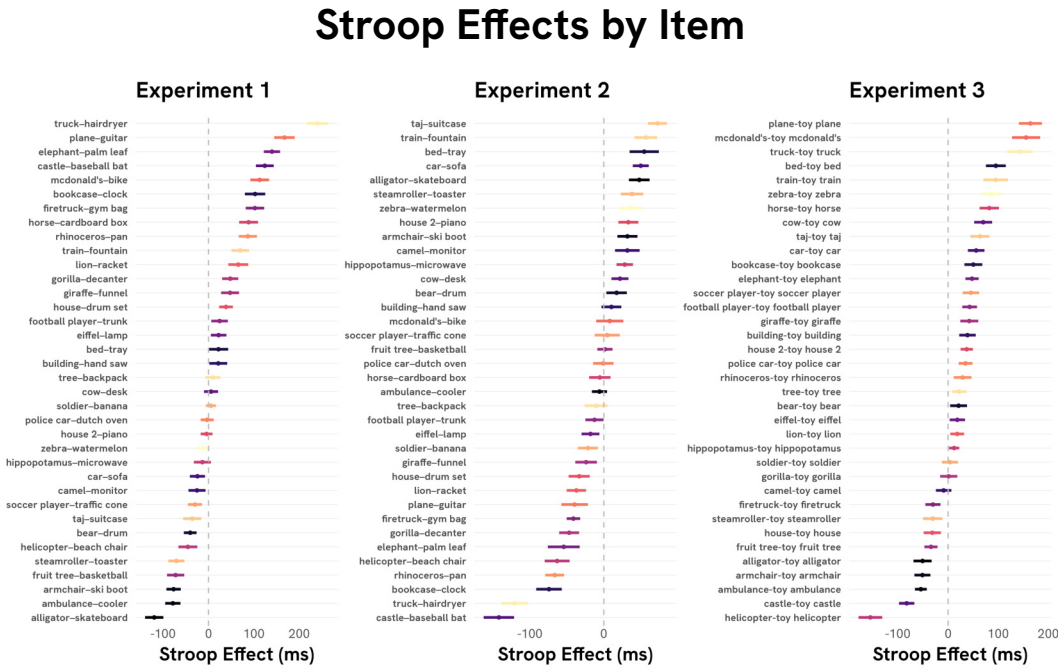

Figure B3. Item effects in Experiments 1–3 sorted by Stroop effect in descending order. Circles represent averages over participants; lines represent ±1 SEM.

B4. Relation Between Size Disparity and Stroop Effect by Item Pair

Size disparities in the depicted objects might also have contributed to the Stroop effect found in Experiments 1–3. I obtained the size of the real-world referent depicted in all 108 images based on the procedure in Konkle and Oliva (2011), either from https://konklab.fas.harvard.edu—for the stimuli which were used in both Konkle and Oliva (2011) and in the current study—or by searching the internet for the typical dimensions of the depicted object. Following Konkle and Oliva (2011), I took the logarithm of the diagonal of the objects' bounding rectangle, ignoring depth. For each trial, I obtained the real-world size difference by subtracting the logarithm of the large image referent size from the logarithm of

the small image referent size. In Experiment 2, I used the real-world size of the small toy and the real-world size of its large referent in separate analyses. Figure B4 depicts the Stroop effects, averaged by pair, against the log-size difference between each pair's large and small objects.

There was no effect of the magnitude of size difference on Stroop effects in Experiments 1 and 3 (*p*s > .49), suggesting that, in general, the Stroop delay was not affected by the magnitude of the perceived incongruency (Figure B4, top row). In Experiment 2, there was a positive correlation between the difference in log-size and the Stroop effect (p = .004) when the size disparity was measured against the toy size. This indicates that the higher the difference between a toy and a mid-sized object was, the more likely it was that participants perceived the trials in which the toy was depicted larger as incongruent. Note, however, that the regression line for this experiment is not analogous to the ones for Experiments 1 and 3. There, the fitted lines are projected toward 0 when the log-size goes to 0, as expected (indicating that the Stroop effect disappears when there is no difference between the actual sizes of the objects).

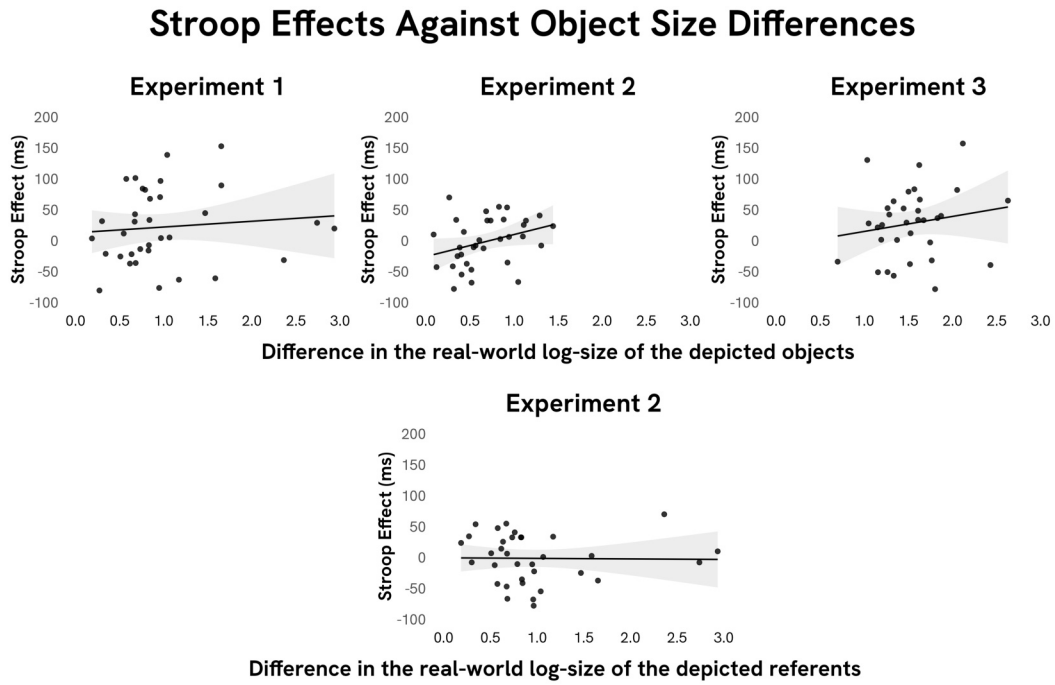

Figure B4. Top: Stroop effect (Incongruent – Congruent reaction times) by size disparity in Experiments 1–3. Bottom: Stroop effect by the size disparity of the toys' referents in Experiment 2.

In Experiment 2, when the correlation is computed with the size measurement of the toys (left column), the regression line predicts a negative Stroop effect when there is no difference between the two objects, which makes no sense. However, if participants interpreted toy images as standing for the referents of the toys, Incongruent trials (in which the toy was displayed larger than the actually larger object) would be easier and thus more congruent for participants. Indeed, when the regression is computed with the size measurements of the toys' referents rather than with the size measurements of the toys (**Figure B4**, bottom row), the regression line looks identical to those of Experiments 1 and 3: a nonsignificant correlation (p = .587), as well as the correct prediction that there will be no Stroop effect when the difference approaches 0.

B5. Pixel Area Differences Control

If larger objects filled more of their bounding box than smaller objects, this could have introduced a bias in the task. In that case, Congruent trials would become easier. The larger object would appear even larger because it would fill more of its bounding box, and the visual judgment could be made faster. On Incongruent trials, the larger object would be depicted at a small size, but it would appear larger because, again, it would fill more of its bounding box, leading to a more difficult judgment. This would result in a Stroop effect even if participants do not compute object sizes at all.

To address this potential confound, I obtained the ratio of nonwhite to white pixels for each of the 108 images in the stimuli set (36 triplets of three images each). Then, for each pair, I obtained the difference between the white-to-nonwhite pixel ratio of both images (large object – small object). The means of the pixel area differences between the paired objects in the three experiments were close to 0 and to each other (Experiment 1: –0.04; Experiment 2: 0.02; Experiment 3: –0.02). Statistical tests confirmed that the null hypotheses for the means being equal to 0 (all p-values > .16) and for the means being equal to each other (p = .309) cannot be rejected. This implies that the pixel area differences may have increased the noise in the measurement but did not bias it.

To confirm this, I recoded pixel difference to reflect the difference between the pixel area filled by the larger image and the pixel area filled by the smaller image on a trial-by-trial basis. I averaged reaction times over participants, item pairs, and trial type, then scaled reaction times and pixel area differ-

ences by dividing them by their standard deviations. I fitted a linear mixed model for each experiment, with trial type and pixel area differences as fixed effects and participant ID and item pair as random intercepts (**Table B1**).

EXPERIMENT	TERM	ESTIMATE	STANDARD ERROR	*P*-VALUE
Experiment 1	Trial type	0.20	0.02	< .001
Experiment 1	Pixel difference	-0.15	0.01	< .001
Experiment 2	Trial type	−0.08	0.02	< .001
Experiment 2	Pixel difference	-0.12	0.01	< .001
Experiment 3	Trial type	0.24	0.02	< .001
Experiment 3	Pixel difference	-0.15	0.01	< .001

Table B1. Subset of the output of the linear mixed model (by experiment): reactionTime ~ trialType + pixelDiff + (1|stimuliPair) + (1|participantID).

Pixel area differences contributed significantly to the Stroop effect in all three experiments—all pixel-difference coefficients are significantly below 0. This is as expected. When the larger image fills more of its bounding box than the smaller image, size judgments are easier and reaction times shorter. However, trial type continues to be a significant predictor after controlling for pixel area differences, indicating that this low-level explanation cannot fully account for the results.