

¹ Action Anticipation Based on an Agent's Epistemic State in Toddlers and Adults

² The ManyBabies² Consortium¹

³ ¹ See author note

⁴ Author Note

5 The ManyBabies2 Consortium consists of Tobias Schuwerk
6 (Ludwig-Maximilians-Universität München), Dora Kampis* (University of Copenhagen),
7 Renée Baillargeon (University of Illinois at Urbana-Champaign), Szilvia Biro (Leiden
8 University), Manuel Bohn (Max Planck Institute for Evolutionary Anthropology), Krista
9 Byers-Heinlein (Concordia University), Sebastian Dörrenberg (University of Bremen),
10 Cynthia Fisher (University of Illinois at Urbana-Champaign), Laura Franchin (University
11 of Trento), Tess Fulcher (University of Chicago), Isa Garbisch (University of Göttingen),
12 Alessandra Geraci (University of Trento), Charlotte Grosse Wiesmann (Max Planck
13 Institute for Human Cognitive and Brain Sciences), J. Kiley Hamlin (University of British
14 Columbia), Daniel Haun (Max Planck Institute for Evolutionary Anthropology) Robert
15 Hepach (University of Oxford), Sabine Hunnius (Radboud University Nijmegen), Daniel C.
16 Hyde (University of Illinois at Urbana-Champaign), Petra Kármán (Central European
17 University), Heather L Kosakowski (MIT), Ágnes M. Kovács (Central European
18 University), Anna Krämer (University of Salzburg), Louisa Kulke
19 (Friedrich-Alexander-University Erlangen-Nürnberg), Crystal Lee (Princeton University),
20 Casey Lew-Williams (Princeton University), Ulf Liszkowski (Universität Hamburg), Kyle
21 Mahowald (University of California, Santa Barbara), Olivier Mascaro (Integrative
22 Neuroscience and Cognition Center, CNRS UMR8002/University of Paris), Marlene Meyer
23 (Radboud University Nijmegen), David Moreau (University of Auckland), Josef Perner
24 (University of Salzburg), Diane Poulin-Dubois (Concordia University), Lindsey J. Powell
25 (University of California, San Diego), Julia Prein (Max Planck Institute for Evolutionary
26 Anthropology), Beate Prielwasser (University of Salzburg), Marina Proft (Universität
27 Göttingen), Gal Raz (MIT), Peter Reschke (Brigham Young University), Josephine Ross
28 (University of Dundee), Katrin Rothmaler (Max Planck Institute for Human Cognitive and
29 Brain Sciences), Rebecca Saxe (MIT), Dana Schneider (Friedrich-Schiller-University Jena,
30 Germany), Victoria Southgate (University of Copenhagen), Luca Surian (University of

³¹ Trento), Anna-Lena Tebbe (Max Planck Institute for Human Cognitive and Brain
³² Sciences), Birgit Träuble (Universität zu Köln), Angeline Sin Mei Tsui (Stanford
³³ University), Annie E. Wertz (Max Planck Institute for Human Development), Amanda
³⁴ Woodward (University of Chicago), Francis Yuen (University of British Columbia),
³⁵ Amanda Rose Yuile (University of Illinois at Urbana-Champaign), Luise Zellner
³⁶ (University of Salzburg), Lucie Zimmer (Ludwig-Maximilians-Universität München),
³⁷ Michael C. Frank (Stanford University), and Hannes Rakoczy (University of Göttingen).

³⁸ Correspondence concerning this article should be addressed to The ManyBabies2
³⁹ Consortium, Leopoldstr. 13, 80802 München, Germany. E-mail:
⁴⁰ tobias.schuwerk@psy.lmu.de

41

Abstract

42 Do toddlers and adults engage in spontaneous Theory of Mind (ToM)? Evidence from
43 anticipatory looking (AL) studies suggests they do. But a growing body of failed
44 replication studies raised questions about the paradigm's suitability, urging the need to test
45 the robustness of AL as a spontaneous measure of ToM. In a multi-lab collaboration we
46 examine whether 18- to 27-month-olds' and adults' anticipatory looks distinguish between
47 two basic forms of epistemic states: knowledge and ignorance. In toddlers [ANTICIPATED
48 n = 520 50% FEMALE] and adults [ANTICIPATED n = 408, 50% FEMALE], we found
49 [SUPPORT/NO SUPPORT] for epistemic state-based action anticipation. Future research
50 can probe whether this conclusion extends to more complex kinds of epistemic states, such
51 as true and false beliefs.

52 *Keywords:* anticipatory looking; spontaneous Theory of Mind; replication

53 Word count: 10243

54 Action Anticipation Based on an Agent's Epistemic State in Toddlers and Adults

55 The capacity to represent epistemic states, known as Theory of Mind (ToM) or
56 mentalizing, plays a central role in human cognition (Dennett, 1989; Frith & Frith, 2006;
57 Premack & Woodruff, 1978). Although ToM has been under intense scrutiny in the past
58 decades, its nature and ontogeny are still the subjects of much controversy. At the heart of
59 these debates are questions about the reliability of the tools used to measure ToM
60 (Baillargeon, Buttelmann, & Southgate, 2018; e.g., Poulin-Dubois et al., 2018), among
61 others, anticipatory looking (AL) paradigms. To address this issue, in a collaborative
62 long-term project we assess the robustness of infants' and adults' tendency to
63 spontaneously take into account different kinds of epistemic states — what they perceive,
64 know, think, or believe — when predicting others' behaviors. This paper reports the first
65 foundational step of this project, which focuses on the most basic epistemic state
66 ascription: the capacity to distinguish between knowledgeable and ignorant individuals.
67 Simple forms of knowledge attribution (such as tracking what other individuals have seen
68 or experienced) are typically assumed to develop early and to operate spontaneously
69 throughout the lifespan (Liszkowski, Carpenter, & Tomasello, 2007; e.g., Luo &
70 Baillargeon, 2007; O'Neill, 1996; Phillips et al., 2021). Thus, evaluating whether ToM
71 measures are sensitive to the knowledge-ignorance distinction is a crucial test case to assess
72 their robustness. The present paper investigates this question in an AL paradigm including
73 18-27-month-old infants and adults.

74 In the following sections we first establish the background and scientific context of
75 this study, namely the reliability and replicability of spontaneous ToM measures. We then
76 introduce a novel way to approach these issues: a large-scale collaborative project targeting
77 the replicability of ToM findings. Finally, we outline the rationale of the present study
78 which uses an AL paradigm to test whether infants and adults distinguish between two
79 basic forms of an agent's epistemic state: knowledge and ignorance.

80 Spontaneous Theory of Mind tasks

81 Humans are proficient at interpreting and predicting others' intentional actions.

82 Adults as well as infants expect agents to act persistently towards the goal they pursue

83 Woodward & Sommerville (2000), and anticipate others' actions based on their goals even

84 before goals are achieved - that is, humans engage in goal-based action anticipation (for

85 review, see Elsner & Adam, 2021; but see Ganglmayer, Attig, Daum, & Paulus, 2019). To

86 predict others' actions, however, it is essential to consider their epistemic state: what they

87 perceive, know, or believe. A number of seminal studies using non-verbal spontaneous

88 measures have suggested that infants, toddlers, older children, and adults show action

89 anticipation and action understanding not only based on other agents' goals (what they

90 want) but also on the basis of their epistemic status (what they perceive, know, or believe).

91 These studies suggest that from infancy onwards, humans spontaneously engage in ToM or

92 mentalizing. For example, studies using violation of expectation methods have

93 demonstrated that infants look longer in response to events in which an agent acts in ways

94 that are incompatible with their (true or false) beliefs, compared to events in which they

95 act in belief-congruent ways (Onishi & Baillargeon, 2005; Surian, Caldi, & Sperber, 2007;

96 Träuble, Marinović, & Pauen, 2010). Other studies have employed more interactive tasks

97 requiring the child to play, communicate, or cooperate with experimenters and, for

98 example, give an experimenter one of several objects as a function of their epistemic status.

99 Such studies have shown that toddlers spontaneously adjust their behavior to the

100 experimenter's beliefs (Buttelmann, Carpenter, & Tomasello, 2009; Király, Oláh, Csibra, &

101 Kovács, 2018; Knudsen & Liszkowski, 2012; Southgate, Johnson, Karoui, & Csibra, 2010).

102 The largest body of evidence for spontaneous ToM comes from studies using AL

103 tasks. In such tasks, participants see an agent who acts in pursuit of some goal (typically,

104 to collect a certain object) and has either a true or a false belief (for example, regarding

105 the location of the target object). A number of studies have shown that infants, toddlers,

106 older children, neurotypical adults, and even non-human primates anticipate (indicated by
107 looks to the location in question) that an agent will go where it (truly or falsely) believes
108 the object to be rather than, irrespective of the actual location of the object (Gliga, Jones,
109 Bedford, Charman, & Johnson, 2014; Grosse Wiesmann, Friederici, Singer, & Steinbeis,
110 2017; Hayashi et al., 2020; Kano, Krupenye, Hirata, Tomonaga, & Call, 2019; Krupenye,
111 Kano, Hirata, Call, & Tomasello, 2016; Meristo et al., 2012; Schneider, Lam, Bayliss, &
112 Dux, 2012; Schneider, Slaughter, Bayliss, & Dux, 2013; Senju et al., 2010; Senju,
113 Southgate, Snape, Leonard, & Csibra, 2011; Senju, Southgate, White, & Frith, 2009;
114 Surian & Franchin, 2020; Thoermer, Sodian, Vuori, Perst, & Kristen, 2012). These studies
115 have revealed converging evidence for spontaneous ToM across the human lifespan and
116 even in other primate species.

117 Across the different measures, the majority of early works on spontaneous ToM in
118 infants and toddlers have reported positive results in the second year of life, and a few
119 studies even within the first year (Kovács, Téglás, & Endress, 2010; Luo & Baillargeon,
120 2010; Southgate & Vernetti, 2014), yielding a rich body of coherent and convergent
121 evidence (for reviews see e.g., Barone, Corradi, & Gomila, 2019; Kampis & Southgate,
122 2020; Scott & Baillargeon, 2017). This growing body of literature has led to a theoretical
123 transformation of the field. In particular, findings with young infants have paved the way
124 for novel accounts of the development and cognitive foundations of ToM. The previous
125 consensus was that full-fledged ToM emerges only at around age 4, potentially as the result
126 of developing executive functions, complex language skills and other factors (e.g., Perner,
127 1991; Wellman & Cross, 2001). In contrast, the newer accounts proposed that some basic
128 forms of ToM may be phylogenetically more ancient and may develop much earlier in
129 ontogeny (e.g., Baillargeon, Scott, & He, 2010; Carruthers, 2013; Kovács, 2016; Leslie,
130 2005).

131 Recently, however, a number of studies have raised uncertainty regarding the
132 empirical foundations of the early-emergence theories, as we review below. In the following

¹³³ sections, we present an overview of the current empirical picture of early understanding of
¹³⁴ epistemic states and then introduce ManyBabies2 (MB2), a large-scale collaborative
¹³⁵ project exploring the replicability of ToM in infancy, of which the current study constitutes
¹³⁶ the first step.

¹³⁷ Replicability of Spontaneous Theory of Mind Tasks

¹³⁸ A number of failures to replicate findings from spontaneous ToM tasks have recently
¹³⁹ been published with infants, toddlers, and adults Kulke & Rakoczy (2019). Besides
¹⁴⁰ conceptual replications, many of these studies involve more direct replication attempts
¹⁴¹ with the original stimuli and procedures. One of these was a two-lab replication attempt of
¹⁴² one of the most influential AL studies (Southgate, Senju, & Csibra, 2007). This failure to
¹⁴³ replicate is especially notable not only because of the influence of the original finding of the
¹⁴⁴ field, but also because of the large sample size and the involvement of some of the original
¹⁴⁵ authors (Kampis et al., 2021). Additional unpublished replication failures have also been
¹⁴⁶ reported. Kulke and Rakoczy (2018) examined 65 published and non-published studies
¹⁴⁷ including 36 AL studies [replications of Schneider, Bayliss, Becker, and Dux (2012);
¹⁴⁸ Southgate et al. (2007); Surian & Geraci, 2012; and Low & Watts, 2013], as well as studies
¹⁴⁹ using other paradigms, and classified them as a successful, partial, or non-replication,
¹⁵⁰ depending on whether all, some, or none of the original main effects were found. Although
¹⁵¹ no formal analysis of effect size was carried out, overall, non-replications and partial
¹⁵² replications outnumbered successful replications, regardless of the method used. In
¹⁵³ addition to the failure to replicate spontaneous anticipation of agents' behaviors based on
¹⁵⁴ their beliefs, many of the replication studies revealed an even more fundamental problem of
¹⁵⁵ spontaneous AL procedures: a failure to adequately anticipate an agent's action in the
¹⁵⁶ absence of a belief. That is, researchers did not find evidence for spontaneous anticipation
¹⁵⁷ of agents' behaviors based on their goals, even in the initial familiarization trials of the
¹⁵⁸ experiments, where the agent's beliefs do not play any role yet (e.g., Kampis et al., 2020;

¹⁵⁹ Kulke, Reiß, et al., 2018; Schuwerk et al., 2018). The familiarization trials are designed to
¹⁶⁰ convey the goal of the agent, as well as the general timing and structure of events, to set
¹⁶¹ up participants' expectations in the test trials where the agent's epistemic state is then
¹⁶² manipulated. Typically, the last familiarization trial can also be used to probe participants'
¹⁶³ spontaneous action anticipation; and test trials can only be meaningfully interpreted if
¹⁶⁴ there is evidence of above-chance anticipation in the familiarization trials. In several AL
¹⁶⁵ studies many participants had to be excluded from the main analyses for failing to
¹⁶⁶ demonstrate robust action anticipation during the familiarization trials (e.g., Kampis et al.,
¹⁶⁷ 2020; Kulke, Reiß, et al., 2018; Schuwerk et al., 2018; Southgate et al., 2007). This raises
¹⁶⁸ the possibility that these paradigms may not be suitable for reliably eliciting spontaneous
¹⁶⁹ action prediction in the first place (for discussion see Baillargeon et al., 2018). In sum, in
¹⁷⁰ light of the complex and mixed state of the evidence, it currently remains unclear whether
¹⁷¹ infants, toddlers, and adults engage in spontaneous ToM. This calls for systematic,
¹⁷² large-scale, a priori designed multi-lab study that stringently tests for the robustness,
¹⁷³ reliability, and replicability of spontaneous measures of ToM.

¹⁷⁴ General Rationale of MB2

¹⁷⁵ To this end, ManyBabies 2 (MB2) was established as an international consortium
¹⁷⁶ dedicated to investigating infants' and toddlers' ToM skills. The main aim is to test the
¹⁷⁷ replicability and thus reliability of findings from spontaneous ToM tasks. In the long-term,
¹⁷⁸ MB2 will build on the initial findings and the aim will be extended to include testing the
¹⁷⁹ validity of these experimental designs and addressing theoretical accounts of spontaneous
¹⁸⁰ ToM. MB2 operates under the general umbrella of ManyBabies (MB), a large-scale
¹⁸¹ international research consortium founded with the aim of probing the reliability of central
¹⁸² findings from infancy research. In particular, MB projects bring together large and
¹⁸³ theoretically diverse groups of researchers to tackle pressing questions of infant cognitive
¹⁸⁴ development, by collaboratively designing and implementing methodologies and

185 pre-registered analysis plans (Frank et al., 2017). The MB2 consortium involves authors of
186 original studies as well as authors of both successful and failed replication studies, and
187 researchers from very different theoretical backgrounds. It thus presents a case of true
188 “adversarial collaboration” (Mellers et al., 2001).

189 **Rationale of the Present Study**

190 Based on both theoretical and practical considerations, the current paper presents
191 the first foundational step in MB2, focusing on AL measures. It investigates whether
192 toddlers and adults anticipate (in their looking behavior) how other agents will act based
193 on their goals (i.e., what they want) and epistemic status (i.e., what they know or do not
194 know). From a practical perspective, we focus on AL since it is a child-friendly and widely
195 used method that is also suitable for humans across the lifespan and even other species.
196 Additionally, as AL is screen-based and standardizable, identical stimuli can be presented
197 in different labs. From a theoretical perspective, given the mixed findings with AL tasks
198 reviewed in the previous section, we take a systematic and bottom-up approach. First, we
199 probe whether AL measures are suitable for measuring spontaneous goal-directed action
200 anticipation. With the aim to improve the low overall rates of anticipatory looks in recent
201 studies, we designed new, engaging stimuli to test whether these are successful in eliciting
202 spontaneous action anticipation. Second, in case reliably elicited action anticipation can be
203 found: we probe whether toddlers and adults take into account the agent’s epistemic status
204 in their spontaneous goal-based action anticipation. That is, do they track whether the
205 agent saw or did not see a crucial event, and therefore whether this agent does or does not
206 know something? In the current study we focus on the most basic form of tracking the
207 epistemic status of agents: considering whether they had access to relevant information,
208 and whether they are thus *knowledgeable* or *ignorant*. We reasoned that only after
209 establishing whether a context can elicit spontaneous tracking of an agent’s epistemic
210 status in a more basic sense (i.e., the agent’s knowledge vs. ignorance) is it eventually

meaningful to ask whether this context also elicits more complex epistemic state tracking (i.e., the agent's beliefs). Answering these first two questions in the present study will allow us, in the long run, to address a third set of questions in subsequent studies, probing the nature of the representations and cognitive mechanisms involved in infant ToM. Do toddlers and adults engage in full-fledged belief-ascription in their spontaneous goal-based action anticipation? What *kind* of epistemic states do toddlers and adults spontaneously attribute to others in their action anticipation (e.g., Horschler et al., 2020; Phillips et al., 2020)? Do the results that prove replicable really assess ToM, or can they be interpreted in alternative ways such as behavioral rules, associations, or simple perceptual preferences (see, e.g., Heyes, 2014; Perner & Ruffman, 2005)? The present study lays the foundation for investigating these questions. Regarding the knowledge-ignorance distinction, many accounts in developmental and comparative ToM research have argued for the ontogenetic and evolutionary primacy of representing *what* agents witness and represent, relative to more sophisticated ways of representing *how* agents represent (and potentially mis-represent) objects and situations (e.g., Apperly & Butterfill, 2009; Flavell, 1988; Kaminski et al., 2008; Martin & Santos, 2016; Perner, 1991; Phillips et al., 2020). For example, it is often assumed that young children and non-human primates may be capable of so-called "Level I perspective-taking" (understanding *who* sees *what*) but only human children from around age 4 may finally develop capacities for "Level II perspective-taking" (understanding *how* a given situation may appear to different agents; Flavell et al., 1981). Empirically, many studies using verbal and/or interactive measures have indicated that children may engage in knowledge-ignorance and related distinctions before they engage in more complex forms of meta-representation (e.g., Flavell et al., 1981; Hogrefe et al., 1986; Moll & Tomasello, 2006; O'Neill, 1996; though for some findings indicating Level II perspective-taking at an early age see Scott & Baillargeon, 2009; Buttelmann et al., 2015; Buttelmann & Kovács, 2019; Kampis et al., 2020; Scott, Richman, & Baillargeon, 2015), and that non-human primates seem to master knowledge-ignorance tasks while not

238 demonstrating any more complex, meta-representational form of ToM (e.g., Hare et al.,
239 2011; Kaminski et al., 2008; Karg et al., 2015). The knowledge-ignorance distinction thus
240 appears to be an ideal candidate for assessing epistemic status-based action anticipation in
241 a wide range of populations. To date, however, no study has probed whether or how
242 children's (and adults') spontaneous action anticipation, as indicated by AL, is sensitive to
243 ascriptions of knowledge vs. ignorance. Most studies that have addressed ToM with AL
244 measures have targeted the more sophisticated true/false belief contrast. As reviewed
245 above, the results of those studies yield a mixed picture regarding replicability of the
246 findings. It has been argued that tasks that reliably replicate are ones which can be solved
247 with the more basic knowledge-ignorance distinction, whereas tasks that do not replicate
248 require more sophisticated belief-ascertainment (Powell et al., 2018)¹, suggesting that only
249 some but not all findings might not be replicable. Based on these considerations, the
250 present study tests whether toddlers and adults engage in knowledge- and ignorance-based
251 AL to probe the most basic form of spontaneous, epistemic state-based action anticipation.

252 Design and Predictions of the Present Study

253 The current study presents 18- to 27-month-old toddlers and adults with animated
254 scenarios while measuring their gaze behavior. Testing adults (and not just toddlers) is
255 crucial to address debates about the validity and interpretation of AL measures of ToM
256 throughout the lifespan (e.g., Schneider et al., 2017). Following the structure of previous
257 AL paradigms, participants are first familiarized to an agent repeatedly approaching a
258 target (familiarization trials). AL is measured during familiarization trials to probe

¹ For example, some studies have found partial replication results, with patterns of the following kind: participants showed systematic anticipation (or appropriate interactive responses) in true belief trials but showed looking (or interactive responses) at chance level in the false belief trials (e.g., Dörrenberg et al., 2019; Kulke, Reiß, et al., 2018; Powell et al., 2018). Such a pattern remains ambiguous since it may merely reflect a knowledge-ignorance distinction.

259 whether participants understood the agent's goal and spontaneously anticipate their
260 actions. Subsequently, during test trials the agent's visual access is manipulated, leading
261 them to be either *knowledgeable* or *ignorant* about the location of the target. Participants'
262 AL will be measured during test trials to determine whether or not they take into account
263 the agent's epistemic access and adjust their action anticipation accordingly. Participants'
264 looking patterns will be recorded using either lab-based corneal reflection eye-tracking or
265 online recording of gaze patterns. We chose to provide the online testing option to increase
266 the flexibility for data collection given the disruption caused by the Covid-19 pandemic.
267 This option will also provide the opportunity to potentially compare in-lab and online
268 testing procedures (Sheskin et al., 2020). Novel animated stimuli were collectively
269 developed within the MB2 consortium on the basis of previous work (e.g., Clements &
270 Perner, 1994) and based on input from collaborators with experience with both successful
271 and failed replication studies (e.g., Grosse Wiesmann et al., 2017; Surian & Geraci, 2012).
272 These animated 3D scenes feature a dynamic interaction aimed to optimally engage
273 participants' attention: a chasing scenario involving two agents, a *chaser* and a *chasee* (see
274 Figures 1 and 2). As part of the chase, the chasee enters from the top of an upside-down
275 Y-shaped tunnel with two boxes at its exits. The tunnel is opaque so participants cannot
276 see the chasee after it enters the tunnel, but can hear noises that indicate movement. The
277 chasee eventually exits from one of the arms of the Y, and goes into the box on that side.
278 The chaser observes the chasee exit the tunnel and go into a box, and then follows it
279 through the tunnel. During familiarization trials, the chaser always exits the tunnel on the
280 same side as the chasee, and approaches the box where the chasee is currently located.
281 Thus, if participants engage in spontaneous action anticipation during familiarization
282 trials, they should reliably anticipate during the period when the chaser is in the tunnel
283 that it will emerge at the exit that leads to the box containing the chasee. During test
284 trials, the chasee always first hides in one of the boxes but shortly thereafter leaves its
285 initial hiding place and hides in the box at the other tunnel exit. Critically, the chaser

either does (*knowledge* condition) or does not (*ignorance* condition) have epistemic access to the chasee's location. During *knowledge* trials, the chaser observes all movements of the chasee. During *ignorance* trials, the chaser observes the chasee enter the tunnel, but then leaves and only returns once the chasee is already hidden inside the second box. The event sequences in the two conditions are thus identical with the only difference between conditions pertaining to what the chaser has or has not seen. They were designed in this way with the long-term aim to implement, in a minimal contrast design, more complex conditions of false/true belief contrasts with the very same event sequences (true belief conditions will then be identical to the knowledge conditions here, but in false belief conditions the chaser witnesses the chasee's placement in the first box, but then fails to witness the re-location)². Participants' AL (their gaze pattern indicating where they expect the chaser to appear) will be assessed during the anticipatory period - that is, the period during which the chaser is going through the tunnel and is not visible. There will be two main dependent measures: first looks, and a differential looking score (DLS). The first look measure will be binary, indicating which of the two tunnel exits participants fixate first: the exit where the chasee is actually hiding, or the other exit. DLS is a measure of the proportion of time spent looking at the correct tunnel exit during the entire anticipatory period. In two pilot studies (see Methods section), we addressed the foundational question

² There is thus a certain asymmetry with regard to the interpretation and the consequences of potentially positive and negative results of the present knowledge-ignorance contrast: in the case of positive results, we can conclude that subjects spontaneously engage in basic epistemic state ascription and can move on to test, with the minimal contrast comparison of knowledge-ignorance vs. false belief-true belief, whether this extends to more complex forms of epistemic state attribution. In the case of negative results, though, we cannot draw firm conclusions to the effect that subjects do not engage in spontaneous epistemic state ascription. More caution is in order since the present knowledge-ignorance contrast has been designed in order to be comparable to future belief contrasts rather than to be the simplest implementation possible. Simpler implementations would then need to be devised that involve fewer steps (i.e. the chasee just goes to one location and this is or is not witnessed by the chasee).

of the current study: whether these stimuli reveal spontaneous goal-directed action anticipation as measured by AL in the above-described familiarization trials (i.e., without a change of location by the chasee or manipulation of the chaser's epistemic state). We found that our paradigm indeed elicited action anticipation and exclusion rates due to lack of anticipation were significantly lower relative to previous (original and replication) AL studies. Both toddlers and adults showed reliable anticipation of the chaser's exit at the chasee's location, indicating that in contrast with many previous AL studies the current paradigm successfully elicits spontaneous goal-based action anticipation. Based on these pilot data we concluded that the paradigm is suitable for examining the second and critical question: whether toddlers and adults, in their spontaneous goal-based action anticipation, take into account the agent's epistemic state. We predict that if participants track the chaser's perceptual access and resulting epistemic state (knowledge/ignorance) and anticipate their actions accordingly, they should look more in anticipation to the exit at the chasee's location than the other exit in the *knowledge* condition, but should not do so (or to a lesser degree; see below) in the *ignorance* condition. We anticipate three potential factors that could influence participant's gaze patterns: Keeping track of the chaser's epistemic status in the *ignorance* condition might either lead to no expectations as to where the chaser will look (resulting in chance level looking between the two exits) or (if participants follow an "ignorance leads to mistakes"-rule, see e.g., Ruffman, 1996) to an expectation that the chaser will go to the wrong location (longer looking to the exit with the empty box; e.g., Fabricius et al., 2010). Either way, participants may still show a 'pull of the real' even in the *ignorance* condition, i.e., reveal a default tendency to look to the side where the chasee is located. But if they truly keep track of the epistemic status of the chaser (*knowledge* vs. *ignorance*), they should show this tendency to look to the side where the chasee really is in the *ignorance* condition to a lesser degree than in the *knowledge* condition. In sum, the research questions of the present study are the following: First, can we observe in a large sample that toddlers and adults robustly anticipate agents' actions

331 based on their goals in this paradigm, as they did in our pilot study? Second, can we find
332 evidence that they take into account the agent's epistemic access (knowledge vs. ignorance)
333 and adjust their action anticipation accordingly? In addressing these questions, the present
334 study will significantly contribute to our knowledge on spontaneous ToM. It will inform us
335 whether the present paradigm and stimuli can elicit spontaneous goal-based and
336 mental-state-based action anticipation in adults and toddlers, based on a large sample of
337 about 800 participants in total from over 20 labs. In the long run, the present study will
338 lay the foundation for future work to address broader questions of what *kind* of epistemic
339 states toddlers and adults spontaneously attribute to others in their action anticipation
340 and what cognitive mechanisms allow them to do so.

341 Methods

342 All materials, and later the collected de-identified data, will be provided on the Open
343 Science Framework (OSF; <https://osf.io/jmuvd/>). All analysis scripts, including the pilot
344 data analysis and simulations for the design analysis, can be found on GitHub
345 (<https://github.com/manybabies/mb2-analysis>). We report how we determined our
346 sample size and we will report all data exclusions, all manipulations, and all measures in
347 the study. Additional methodological details can be found in the Supplemental Material.

348 Stimuli

349 Figures 1 and 2 provide an overview of the paradigm. For the stimuli, 3D animations
350 were created depicting a chasing scenario between two agents (chaser and chasee) who start
351 in the upper part of the scene. At the very top of the scene a door leads to outside the
352 visible scene. Below this area, a horizontal fence separates the space, and thus the lower
353 part of the space can be reached by the Y-shaped tunnel only. Additional information on
354 the general scene setup, events, and timings in the familiarization and the test trials, as
355 well as trial randomization can be found in the Supplemental Material.

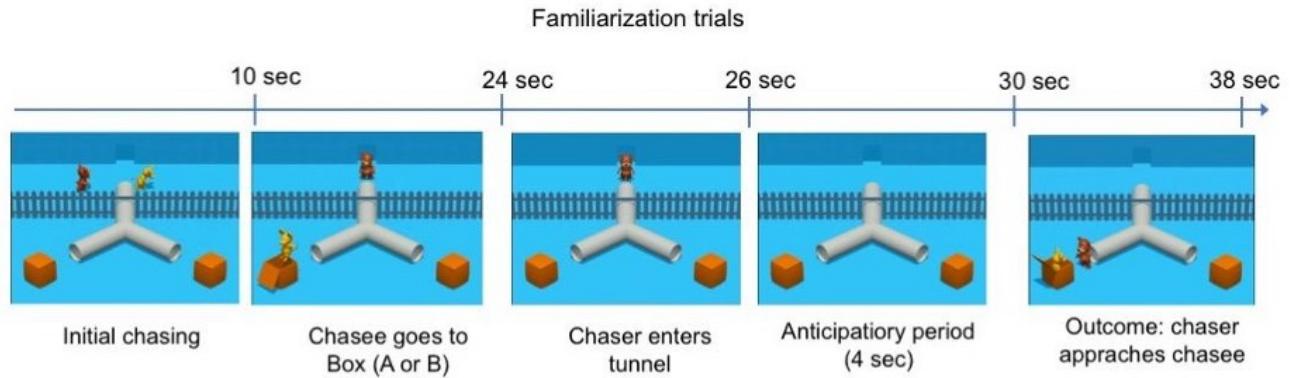


Figure 1. Timeline of the familiarization trials.

356 **Familiarization Trials.** All participants will view four familiarization trials (for an

357 overview of key events see Figure 1). During familiarization trials, after a brief chasing

358 introduction, the chasee enters an upside-down Y-shaped tunnel with a box at both of its

359 exits. The chasee then leaves the tunnel through one of the exits and hides in the box on

360 the corresponding side. Subsequently, the chaser enters the tunnel (to follow the chasee),

361 and participants' AL to the tunnel exits is measured before the chaser exits on the side the

362 chasee is hiding, as an index of their goal-based action anticipation. In these familiarization

363 trials, if participants engage in spontaneous action anticipation, they should reliably

364 anticipate that the chaser should emerge at the tunnel exit that leads to the box where the

365 chasee is. After leaving the tunnel, the chaser approaches the box in which the chasee is

366 hiding and knocks on it. Then, the chasee jumps out of the box and the two briefly interact.

367 **Familiarization Phase Pilot Studies.** In a pilot study with 18- to

368 27-month-olds ($n = 65$) and adults ($n = 42$), seven labs used in-lab corneal reflection

369 eye-tracking to collect data on gaze behavior in the familiarization phase. A key

370 desideratum of our paradigm is that it should produce sufficient AL, as a low rate of AL in

371 previous studies has led to high exclusion rates. The goals of the pilot study were to 1)

372 estimate the level of correct goal-based action predictions in the familiarization phase, 2)

373 determine the optimal number of familiarization trials, 3) check for issues with perceptual

374 properties of stimuli (e.g., distracting visual saliences), and 4) test the general procedure

375 including preprocessing and analyzing raw gaze data from different eye-tracking systems.

376 We found that the familiarization stimuli elicited a relatively high proportion of

377 goal-directed action anticipations, but we were concerned about the effects of some minor

378 properties of the stimulus (in particular, a small rectangular window in the tunnel tube

379 that allowed participants to see the agents at one point on their path to the tunnel exits).

380 In a second pilot study with 18- to 27-month-olds ($n = 12$, three participating labs), slight

381 changes of stimulus features (the removal of the window in the tube; temporal changes of

382 auditory anticipation cue) did not cause major changes in the AL rates. Sixty-eight percent

383 of toddlers' first looks in the first pilot, 69% of toddlers' first looks in the second pilot, and

384 69% of adults' first looks were toward the correct area of interest (AOI) during the

385 anticipatory period. The average proportion of looking towards the correct AOI during the

386 anticipatory period was 70.7% ($CI_{95\%} = 67.6\% - 73.8\%$) in toddlers in the first pilot, 70.5%

387 ($CI_{95\%} = 62.8\% - 78.2\%$) in the second pilot for toddlers, and 75.3% ($CI_{95\%} = 71.0\% -$

388 79.5%) in adults. In Bayesian analyses, we found strong evidence that toddlers and adults

389 looked more towards the target than towards the distractor during the anticipation period.

390 Based on conceptual and practical methodological considerations while also considering

391 previous studies, we decided to include four trials in the final experiment. The pilot data

392 results of the toddlers supported this decision insofar as we observed a looking bias towards

393 the correct location already in trials 1-4, without additional benefit of trials 5-8. Further,

394 prototypical analysis pipelines were established for combining raw gaze data from different

395 eye-trackers. In short, we developed a way to resample gaze data from different

396 eye-trackers to be at a common Hz rate and to define proportionally correct AOIs for

397 different screen dimensions with the goal to merge all raw data into one data set for

398 inferential statistics. The established analysis procedure is described further in the Data

399 Preprocessing section below. In sum, we concluded that this paradigm sufficiently elicits

400 goal-directed action predictions, an important prerequisite for drawing any conclusion on

401 AL behavior in the test trials of this study. A detailed description of the two pilot studies

402 can be found in the Supplemental Material.

403 **Test Trials.** All participants will see two test trials, one *knowledge* and one
404 *ignorance* trial. However, in line with common practice in ToM studies, the main
405 comparison concerns the first test trial between-participants to avoid potential carryover
406 effects. In addition, in exploratory analyses, we plan to assess whether results remain the
407 same if both trials are taken into account and whether gaze patterns differ between the two
408 trials (see Exploratory Analyses). If the results remain largely unchanged across the two
409 trials, it may suggest that future studies could increase power by including multiple test
410 trials. In test trials, the chasee first hides in one of the boxes, but shortly thereafter the
411 chasee leaves this box and hides in the second box, at the other tunnel exit. Critically, the
412 chaser either witnesses (*knowledge* condition) or does not witness (*ignorance* condition)
413 from which tunnel exit the chasee exited and thus where the chasee is currently hiding (for
414 an overview, see Figure 2). In the *knowledge* trials, the chaser observes all movements of
415 the chasee. The chaser leaves for a brief period of time after the chasee entered the tunnel,
416 but it returns before the chasee exits the tunnel. Therefore, no events take place in the
417 chaser's absence. In the *ignorance* trials, the chaser sees the chasee enter the tunnel, but
418 then leaves. Therefore, the chaser does not see the chasee entering either box and only
419 returns once the chasee is already hidden in the final location. Finally, the chaser enters
420 the tunnel but does not appear in either exit. Rather, the scene "freezes" for four seconds
421 and participants' AL is measured. Thus, the *knowledge* and *ignorance* conditions are
422 matched for the chaser leaving for a period of time, but they differ in whether they warrant
423 the chaser's epistemic access to the location of the chasee. No outcome is shown in either
424 test trials. When designing the *knowledge* and *ignorance* condition, we aimed at keeping all
425 events and their timings parallel, except the crucial manipulation. We show the same
426 events in both conditions. Where possible, all events also have the same duration. In the
427 case of the chaser's absence in the *knowledge* condition, there were two main options, both
428 with inevitable trade-offs. First, we could have increased the duration of the chaser's

absence in the *knowledge* condition to match the duration of the chaser's absence in both conditions. Yet, this would potentially disrupt the flow of events, such as keeping track of the chasee's actions and the general scene dynamics, since nothing would happen for a substantial amount of time. Second, the chaser can be absent for a shorter time in the *knowledge* than in the *ignorance* condition, in which case the flow of events – the chasee's actions and the general scene dynamics – remains natural. We chose the second option because we reasoned that the artificial break in the *knowledge* condition could disrupt the participant's tracking of the chaser's epistemic state, thus being a confound that would be more detrimental than the difference in the duration of absence. Further, the current contrast has the advantage that the chasee's sequence and timing of actions are identical in both conditions, thus minimizing the difference between conditions. Finally, with the current design, the duration of the chaser's absence will be closely matched in the later planned false belief - true belief contrast, because in the future false belief condition, the chaser has to be absent for fewer events (because the chaser witnesses the first hiding events after the chasee reappeared at the other side of the tunnel).

Trial Randomization. We will vary the starting location of the chasee (left or right half of the upper part of the scene) and the box the chasee ended up (left or right box) in both familiarization and test trials. The presentation of the familiarization trials will be counterbalanced in two pseudo-randomized orders. Each lab signs up for one or two sets of 16-trial-combinations, for each of their tested age groups.

Lab Participation Details

Time-Frame. The contributing labs will start data collection as soon as they are able to once our Registered Report receives an in-principle acceptance. The study will be submitted for Stage 2 review within one year after in-principle acceptance (i.e., post-Stage 1 review). We anticipate that this time window gives the individual labs enough flexibility to contribute the committed sample sizes; however, if this timeline needs adjusting due to

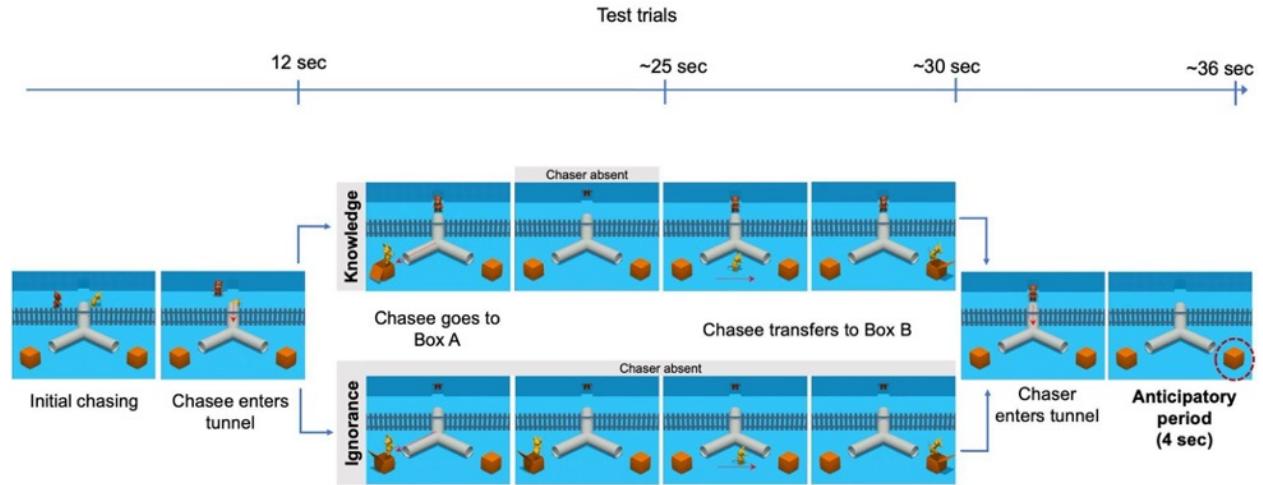


Figure 2. Schematic overview of stimuli and conditions of the test trials.

Note. After the familiarization phase, participants know about the agent's goal (chaser wants to find chasee), perceptual access (chaser can see what happens on the other side of the fence), and situational constraints (boxes can be reached by walking through the forking tunnel). In the *knowledge* condition, the chaser witnesses the chasee walking through the tunnel and jumping in and out of the first box. While the chasee is in the box, the chaser briefly leaves the scene through the door in the back and returns shortly after. Subsequently, the chaser watches the chasee jumping out of the box again and hiding in the second box. In the *ignorance* condition, the chaser turns around and stands on the other side of the door in the back of the scene, thus unable to witness any of the chasee's actions. The chaser then returns and enters the tunnel to look for the chasee. During the test phase (4 seconds still frame), AL towards the end of the tunnels is measured.

455 the Covid-19 pandemic this decision will be made prior to any data analysis.

456 **Participation Criterion.** The participating labs were recruited from the MB2
457 consortium. In July 2020, we asked via the MB2 listserv which labs plan to contribute how
458 many participants for the respective age group (toddlers and/or adults). The Supplemental
459 Material provides an overview of participating labs. Each lab made a commitment to
460 collecting data from at least 16 participants (toddlers or adults), but we will not exclude
461 any contributed data on the basis of the total sample size contributed by that lab. Labs
462 will be allowed to test using either in-lab eye-tracking or online methods.

463 **Ethics.** All labs will be responsible for obtaining ethics approval from their
464 appropriate institutional review board. The labs will contribute de-identified data for
465 central data analysis (i.e., eye-tracking raw data/coded gaze behavior, demographic
466 information). Video recordings of the participants will be stored at each lab according to
467 the approved local data handling protocol. If allowed by the local institutional review
468 board, video recordings will be made available to other researchers via the video library
469 DataBrary (<https://nyu.databrary.org/>).

470 **Participants.** In a preliminary expression of interest, 26 labs signed up to
471 contribute a minimal sample size of 16 toddlers and/or adults. Based on this information,
472 we expect to recruit a total sample of 520 toddlers (ages 18-27 months) and 408 adults
473 (ages 18-55 years). To avoid an unbalanced age distribution in the toddlers sample, labs
474 will sign up for testing at least one of two age bins (bin 1: 18-22 months, bin 2: 23-27
475 months), and will be asked to ensure approximately equal distribution of participants' age
476 in their collected sample if possible. They will be asked to try to ensure that the mean age
477 of their sample lies in the middle of the range of the chosen bin and that participant ages
478 are distributed across their whole bin. Both for adults and toddlers, basic demographic
479 data will be collected on a voluntary basis with a brief questionnaire (see Supplemental
480 Material for details). The requested demographic information that is not used in the
481 registered confirmatory and/or exploratory analyses of this study will be collected for

482 further potential follow-up analyses in spin-off projects within the MB framework. After
483 completing the task, adult participants will be asked to fill a funneled debriefing
484 questionnaire. This questionnaire asks what the participant thinks the purpose of the
485 experiment was, whether the participant had any particular goal or strategy while watching
486 the videos, and whether the participant consciously tracked the chaser's epistemic state.
487 Additionally, we collect details regarding each testing session (see Supplemental Material).

488

489

490 Our final dataset consisted of 1224 participants, with an overall exclusion rate of
491 24.16% (toddlers: 35.60%, adults: 12.67%). Tables 1 A. and B. show the distribution of
492 included participants across labs, eye-tracking methods, and ages. A final sample of 521
493 toddlers (49.14% female) that were tested in 37 labs (mean lab sample size = 14.08, $SD =$
494 5.56, range: 2 - 32) was analyzed. The average age of toddlers in the final sample was 22.49
495 months ($SD: 2.53$, range: 18 - 27.01). The final sample size of included adults was $N = 703$
496 (68.85% female), tested in 34 labs (mean lab sample size = 20.68, $SD = 12.14$, range: 8 -
497 65). Their mean age was 24.61 years ($SD: 7.36$, range: 18 - 55).

498 **Apparatus and Procedure**

499 **Eye-tracking Methods.** We expect that participating labs will use one of three
500 types of eye-tracker brands to track the participant's gaze patterns: Tobii, EyeLink, or
501 SMI. Thus, apparatus setup will slightly vary in individual labs (e.g., different sampling
502 rates and distances at which the participants are seated in front of the monitor).
503 Participating labs will report their eye-tracker specifications and study procedure alongside
504 the collected data. To minimize variation between labs, all labs using the same type of
505 eye-tracker will use the same presentation study file specific to that eye-tracker type. The

506 Supplemental Material will provide an overview of employed eye-trackers, stimulus
507 presentation softwares, sampling rates and screen dimensions.

508 **Online Gaze Recording.** To allow for the participation of labs that do not have
509 access to an eye-tracker, or are not able to invite participants to their facilities due to
510 current restrictions regarding the COVID-19 pandemic, labs can choose to collect data via
511 online testing. Specifically, labs may choose to manually code gaze direction during
512 stimulus presentation on a frame-by-frame basis from video recordings of a camera facing
513 the participant (e.g., a webcam). Labs that choose to collect data virtually will utilize the
514 platform of their choice (e.g., LookIt, YouTube, Zoom, Labvanced, etc.). Further, labs may
515 also choose to use webcam eye-tracking with tools like WebGazer.js (Papoutsaki et al.,
516 2016). In our analyses, we control for and quantify potential sources of variability due to
517 these different methods.

518 **Testing Procedure.** Toddlers will be seated either on their caregiver's lap or in a
519 highchair. The distance from the monitor will depend on the data collection method.
520 Caregivers will be asked to refrain from interacting with their child and close their eyes
521 during stimulus presentation or wear a set of opaque sunglasses. Adult participants will be
522 seated on a chair within the respective appropriate distance from the monitor. Once the
523 participant is seated, the experimenter will initiate the eye-tracker-specific calibration
524 procedure. Additionally, we will present another calibration stimulus before and after the
525 presentation of the task. This allows for evaluating the accuracy of the calibration
526 procedure across labs (cf., Frank et al., 2012).

527 **General Lab Practices**

528 To ensure standardization of procedure, materials for testing practices and
529 instructions will be prepared and distributed to the participating labs. Each lab will be
530 responsible for maintaining these practices and report all relevant details on testing
531 sessions (for details see the Supplemental Material).

532 **Videos of Participants.** As with all MB projects, we strongly encourage labs to
533 record video data of their own lab procedures and each testing session, provided that this is
534 in line with regulations of the respective institutional ethics review board and the given
535 informed consent. Participating labs that cannot contribute participant videos will be
536 asked to provide a video walk-through of their experimental set-up and procedure instead.
537 If no institutional ethics review board restrictions occur, labs are encouraged to share video
538 recordings of the test sessions via DataBrary.

539 Design Analysis

540 Here we provide a simulation of the predicted findings because a traditional
541 frequentist power analysis is not applicable for our project for two reasons. First, we use
542 Bayesian methods to quantify the strength of our evidence for or against our hypotheses,
543 rather than assessing the probability of rejecting the null hypothesis. In particular, we
544 compute a Bayes factor (BF; a likelihood ratio comparing two competing hypotheses),
545 which allows us to compare models. Second, because of the many-labs nature of the study,
546 the sample size will not be determined by power analysis, but by the amount of data that
547 participating labs are able to contribute within the pre-established timeframe. Even if the
548 effect size is much smaller than what we anticipate (e.g., less than Cohen's $d = 0.20$), the
549 results would be informative as our study is expected to be dramatically larger than any
550 previous study in this area. If, due to unforeseen reasons, the participating labs will not be
551 able to collect a minimum number of 300 participants per age group within the proposed
552 time period, we plan to extend the time for data collection until this minimum number is
553 reached. Or in contrast, if the effect size is large (e.g., more than Cohen's $d = 0.80$), the
554 resulting increased precision of our model will allow us to test a number of other
555 theoretically and methodologically important hypotheses (see Results section). Although
556 we did not determine our sample size based on power analysis, here we provide a
557 simulation-based design analysis to demonstrate the range of BFs we might expect to see,

given a plausible range of effect sizes and parameters. We focus this analysis on our key analysis of the test trials (as specified below), namely the difference in AL on the first test trial that participants saw. We describe below the simulation for the child sample, but based on our specifications, we expect that a design analysis for adult data would produce similar results. We first ran a simulation for the first look analysis. In each iteration of our simulation, we used a set of parameters to simulate an experiment, using a first look (described below) as the key measure. For the key effect size parameter for condition (*knowledge* vs. *ignorance*), we sampled a range of effect sizes in logit space spanning from small to large effects (Cohen's $d = 0.20 - 0.80$; log odds from 0.36 - 1.45). For each experiment, the betas for age and the age x condition interaction were sampled uniformly between -0.20 and 0.20. The age of each participant was sampled uniformly between 18 and 27 months and then centered. The intercept was sampled from a normal distribution (1, 0.25), corresponding to an average looking proportion of 0.73. Lab intercepts and the lab slope by condition were set to 0.1, and other lab random effects were set to 0 as we do not expect them to be meaningfully non-zero. These values were chosen based on pilot data (average looking proportion), but also to have a large range of possible outcomes (lab intercept, age and age x condition interaction). We are confident that the results would be robust to different choices. We then used these simulated data to simulate an experiment with 22 labs and 440 toddlers and computed the resulting BFs, as specified in the analysis plan below. We adopted all of the priors specified in the results section below³. We ran 349 simulations and, in 72% of them, the BF showed strong evidence in favor of the full model ($BF > 10$); in 6% the BF showed substantial evidence ($10 > BF > 3$); it was inconclusive 14% of the time ($1/10 > BF > 3$), and in 8% of cases the null model was substantially

³ After the design analysis, additional labs expressed their interest in contributing data, which is why the anticipated sample sizes and the numbers this design analysis is based on differ. Given the uncertainty in determining the final sample size in this project, we kept the design analysis as is to have a more conservative estimate of the study's power.

581 favored (see Figure 3). In none of the simulations the BF was $< 1/10$. Thus, under the
582 parameters chosen here for our simulations, it is likely that the planned experiment is of
583 sufficient size to detect the expected effect. We also ran a design analysis for the
584 proportional looking analysis. We used the same experimental parameters (number of labs,
585 participants, ages, etc.). For generating simulated data, we drew the condition effect from
586 a uniform distribution between .05 and .20 (in proportion space). The age and
587 age:condition effects were drawn from uniform distributions between -.05 and .05. Sigma,
588 the overall noise in the experiment, was drawn from a uniform distribution between .05 and
589 .1. The intercept was drawn from a normal distribution with mean .65 and a standard
590 deviation of .05. The by-lab standard deviation for the intercept and condition slope was
591 set to .01. Priors were as described in the main text. We ran 119 simulations, and in all
592 119 we obtained a BF greater than 10, suggesting that, under our assumptions, the study
593 is well-powered.

594 Data Preprocessing

595 **Eye-tracking.** Raw gaze position data (x- and y-coordinates) will be extracted in
596 the time window starting from the first frame at which the chaser enters the tunnel until
597 the last frame before it exits the tunnel in the last familiarisation trial and in the test trial.
598 For data collected from labs using a binocular eye-tracker, gaze positions of the left and the
599 right eye will be averaged. We will use the peekds R package
600 (<http://github.com/langcog/peekds>) to convert eye-tracking data from disparate trackers
601 into a common format. Because not all eye-trackers record data with the same frequency or
602 regularity, we will resample all data to be at a common rate of 40 Hz (samples per second).
603 We will exclude individual trials if more than 50% of the gaze data is missing (defined as
604 off-screen or unavailable point of gaze during the whole trial, not just the anticipatory
605 period). Applying this criterion would have caused us to exclude 4% of the trials in our
606 pilot data, which inspection of our pilot data suggested was an appropriate trade-off

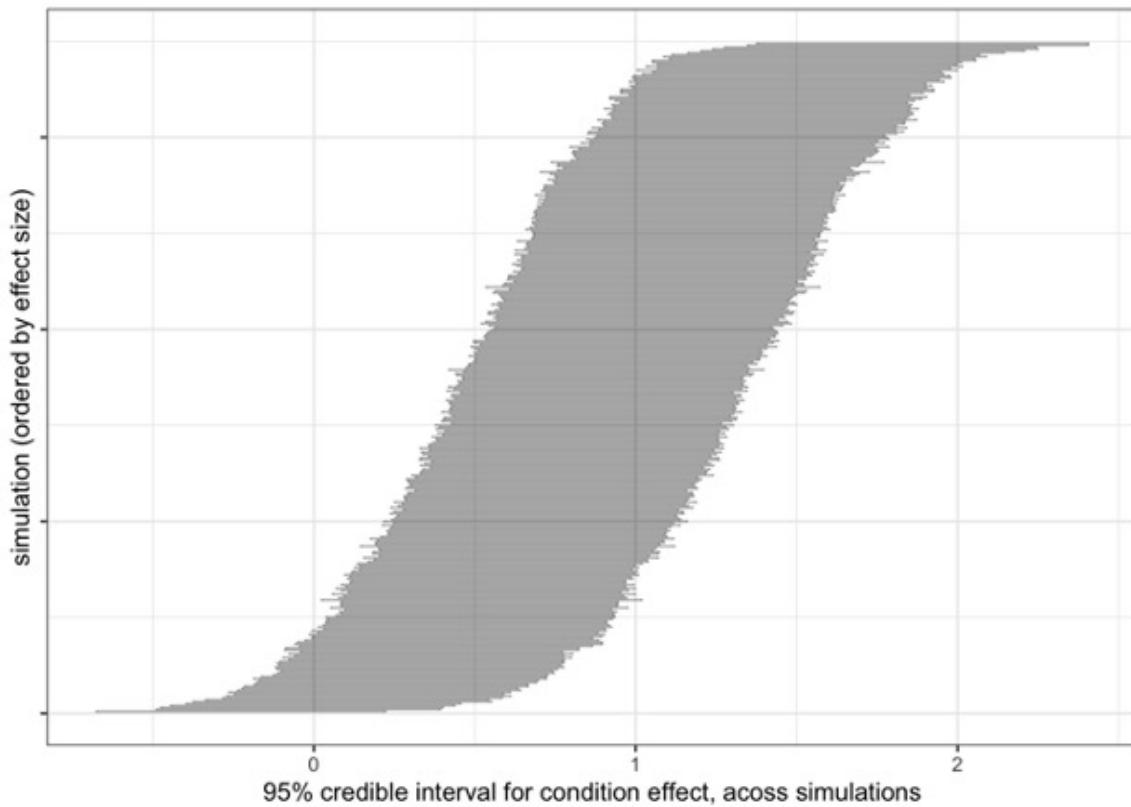


Figure 3. Effect sizes of simulated experiments.

Note. Ordered by effect size (from left to right), 95% credible intervals for the key effect (in logit space) for our simulated experiments that use first look as the dependent variable.

607 between not excluding too much usable data and not analyzing trials which were
 608 uninformative. For each monitor size, we will determine the specific AOIs and compute
 609 whether the specific x- and y-position for each participant, trial, and time point fall within
 610 their screen resolution-specific AOIs. Our goal is to determine whether participants are
 611 anticipating the emergence of the chaser from one of the two tunnel exits. Thus, we defined
 612 AOIs on the stimulus by creating a rectangular region around the tunnel exit that is D
 613 units from the top, bottom, left, and right of the boundary of the tunnel exit, where D is
 614 the diameter of the tunnel exits. We then expanded the sides of the AOI rectangles by 25%
 615 in all directions to account for tracker calibration error. Our rationale was that, if we made
 616 the AOI too small, we might fail to capture anticipations by participants with poor

617 calibrations. In contrast, if we made the regions too large, we might capture some fixations
 618 by participants looking at the box where the chasee actually is. On the other hand, these
 619 chasee looks would not be expected to vary between conditions and so would only affect our
 620 baseline level of looking. Thus, the chosen AOIs aim at maximizing our ability to capture
 621 between-condition differences. For an illustration of the tunnel exit AOIs see Figure 4. We
 622 are not analyzing looks to the boxes, since they can less unambiguously be interpreted as
 623 epistemic state-based action predictions and because we observed few anticipatory looks to
 624 the boxes in the pilot studies. For more detailed information about the AOI definition
 625 process see the description of the pilot study results in the Supplemental Material.

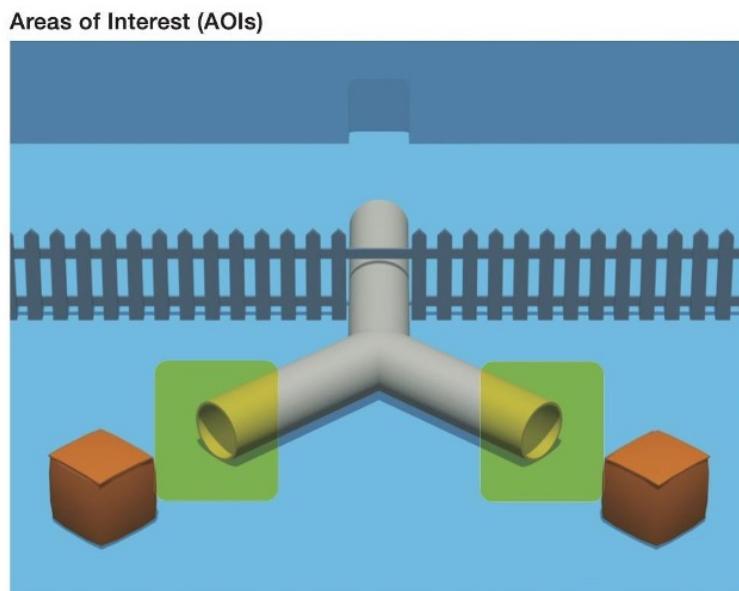


Figure 4. Illustration of Areas of Interest (AOIs) for gaze data analysis during the anticipatory period.

Note. The light green rectangles show the dimensions of the AOIs used for the analysis of AL during the test period.

626 **Manual Coding.** For data gathered without an eye-tracker (e.g., videos of
 627 participants gathered from online administration), precise estimation of looks to specific
 628 AOIs will not be possible. Instead, videos will be coded for whether participants are looking

629 to the left or the right side of the screen (or “other/off screen”). In our main analysis,
630 during the critical anticipatory window, we will treat these looks identically to looks to the
631 corresponding AOI. See exploratory analyses for analysis of data collected online.

632 **Temporal Region of Interest.** For familiarization trials, we define the start of
633 the anticipatory period (total length = 4000 ms) as starting 120 ms after the first frame
634 after which the chaser has completely entered the tunnel and lasting until 120 ms after the
635 first frame at which the chaser is visible again (we chose 120 ms as a conservative value for
636 cutting off reactive saccades; cf., Yang et al., 2002). For test trials, we define the start of
637 the anticipatory period in the same way, with a total duration of 4000 ms.

638 **Dependent Variables.** We define two primary dependent variables: 1. First look.
639 First saccades will be determined as the first change in gaze occurring within the
640 anticipatory time window that is directed towards one of the AOIs. The first look is then
641 the binary variable denoting the target of this first saccade (i.e., either the correct or
642 incorrect AOI) and is defined as the first AOI where participants fixated at for at least 150
643 ms, as in Rayner et al. (2009). The rationale for this definition was that, if participants are
644 looking at a location within the tunnel exit AOIs before the anticipation period, they
645 might have been looking there for other reasons than action prediction. We therefore count
646 only looks that start within the anticipation period because they more unambiguously
647 reflect action predictions. This further prevents us from running into a situation where we
648 would include a lot of fixations on regions other than the tunnel exit AOIs because
649 participants are looking somewhere else before the anticipation period begins. 2.
650 Proportion DLS (also referred to as total relative looking time; Senju et al., 2009). We
651 compute the proportion looking (p) to the correct AOI during the full 4000 ms anticipatory
652 window ($\text{correct looking time} / (\text{correct looking time} + \text{incorrect looking time})$), excluding
653 looks outside of either AOI.

654

Results

655 Confirmatory Analyses

656 **Approach.** As discussed in the Methods section, we adopted a Bayesian analysis
657 strategy so as to maximize our ability to make inferences about the presence or absence of
658 a condition effect (i.e., our key effect of interest). In particular, we fit Bayesian mixed
659 effects regressions using the package brms in R (Bürkner, 2017). This framework allows us
660 to estimate key effects of interest while controlling for variability across grouping units (in
661 our case, labs). To facilitate interpretation of individual coefficients, we report means and
662 credible intervals. For key inferences in our confirmatory analysis, we use the bridge
663 sampling approach (Gronau et al., 2017) to compute BFs comparing different models. As
664 the ratio of the likelihood of the observed data under two different models, BFs allow us to
665 quantify the evidence that our data provide with respect to key comparisons. For example,
666 by comparing models with and without condition effects, we can quantify the strength of
667 the evidence for or against such effects. Bayesian model comparisons require the
668 specification of proper priors on the coefficients of individual models. Here, for our first
669 look analysis, we use a set of weakly informative priors that capture the expectation that
670 the effects that we observe (of condition and, in some cases, trial order) are modest. For
671 coefficients, we choose a normal distribution with mean of 0 and *SD* of 2. Based on our
672 pilot testing and the results of MB1, we assume that lab and participant-level variation will
673 be relatively small, and so for the standard deviation of random effects (i.e., variation in
674 effects across labs and, in the case of the familiarization trials, participants) we set a
675 Normal prior with mean of 0 and *SD* of 0.1. We set an LKJ(2) prior on the correlation
676 matrix in the random effect structure, a prior that is commonly used in Bayesian analyses
677 of this type (Bürkner, 2017). Because the BF is sensitive to the choice of prior, we also ran
678 a secondary analysis with a less informative prior: fixed effect coefficients chosen from a
679 normal distribution with mean 0 and *SD* of 3, and random effect standard deviations

680 drawn from a normal prior with a mean of 0 and SD of 0.5. With respect to the
681 specification of random effects, we followed the approach advocated by Barr et al. (2013),
682 that is, specifying the maximal random effect structure justified by our design. Since we
683 are interested in lab-level variation, we will fit random effect coefficients for fixed effects of
684 interest within labs (e.g., condition within lab). Further, where there were participant-level
685 repeated measure data (e.g., familiarization trials), we fitted random effects of participants.
686 For the proportional looking score analysis, we used a uniform prior on the intercept
687 between -0.5 and 0.5 (corresponding to proportional looking scores between 0 and 1: the
688 full possible range). For the priors on the fixed effect coefficients, we used a normal prior
689 with a mean of 0 and an SD of 0.1. Because these regressions are in proportion space, 0.10
690 corresponds to a change in proportion of 10%. For the random effect priors, we used a
691 normal distribution with mean 0 and standard deviation .05. The LKJ prior was specified
692 as above.

693

694

695 **Familiarization Trials.** Figure 5 shows the proportion of total relative looking
696 time (non-logit transformed) and proportion of first looks for toddlers and adults plotted
697 across familiarization trials and test trials. Our first set of analyses examined data from
698 the four familiarization trials and asked whether participants anticipated the chaser's
699 reappearance at one of the tunnel exits. In our first analysis, we were interested in whether
700 participants engage in AL during the familiarization trials. To quantify the level of
701 familiarization, we fitted Bayesian mixed effect models predicting target looks based on
702 trial number (1-4) with random effects for lab and participants and random slopes for trial
703 number for each. In R formula notation (which we adopt here because of its relative
704 concision compared with standard mathematical notation), our base model was as follows:

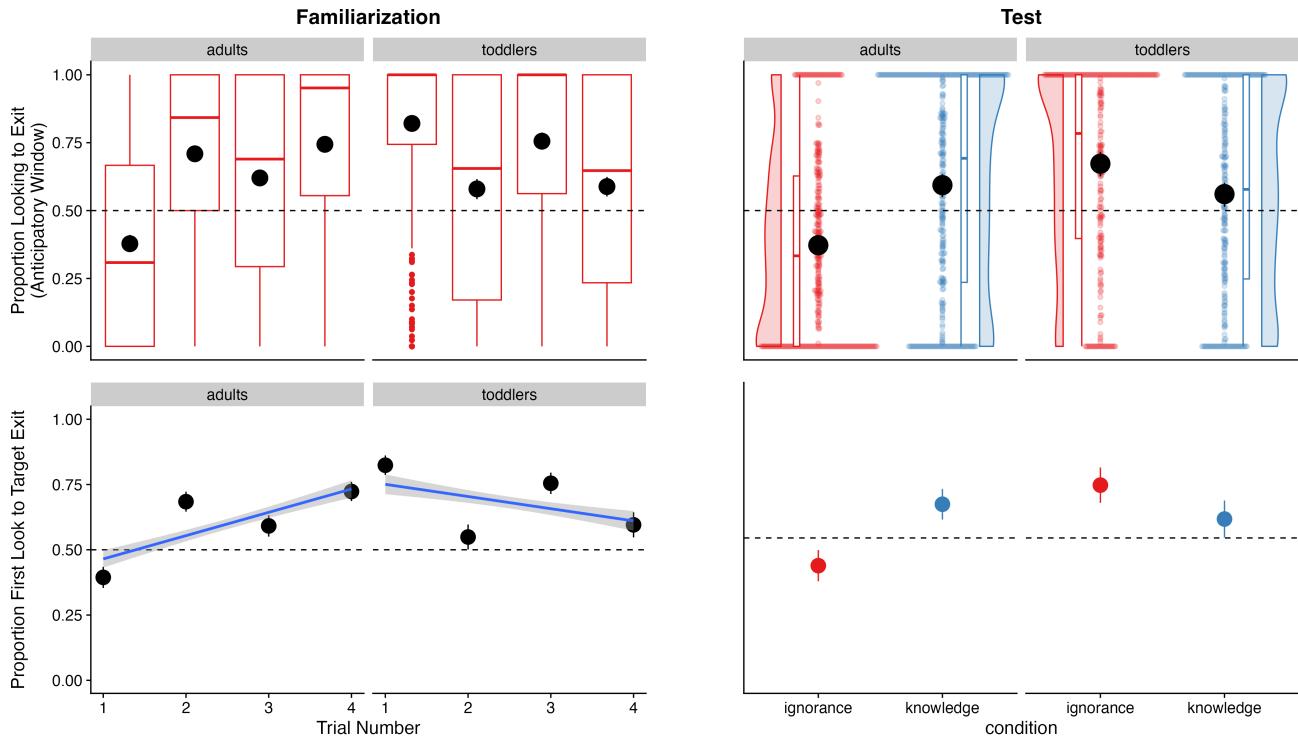


Figure 5. Proportional target looking and proportion of first looks for toddlers and adults during familiarization and test.

705 measure $\sim 1 + \text{trial_number} + (\text{trial_number}|lab) + (\text{trial_number}|participant)$ We
 706 fitted a total of four instances of this model, one for each age group (toddlers vs. adults)
 707 and dependent measure (proportion looking score vs. first look). First look models were
 708 fitted using a logistic link function. The proportion looking score models were Gaussian.
 709 Our key question of interest was whether overall anticipation is higher than chance levels
 710 on the familiarization trial immediately before the test trials, in service of evaluating the
 711 evidence that participants are attentive and making predictive looks immediately prior to
 712 test. To evaluate this question across the four models, we coded trial number so that the
 713 last trial before the test trials (trial 4) was set to the intercept, allowing the model
 714 intercept to encode an estimate of the proportion of correct anticipation immediately
 715 before test. We then fitted a simpler model for comparison
 716 measure $\sim 0 + \text{trial_number} + (\text{trial_number}|lab) + (\text{trial_number}|participant)$, which

717 included no intercept term. We then computed the BF comparing this model to the full
718 model. This BF quantified the evidence for an anticipation effect for each group and
719 measure.

720 ***Proportion of total relative looking time.***

721 *Toddlers.* As first model, we used a Bayesian mixed effects models to predict PTL
722 based on trial number (1-4) for toddlers, with random effects for lab and participants and
723 random slopes for trial number for each. The Bayes factor comparing this model to the
724 simpler null model without the intercept was estimated to be > 1000 , strongly favoring the
725 full model over the null model. See also Table 3 for regression coefficients for the base
726 model. These results suggest a significant effect of trial number on PTL, with the negative
727 coefficient indicating a decrease in PTL across the familiarization trials.

728 *Adults.* Next, we used a Bayesian mixed effects model to predict PTL based on trial
729 number (1-4) for adults, again with random effects for lab and participants and random
730 slopes for trial number for each. The Bayes factor for the full model against the null model
731 was > 1000 , suggesting strong evidence for the full model. These results suggest a
732 significant effect of trial number on PTL, with the positive coefficient indicating an
733 increase in target looks across the familiarization trials.

734 ***Proportion of first looks.***

735 *Toddlers.* Investigating proportion of first looks to the target location for toddlers,
736 we again used a Bayesian mixed effects model to predict whether toddlers first look was to
737 the target exit based on trial number (1-4), with random effects for lab and participants
738 and random slopes for trial number for each. The Bayes factor comparing the full model to
739 the simpler model was estimated to be > 1000 , strongly favoring the full model over the
740 null model. The model also provided support for an effect of trial number on proportion of
741 first looks, with the negative coefficient indicating a decrease in target looks across the
742 familiarization trials.

743 Adults. Comparing the Bayesian mixed effects model of adults predicting proportion

744 of first looks based on trial number (1-4), with random effects for lab and participants and

745 random slopes for trial number for each with the simpler model without an intercept, we

746 computed a Bayes factor of > 1000, strongly favoring the base model over the full model.

747 There was again support for an effect of trial number on proportion of first looks, with the

748 positive coefficient indicating an increase in proportion of first target looks across the

749 familiarization trials.

*750 **Test Trials.*** We focused our confirmatory analysis on the first test trial (see

751 Exploratory Analysis section for an analysis of both trials). Our primary question of

752 interest was whether AL differs between conditions (knowledge vs. ignorance, coded as

753 -.5/.5) and by age (in months, centered). For child participants, we fitted models with the

754 specification:

755 measure 1 + condition + age + condition : age + (1 + condition + age + condition : age | lab).

756 For adult participants, we fitted models with the specification

757 measure 1 + condition + (1 + condition | lab). Again, we fitted models with a logistic link

758 for first look analyses and with a standard linear link for DLS. In each case, our key BF

759 was a comparison of this model with a simpler “null” model that did not include the fixed

760 effect of condition but still included other terms. We take a BF > 3 in favor of a particular

761 model as substantial evidence and a BF > 10 in favor of strong evidence. A BF < 1/3 is

762 taken as substantial evidence in favor of the simpler model, and a BF < 1/10 as strong

763 evidence in favor of the simpler model. For the model of data from toddlers, we

764 additionally were interested in whether the model shows changes in AL with age. We

765 assessed evidence for this by computing BFs related to the comparison with a model that

766 did not include an interaction between age and condition as fixed effects

measure 1 + condition + age + (1 + condition + age + condition : age | lab).

767 These BFs captured the evidence for age-related changes in the difference in action

768 anticipation between the two conditions. It is important to note that in the case of a null

769 effect, there are two main explanations: (1) toddlers and adults in our study do not
770 distinguish between knowledgeable and ignorant agents when predicting their actions. (2)
771 The method used is not appropriate to reveal knowledge/ignorance understanding. By
772 using Bayesian analyses, we are able to better evaluate the first of these two possibilities:
773 The BF provides a measure of our statistical confidence in the null hypothesis, i.e., no
774 difference between experimental conditions, given the data in ways that standard null
775 hypothesis significance testing does not. In other words, instead of merely concluding that
776 we did not find a difference between conditions, we would be able to find
777 no/anecdotal/moderate/strong/very strong/extreme evidence for the null hypothesis that
778 our participants did not distinguish between knowledgeable and ignorant agents when
779 predicting their actions (Schönbrodt & Wagenmakers, 2018). We therefore consider this
780 analysis an important addition to our overall analysis strategy. Yet, even our Bayesian
781 analyses are not able to rule out the second possibility that participants may well show
782 such knowledge/ignorance understanding with different methods, or that this ability may
783 not be measurable with any methods available at the current time. Addressing this
784 alternative explanation warrants follow up experiments.

785 ***Proportion of total relative looking time.***

786 *Toddlers.* As first model, we used a Bayesian mixed effects models to predict
787 toddlers' PTL based on condition, age, and the interaction of condition and age, while
788 accounting for variability across labs. The Bayes factor comparing this model to the simpler
789 null model without the main effect of condition was estimated to be $\text{BF} = 33.3$, favoring
790 the full model over the null model. Table 4 shows the statistics for regression coefficients of
791 the full model. These results suggest a significant effect of condition on PTL, with the
792 positive coefficient indicating higher PTL for ignorance trials compared to knowledge trials.

793 *Adults.* Next, we used a Bayesian mixed effects model to predict PTL based on
794 condition for adults, again with random effects for lab. The Bayes factor comparing this
795 model to the simpler null model without the main effect of condition was estimated to be

796 > 1000, strongly favoring the full model over the null model. These results suggest a
797 significant main effect of condition on PTL, with the negative coefficient indicating a
798 higher number of target looks for knowledge than for ignorance trials.

799 ***Proportion of first looks.***

800 *Toddlers.* Investigating proportion of first looks for toddlers, we again used a
801 Bayesian mixed effects model to predict target looks based on condition, with random
802 effects for lab. The Bayes factor comparing the full model to the simpler model was
803 estimated to be $BF = 2.7$, providing no substantial evidence in favor of the full model over
804 the null model.

805 *Adults.* We compared a Bayesian mixed-effects model predicting the proportion of
806 first looks based on condition, including random effects for lab to a simpler model without
807 the main effect of condition. The analysis yielded a Bayes factor of > 1000, providing
808 strong evidence in favor of the full model over the null model. Results indicated that first
809 looks to the target were significantly more frequent in the knowledge condition compared
810 to the ignorance condition.

811 **Exploratory Analyses**

812 [WE LIST POTENTIAL EXPLORATORY ANALYSES HERE TO SIGNAL OUR
813 INTEREST AND INTENTIONS BUT DO NOT COMMIT TO THEIR INCLUSION,
814 DUE TO LENGTH AND OTHER CONSIDERATIONS]

815 1. Spill-over: we will analyze within-participants data from the second test trial that
816 participants saw, using exploratory models to assess whether (1) findings are
817 consistent when both trials are included (overall condition effect), (2) whether effects
818 are magnified or diminished on the second trial (order main effect), and (3) whether
819 there is evidence of “spillover” - dependency in anticipation on the second trial
820 depending on what the first trial is (condition x order interaction effect).

- 821 2. We will explore whether condition differences vary for participants who show higher
822 rates of anticipation during the four familiarization trials. For example, we might
823 group participants according to whether they did or did not show correct AL at the
824 end of the familiarization phase, defined as overall longer looking at the correct AOI
825 than the incorrect AOI on average in trials 3 and 4 of the familiarization phase.
- 826 3. In analyses introducing model terms for certain measurement characteristics (e.g.,
827 types of eye-tracker manufacturers, screen dimensions), we will quantify potential
828 variability between different in-lab data acquisition methods (cf., ManyBabies
829 Consortium, 2020). If we have a sufficiently large sample of participants tested with
830 online sources (e.g., contributions of at least 32 participants), we will conduct a
831 separate analysis with a model term for online participants that estimates whether
832 condition effects are different in this population. We will further report whether
833 exclusion rates are different for this population.
- 834 4. If we observe substantial looking (defined *post hoc* by evaluating scatter plot videos
835 of gaze data) to the boxes as well as the tunnel exit AOIs, we will conduct an
836 exploratory analysis using tighter AOIs around tunnel exits and boxes, asking
837 whether box and tunnel looking vary separately by age or by condition. In particular,
838 we expect that the difference in AL between the two conditions will be bigger for the
839 tunnel exits than for the box (as looks to the correct box might indicate looks to the
840 target, which is in the same box for both conditions, rather than action anticipation).
- 841 5. To examine whether participants monitor both the bear and the mouse during the
842 mouse's location change, and how this may influence AL in the test phase, we define
843 new time windows of interest (TOIs) corresponding to the mouse's location change in
844 each condition and areas of interest (AOIs) for both the mouse and bear. We
845 hypothesize that participants who attend to both AOIs will exhibit greater AL
846 compared to those who predominantly track the mouse during its location change.

847 Specifically, we will analyze the frequency of gaze shifts between the mouse and bear,
848 as well as the duration of gaze directed toward each AOI during the mouse's location
849 change.

850 **Spill-over.** Analyzing condition-effects of within-participants data for both test
851 trials, we fitted a Bayesian mixed-effects model with the dependent variable of PTL and
852 main effects of condition and age and their interaction for toddlers. Comparing this full
853 model to a null model that did not include the fixed effect of condition, we obtained a
854 Bayes Factor of > 1000 , providing very strong evidence in favor of the full model. The
855 effect of condition was positive and credible, indicating PTL was higher in the ignorance
856 condition compared to the knowledge condition. The main effect of age was small and
857 uncertain, suggesting minimal influence of age on PTL. The interaction between condition
858 and age was also small and inconclusive, indicating that the effect of condition on PTL did
859 not differ substantially with age.

860 For adults, we also fitted a Bayesian mixed-effects model to predict their PTL for
861 both test trials with the main effect of condition and random effects for participant and
862 lab. Again, the data provided very strong evidence for the inclusion of the main effect of
863 condition with a Bayes Factor of > 1000 . The effect of condition was negative and credible,
864 suggesting that PTL was significantly lower in the ignorance condition compared to the
865 knowledge condition.

866 In order to investigate whether there's an interaction of condition and test trial
867 number, we fitted Bayesian mixed-effects model to predict PTL with fixed effects for
868 condition, test trial number, and their interaction, along with random intercepts and slopes
869 for these variables across labs, for toddlers and adults separately. While for toddlers, the
870 results were inconclusive ($BF = 0.6$), for adults, the Bayes Factor of > 1000 provided
871 strong evidence for including the interaction of condition and test trial number as fixed
872 effect. Overall, the results demonstrate that while PTL increased over trials, this effect was

873 moderated by the condition, with the ignorance condition showing a slower rate of increase
 874 compared to the knowledge condition.

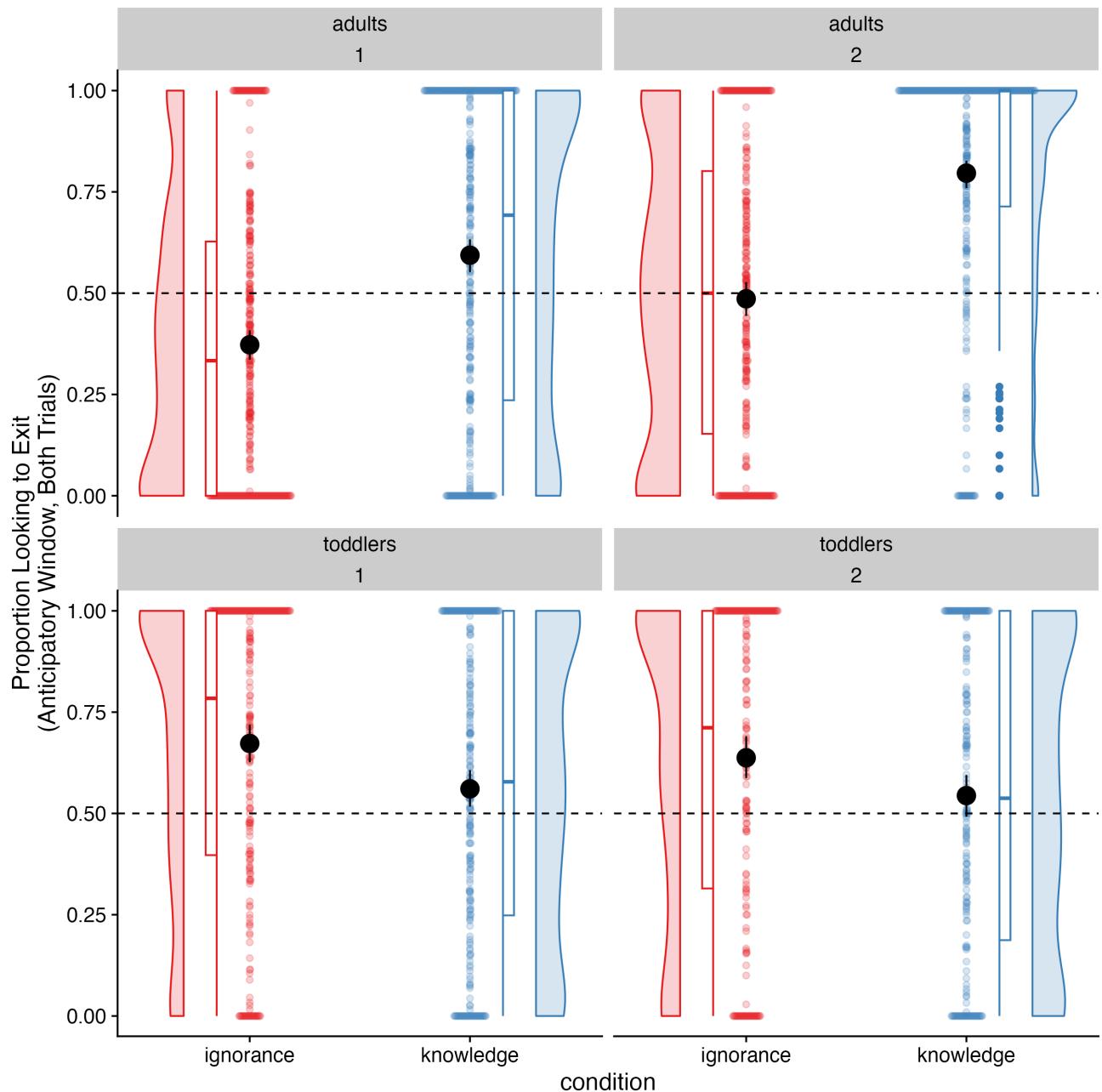


Figure 6. Proportional exit looking for the first and second test trial for toddlers and adults in the ignorance and knowledge condition.

875 **Relationship between familiarization and test.** To investigate whether only

876 participants that show anticipatory looking within the familiarization also display

877 anticipatory looking in test, we explored three different measures. First, we assessed

878 anticipatory looking for participants that successfully anticipated during the last

879 familiarization trial, that is, whose first fixation was on the target. Second, we

880 ***Only anticipators on final familiarization trial.*** We fitted a main Bayesian

881 hierarchical model testing the effect of condition (ignorance vs. knowledge) on first-trial

882 proportion target looking during the anticipatory window for only those participants who

883 anticipated correctly during the last familiarization trial (trial 4, first look to target) for

884 toddlers and adults separately. The results revealed a very similar pattern to the analysis

885 with all participants. However, for toddlers, the Bayes factor comparing this model to the

886 simpler null model without the main effect of condition was inconclusive ($BF = 0.5$). For

887 adults, the Bayes factor comparing this model to the simpler null model without the main

888 effect of condition was estimated to be > 1000 , strongly favoring the base model over the

889 null model. Again, this result suggests a significant main effect of condition on PTL, with

890 the negative coefficient indicating a higher number of target looks for knowledge than for

891 ignorance trials.

892 ***Only >50% looking to target during familiarization trials.*** In addition, we

893 fitted main Bayesian hierarchical models testing the effect of condition (ignorance

894 vs. knowledge) on first-trial proportion target looking during the anticipatory window for

895 only those participants who fixated the target more than half of the time during all

896 familiarization trial. Comparing the full model to the null model of toddlers revealed a

897 Bayes Factor of $BF = 14.9$, providing evidence in favor of the full model that included the

898 fixed effect of condition. The estimated Bayes factor in favor of the full model of adults

899 over the null model was approximately > 1000 , indicating that the inclusion of condition in

900 the full model substantially improved the explanation of the observed data.

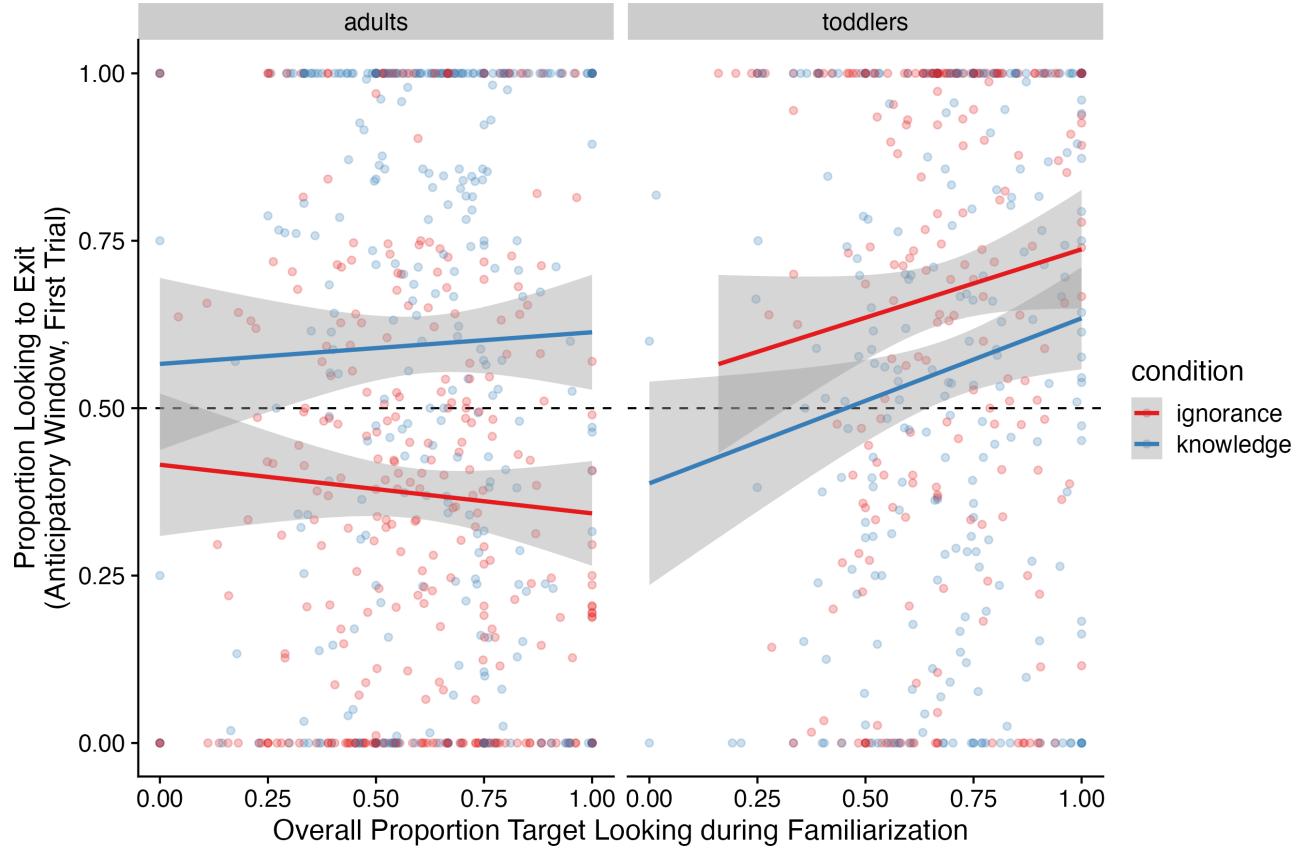


Figure 7. Relationship of anticipatory looking during familiarization and test for both age cohorts and conditions.

Correlation between familiarization and test. We also examined the

correlation between familiarization and test performance across the two age cohorts and conditions (see Figure 7). While no significant correlations were found for adults in either condition, toddlers in the knowledge condition exhibited a significant positive correlation of anticipatory looking in familiarization and test, $r=0.15$, $t(254)=2.35$, $p=0.02$.

Data collection type: in-lab vs. web-based. Bayesian mixed-effects model were

used to evaluate the effects of condition, method, and their interaction on anticipatory looking. The models included fixed effects for condition, method, and their interaction. For toddlers, the effect of method was small and uncertain, with the credible interval including zero, indicating no clear effect of method. The interaction between condition and method

911 was minimal and also uncertain, suggesting no strong evidence that the effect of condition
912 varied by method. The estimated Bayes factor comparing the full model to the null model
913 was approximately $BF = 0.8$, which indicates that the data slightly favors the null model
914 over the full model. This suggests that the predictors included in the full model do not
915 substantially improve the explanation of the observed data compared to the null model.

916 For adults, the main effect of method was slightly negative but uncertain, suggesting
917 that the method had little to no clear effect on the outcome. The interaction between
918 condition and method was negative but with a wide credible interval crossing zero,
919 indicating uncertainty about whether the effect of condition varied by method. The
920 estimated Bayes factor in favor of the full model over the null model was approximately BF
921 $= 3.1$. This Bayes factor indicates that the evidence in favor of the model is modest but
922 not strong. While the model is more likely than the null model to explain the observed
923 data, the support is relatively weak, suggesting that the predictors in the model provide
924 only a small improvement in explaining the data compared to the null model.

925 In sum, the analysis suggests that the method used (web-based vs. in-lab) does not
926 have a strong impact on anticipatory looking, as the effect of method and its interaction
927 with condition were small and uncertain. Additionally, the results should be interpreted
928 with caution due to the relatively small sample size for web-based data compared to in-lab
929 data collection, which may limit the robustness of the findings.

930 **Box and tunnel looking vary separately by age or by condition.** We will
931 conduct an exploratory analysis using tighter AOIs around tunnel exits and boxes, asking
932 whether box and tunnel looking vary separately by age or by condition. In particular, we
933 expect that the difference in AL between the two conditions will be bigger for the tunnel
934 exits than for the box (as looks to the correct box might indicate looks to the target, which
935 is in the same box for both conditions, rather than action anticipation).

936 **Looking patterns during mouse's change of location.**

937 ***Comparing the number of shifts of toddlers and adults during the***

938 ***location change of the mouse.*** We fitted a Bayesian mixed-effects model to examine
939 the relationship between the number of shifts between mouse and bear and age cohort
940 during location change of the mouse, while accounting for random effects by lab. The effect
941 of condition was negative and approached significance, suggesting a potential reduction in
942 the number of shifts for the ignorance condition compared to the knowledge condition. The
943 main effect of age cohort was positive and credible, Estimate=0.34, indicating that the the
944 number of shifts was higherfor adults than for toddlers. Importantly, the interaction
945 between condition and age cohort was negative and credible, indicating that the negative
946 effect of condition was more pronounced in the adult cohort (see Figure 8). Comparing this
947 model to a simpler model without the interaction of condition and age cohort, a Bayes
948 Factor of $BF = 694.8$ was computed. This provides strong evidence in favor of including
949 the interaction of condition and age cohort in the model.

950 ***Number of gaze shifts between mouse and bear during location change as***

951 ***a function of AL.*** In order to examine the effect of condition and the number of shifts
952 between mouse and bear during location change of the mouse on anticipatory looking, we
953 fitted Bayesian mixed-effects models for each age cohort separately. The dependent
954 variable was PTL in the anticipation period. The fixed effects included the main effects of
955 condition, the number of shifts, and their interaction. We also included random intercepts
956 and slopes for number of shifts within each participant and within each lab, allowing us to
957 account for the hierarchical structure of the data and potential variability between
958 participants. For toddlers, comparing this model to a simpler model without the
959 interaction of condition and number of shifts, a Bayes Factor of $BF = 0.0$ was computed,
960 indicating that the data strongly favors the null model over the full model. This suggests
961 that the predictors number of shifts and the interaction with condition included in the full
962 model do not improve the explanation of the observed data compared to the null model.

963 For adults, the number of shifts showed a small but credible positive effect,

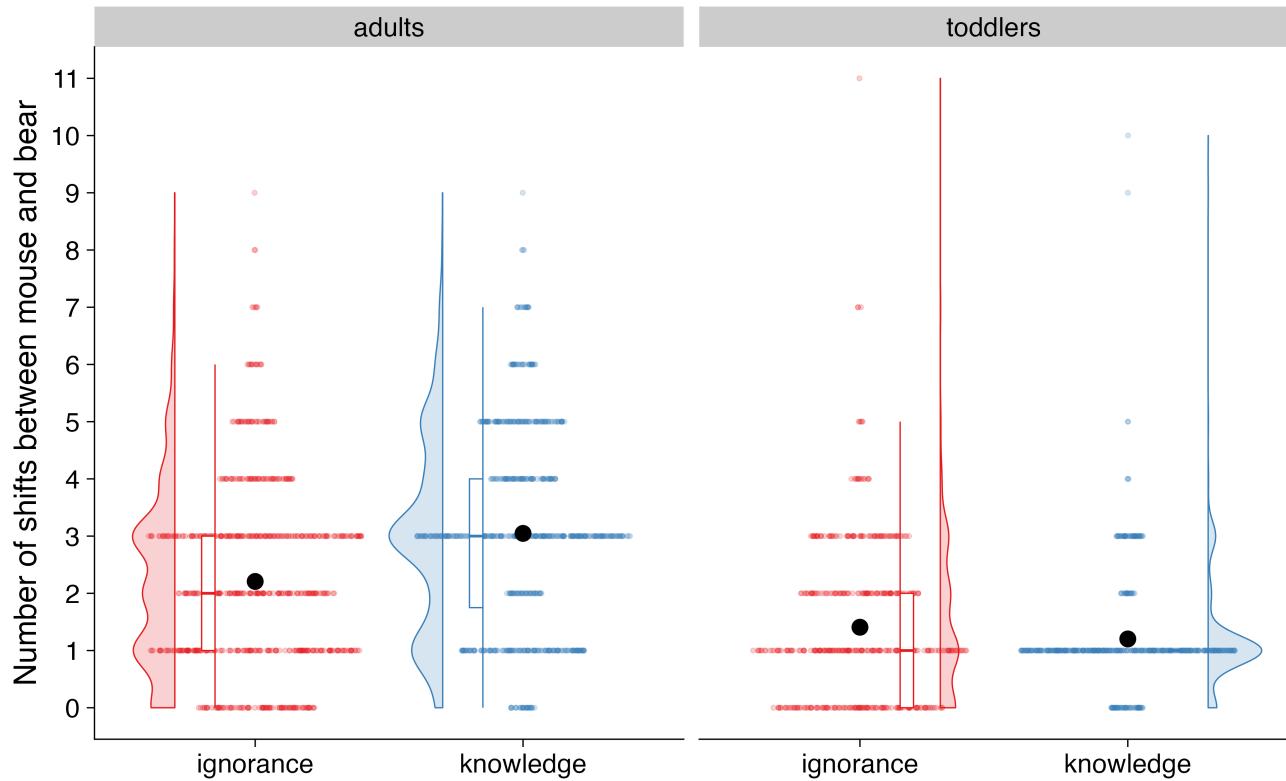


Figure 8. Number of shifts between mouse and bear during location change of mouse in the test phase for toddlers and adults in the ignorance and knowledge condition.

964 suggesting that more shifts were associated with an increase in PTL. The interaction
 965 between condition and the number of shifts was negative and credible, indicating that the
 966 effect of condition became more negative as the number of shifts increased. The estimated
 967 Bayes factor comparing the full model to the null model was approximately > 1000 ,
 968 providing strong evidence in favor of the full model over the null model.

969 ***Differential fixation times of bear and mouse during location change of***
 970 ***mouse as a function of AL.*** In order to examine the effect of condition and the
 971 difference in looking times for mouse and bear during location change of the mouse on
 972 anticipatory looking, we fitted a Bayesian mixed-effects model. The dependent variable was
 973 the proportion of target looking. The fixed effects included the main effects of condition,
 974 the difference in fixation times of mouse and bear, and their interaction. We also included

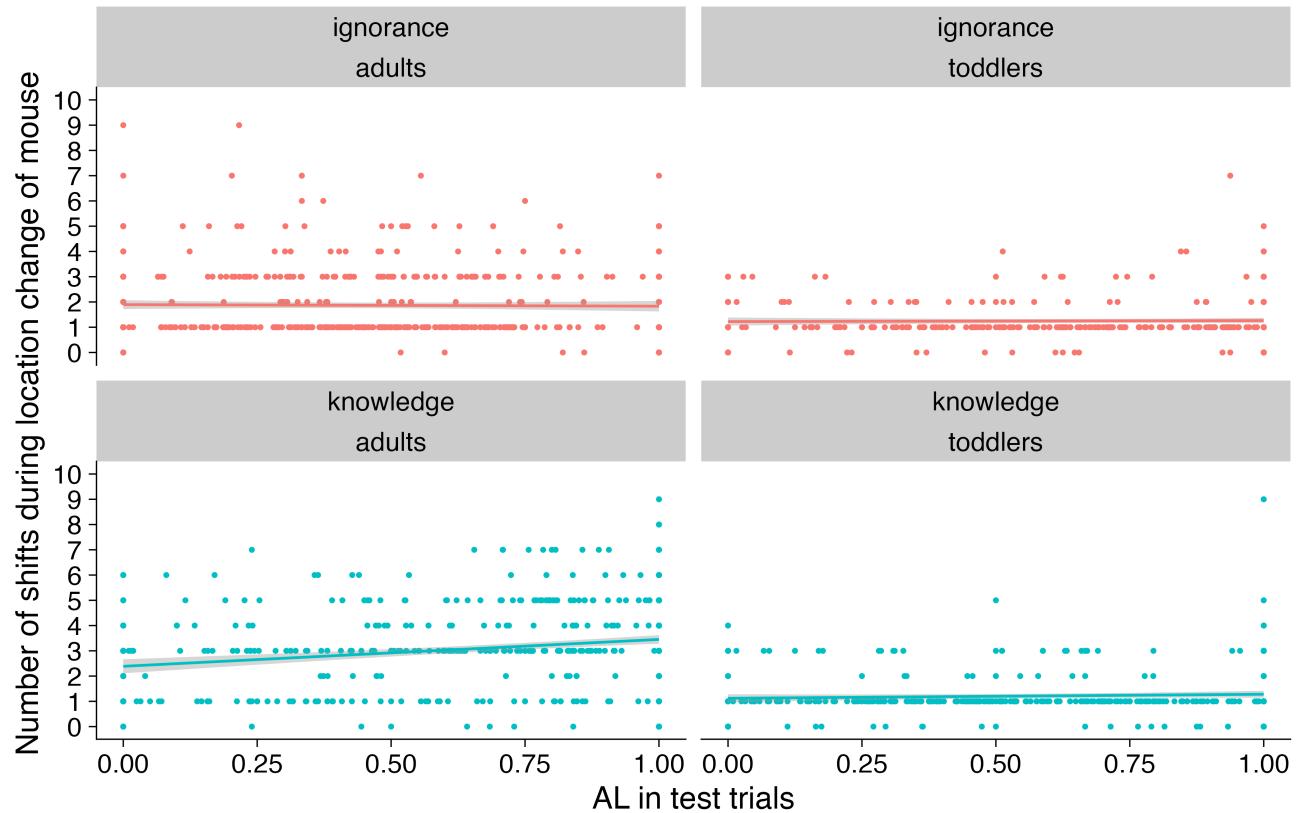


Figure 9. Number of shifts between mouse and bear during location change of mouse as a function of AL in the test phase for toddlers and adults.

975 random intercepts and slopes for differences in fixation times of mouse and bear within
 976 each participant and within each lab, allowing us to account for the hierarchical structure
 977 of the data and potential variability between participants. The fixed effect of difference in
 978 mouse-bear looking on anticipatory looking is estimated to be 0. Comparing this model to
 979 a simpler model without the difference in mouse-bear looking, a Bayes Factor of $BF = 0.0$
 980 was computed. This provides extremely strong evidence against including the difference in
 981 mouse-bear looking in the model.

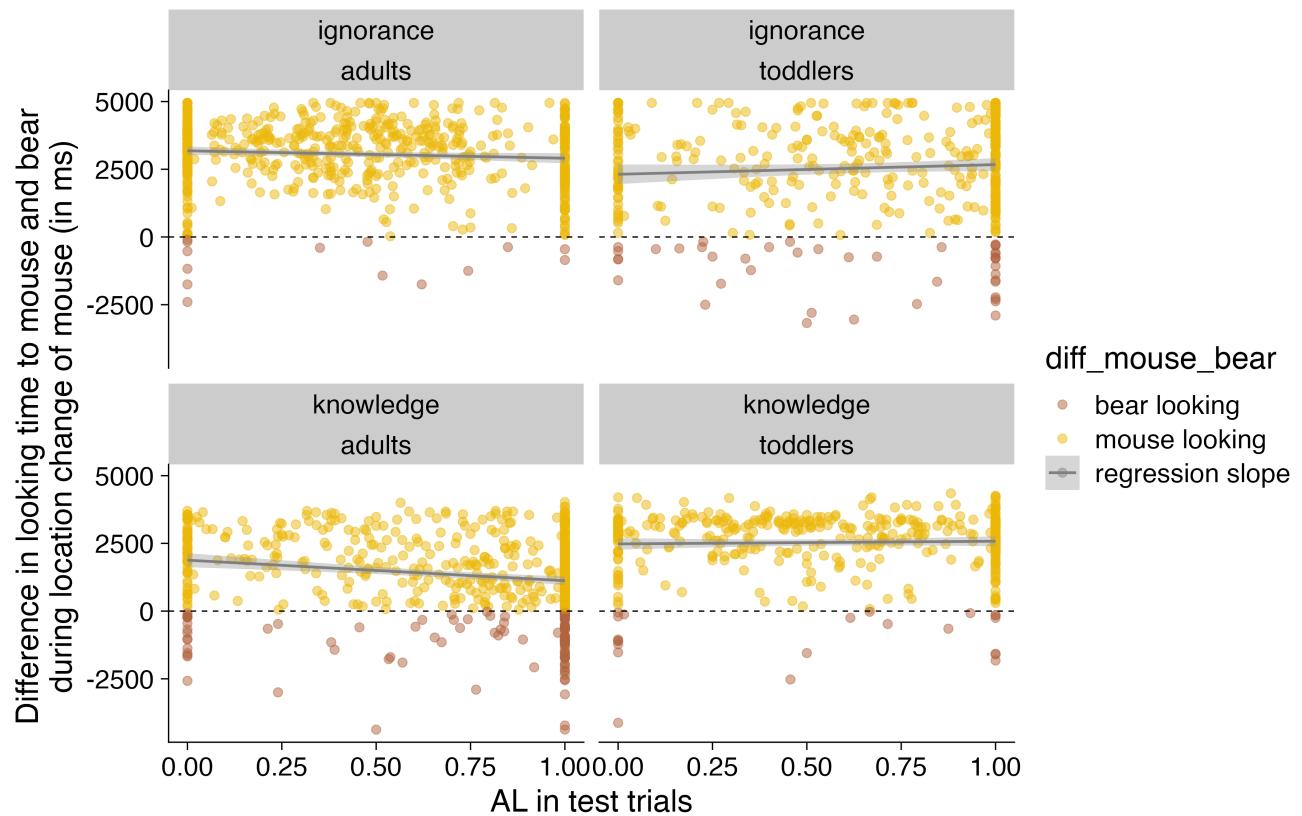


Figure 10. Difference in looking time to mouse and bear during location change of mouse (in ms) as a function of AL for each age cohort and each condition. Higher looking times at mouse are colored in yellow, and higher looking times at bear are colored in brown.

982

General Discussion

983 The current large-scale, multi-lab study set out to examine whether toddlers and
 984 adults engage in spontaneous ToM. In particular, we used an anticipatory looking
 985 paradigm to explore whether 18- to 27-month-old toddlers and adults distinguish between
 986 two basic forms of epistemic states: knowledge and ignorance. Our call for participation
 987 resulted in contributions from 47 labs, representing a total of 809 toddlers from xyz
 988 countries and 805 adults from xyz countries, of which 1224 were included in the final
 989 sample used for analysis (see Table 1). We begin our discussion by summarizing the
 990 principal results of the study with respect to confirmatory analysis and then discuss

⁹⁹¹ limitations of the study as well as future directions.

⁹⁹² **Conclusion**

References

- 993
- 994 Baillargeon, R., Buttelmann, D., & Southgate, V. (2018). Invited commentary:
995 Interpreting failed replications of early false-belief findings: Methodological and
996 theoretical considerations. *Cognitive Development*, 46, 112–124.
- 997 Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants.
998 *Trends in Cognitive Sciences*, 14(3), 110–118.
- 999 Barone, P., Corradi, G., & Gomila, A. (2019). Infants' performance in
1000 spontaneous-response false belief tasks: A review and meta-analysis. *Infant
1001 Behavior and Development*, 57, 101350.
- 1002 Burnside, K., Ruel, A., Azar, N., & Poulin-Dubois, D. (2018). Implicit false belief
1003 across the lifespan: Non-replication of an anticipatory looking task. *Cognitive
1004 Development*, 46, 4–11.
- 1005 Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old
1006 infants show false belief understanding in an active helping paradigm. *Cognition*,
1007 112(2), 337–342.
- 1008 Carruthers, P. (2013). Mindreading in infancy. *Mind & Language*, 28(2), 141–172.
- 1009 Csibra, G., & Gergely, G. (2007). ‘Obsessed with goals’: Functions and mechanisms
1010 of teleological interpretation of actions in humans. *Acta Psychologica*, 124(1),
1011 60–78.
- 1012 Dennett, D. C. (1989). *The intentional stance*. MIT press.
- 1013 Dörrenberg, S., Rakoczy, H., & Liszkowski, U. (2018). How (not) to measure infant
1014 theory of mind: Testing the replicability and validity of four non-verbal
1015 measures. *Cognitive Development*, 46, 12–30.
- 1016 Elsner, B., & Adam, M. (2021). Infants' goal prediction for simple action events:
1017 The role of experience and agency cues. *Topics in Cognitive Science*, 13(1),
1018 45–62.
- 1019 Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., ...

- 1020 Yurovsky, D. (2017). A collaborative approach to infant research: Promoting
1021 reproducibility, best practices, and theory-building. *Infancy*, 22(4), 421–435.
1022 <https://doi.org/10.1111/infa.12182>
- 1023 Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron*, 50(4),
1024 531–534.
- 1025 Ganglmayer, K., Attig, M., Daum, M. M., & Paulus, M. (2019). Infants' perception
1026 of goal-directed actions: A multi-lab replication reveals that infants anticipate
1027 paths and not goals. *Infant Behavior and Development*, 57, 101340.
- 1028 Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naïve
1029 theory of rational action. *Trends in Cognitive Sciences*, 7(7), 287–292.
- 1030 Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional
1031 stance at 12 months of age. *Cognition*, 56(2), 165–193.
- 1032 Gliga, T., Jones, E. J., Bedford, R., Charman, T., & Johnson, M. H. (2014). From
1033 early markers to neuro-developmental mechanisms of autism. *Developmental
1034 Review*, 34(3), 189–207.
- 1035 Grosse Wiesmann, C., Friederici, A. D., Singer, T., & Steinbeis, N. (2017). Implicit
1036 and explicit false belief development in preschool children. *Developmental
1037 Science*, 20(5), e12445.
- 1038 Hayashi, T., Akikawa, R., Kawasaki, K., Egawa, J., Minamimoto, T., Kobayashi,
1039 K., ... others. (2020). Macaques exhibit implicit gaze bias anticipating others'
1040 false-belief-driven actions via medial prefrontal cortex. *Cell Reports*, 30(13),
1041 4433–4444.
- 1042 Kampis, D., Karman, P., Csibra, G., Southgate, V., & Hernik, M. (2021). A two-lab
1043 direct replication attempt of southgate, senju and csibra (2007). *Royal Society
1044 Open Science*, 8(8), 210190.
- 1045 Kampis, D., & Southgate, V. (2020). Altercentric cognition: How others influence
1046 our cognitive processing. *Trends in Cognitive Sciences*, 24(11), 945–959.

- 1047 Kano, F., Krupenye, C., Hirata, S., Tomonaga, M., & Call, J. (2019). Great apes
1048 use self-experience to anticipate an agent's action in a false-belief test.
1049 *Proceedings of the National Academy of Sciences*, 116(42), 20904–20909.
- 1050 Király, I., Oláh, K., Csibra, G., & Kovács, Á. M. (2018). Retrospective attribution
1051 of false beliefs in 3-year-old children. *Proceedings of the National Academy of*
1052 *Sciences*, 115(45), 11477–11482.
- 1053 Knudsen, B., & Liszkowski, U. (2012). 18-month-olds predict specific action
1054 mistakes through attribution of false belief, not ignorance, and intervene
1055 accordingly. *Infancy*, 17(6), 672–691.
- 1056 Kovács, Á. M. (2016). Belief files in theory of mind reasoning. *Review of Philosophy*
1057 *and Psychology*, 7, 509–527.
- 1058 Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility
1059 to others' beliefs in human infants and adults. *Science*, 330(6012), 1830–1834.
- 1060 Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes
1061 anticipate that other individuals will act according to false beliefs. *Science*,
1062 354(6308), 110–114.
- 1063 Kulke, L., Duhn, B. von, Schneider, D., & Rakoczy, H. (2018). Is implicit theory of
1064 mind a real and robust phenomenon? Results from a systematic replication
1065 study. *Psychological Science*, 29(6), 888–900.
- 1066 Kulke, L., & Hinrichs, M. A. B. (2021). Implicit theory of mind under realistic
1067 social circumstances measured with mobile eye-tracking. *Scientific Reports*,
1068 11(1), 1215.
- 1069 Kulke, L., Johannsen, J., & Rakoczy, H. (2019). Why can some implicit theory of
1070 mind tasks be replicated and others cannot? A test of mentalizing versus
1071 submentalizing accounts. *PloS One*, 14(3), e0213772.
- 1072 Kulke, L., & Rakoczy, H. (2018). Implicit theory of mind—an overview of current
1073 replications and non-replications. *Data in Brief*, 16, 101–104.

- 1074 Kulke, L., & Rakoczy, H. (2019). Testing the role of verbal narration in implicit
1075 theory of mind tasks. *Journal of Cognition and Development*, 20(1), 1–14.
- 1076 Kulke, L., Reiß, M., Krist, H., & Rakoczy, H. (2018). How robust are anticipatory
1077 looking measures of theory of mind? Replication attempts across the life span.
1078 *Cognitive Development*, 46, 97–111.
- 1079 Kulke, L., Wübker, M., & Rakoczy, H. (2019). Is implicit theory of mind real but
1080 hard to detect? Testing adults with different stimulus materials. *Royal Society
1081 Open Science*, 6(7), 190068.
- 1082 Leslie, A. M. (2005). Developmental parallels in understanding minds and bodies.
1083 *Trends in Cognitive Sciences*, 9(10), 459–462.
- 1084 Liszkowski, U., Carpenter, M., & Tomasello, M. (2007). Pointing out new news, old
1085 news, and absent referents at 12 months of age. *Developmental Science*, 10(2),
1086 F1–F7.
- 1087 Luo, Y., & Baillargeon, R. (2007). Do 12.5-month-old infants consider what objects
1088 others can see when interpreting their actions? *Cognition*, 105(3), 489–512.
- 1089 Luo, Y., & Baillargeon, R. (2010). Toward a mentalistic account of early
1090 psychological reasoning. *Current Directions in Psychological Science*, 19(5),
1091 301–307.
- 1092 ManyBabies Consortium. (2020). Quantifying sources of variability in infancy
1093 research using the infant-directed-speech preference. *Advances in Methods and
1094 Practices in Psychological Science*, 3(1), 24–52.
- 1095 Meristo, M., Morgan, G., Geraci, A., Iozzi, L., Hjelmquist, E., Surian, L., & Siegal,
1096 M. (2012). Belief attribution in deaf and hearing infants. *Developmental
1097 Science*, 15(5), 633–640.
- 1098 O'Neill, D. K. (1996). Two-year-old children's sensitivity to a parent's knowledge
1099 state when making requests. *Child Development*, 67(2), 659–677.
- 1100 Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false

- beliefs? *Science*, 308(5719), 255–258.
- Perner, J. (1991). *Understanding the representational mind*. The MIT Press.
- Phillips, J., Buckwalter, W., Cushman, F., Friedman, O., Martin, A., Turri, J., ...
- Knobe, J. (2021). Knowledge before belief. *Behavioral and Brain Sciences*, 44, e140.
- Poulin-Dubois, D., Rakoczy, H., Burnside, K., Crivello, C., Dörrenberg, S., Edwards, K., ... others. (2018). Do infants understand false beliefs? We don't know yet—a commentary on baillargeon, buttelmann and southgate's commentary. *Cognitive Development*, 48, 302–315.
- Powell, L. J., Hobbs, K., Bardis, A., Carey, S., & Saxe, R. (2018). Replications of implicit theory of mind tasks with varying representational demands. *Cognitive Development*, 46, 40–50.
- Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4), 515–526.
- Priewasser, B., Fowles, F., Schweller, K., & Perner, J. (2020). Mistaken max befriends duplo girl: No difference between a standard and an acted-out false belief task. *Journal of Experimental Child Psychology*, 191, 104756.
- Priewasser, B., Rafetseder, E., Gargitter, C., & Perner, J. (2018). Helping as an early indicator of a theory of mind: Mentalism or teleology? *Cognitive Development*, 46, 69–78.
- Schneider, D., Bayliss, A. P., Becker, S. I., & Dux, P. E. (2012). Eye movements reveal sustained implicit processing of others' mental states. *Journal of Experimental Psychology: General*, 141(3), 433.
- Schneider, D., Lam, R., Bayliss, A. P., & Dux, P. E. (2012). Cognitive load disrupts implicit theory-of-mind processing. *Psychological Science*, 23(8), 842–847.
- Schneider, D., Slaughter, V. P., Bayliss, A. P., & Dux, P. E. (2013). A temporally sustained implicit theory of mind deficit in autism spectrum disorders.

- 1128 *Cognition*, 129(2), 410–417.
- 1129 Schuwerk, T., Priewasser, B., Sodian, B., & Perner, J. (2018). The robustness and
1130 generalizability of findings on spontaneous false belief sensitivity: A replication
1131 attempt. *Royal Society Open Science*, 5(5), 172273.
- 1132 Scott, R. M., & Baillargeon, R. (2017). Early false-belief understanding. *Trends in*
1133 *Cognitive Sciences*, 21(4), 237–249.
- 1134 Senju, A., Southgate, V., Miura, Y., Matsui, T., Hasegawa, T., Tojo, Y., ... Csibra,
1135 G. (2010). Absence of spontaneous action anticipation by false belief attribution
1136 in children with autism spectrum disorder. *Development and Psychopathology*,
1137 22(2), 353–360.
- 1138 Senju, A., Southgate, V., Snape, C., Leonard, M., & Csibra, G. (2011). Do
1139 18-month-olds really attribute mental states to others? A critical test.
1140 *Psychological Science*, 22(7), 878–880.
- 1141 Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: An
1142 absence of spontaneous theory of mind in asperger syndrome. *Science*,
1143 325(5942), 883–885.
- 1144 Southgate, V., Johnson, M. H., Karoui, I. E., & Csibra, G. (2010). Motor system
1145 activation reveals infants' on-line prediction of others' goals. *Psychological*
1146 *Science*, 21(3), 355–359.
- 1147 Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through
1148 attribution of false belief by 2-year-olds. *Psychological Science*, 18(7), 587–592.
- 1149 Southgate, V., & Vernetti, A. (2014). Belief-based action prediction in preverbal
1150 infants. *Cognition*, 130(1), 1–10.
- 1151 Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old
1152 infants. *Psychological Science*, 18(7), 580–586.
- 1153 Surian, L., & Franchin, L. (2020). On the domain specificity of the mechanisms
1154 underpinning spontaneous anticipatory looks in false-belief tasks. *Developmental*

- 1155 *Science*, 23(6), e12955.
- 1156 Thoermer, C., Sodian, B., Vuori, M., Perst, H., & Kristen, S. (2012). Continuity
1157 from an implicit to an explicit understanding of false belief from infancy to
1158 preschool age. *British Journal of Developmental Psychology*, 30(1), 172–187.
- 1159 Träuble, B., Marinović, V., & Pauen, S. (2010). Early theory of mind competencies:
1160 Do infants understand others' beliefs? *Infancy*, 15(4), 434–444.
- 1161 Wellman, H. M., & Cross, D. (2001). Theory of mind and conceptual change. *Child
1162 Development*, 72(3), 702–707.
- 1163 Wiesmann, C. G., Friederici, A. D., Disla, D., Steinbeis, N., & Singer, T. (2018).
1164 Longitudinal evidence for 4-year-olds' but not 2-and 3-year-olds' false
1165 belief-related action anticipation. *Cognitive Development*, 46, 58–68.
- 1166 Woodward, A. L., & Sommerville, J. A. (2000). Twelve-month-old infants interpret
1167 action in context. *Psychological Science*, 11(1), 73–77.

Table 1

Lab and Participant information.

Lab	N collected	N included	Sex (N Female)	Mean Age (years)	Method
CogConcordia	21	16	11	22.12	In-lab
CorbitLab	16	15	14	19.87	In-lab
DevlabAU	20	20	15	25.15	In-lab
MEyeLab	53	53	39	24.47	In-lab
MiniDundee	15	13	10	30.23	In-lab
PKUSu	39	32	19	22.66	In-lab
SkidLSDLab	11	8	3	21.62	In-lab
ToMcdlSalzburg	33	31	22	27.23	In-lab
UIUCinfantlab	36	32	25	19.06	In-lab
WSUMARCS	18	13	8	29.85	In-lab
affcogUTSC	23	8	5	20.88	web-based
babyLeidenEdu	20	16	12	23.31	In-lab
babylabAmsterdam	17	16	13	24.00	In-lab
babylabBrookes	67	65	49	21.78	In-lab
babylabINCC	18	18	12	31.00	In-lab
babylabMPIB	16	16	11	27.44	In-lab
babylabNijmegen	19	15	13	22.13	In-lab
babylabTrento	16	16	9	21.69	In-lab
babylabUmassb	33	11	10	19.00	In-lab
babyuniHeidelberg	16	16	14	22.06	In-lab
beinghumanWroclaw	19	16	9	32.75	web-based
careylabHarvard	18	15	12	19.80	In-lab
cclUNIRI	32	32	17	30.53	In-lab
childdevlabAshoka	16	16	8	30.88	In-lab
collabUIOWA	16	16	10	19.19	In-lab
gaugGöttingen	30	28	18	31.71	In-lab
jmuCDL	32	32	22	18.81	In-lab
kidsdevUniofNewcastle	15	14	7	33.57	In-lab
labUNAM	20	11	8	22.45	In-lab

Table 2 continued

Lab	N collected	N included	Sex (N Female)	Mean Age (years)	Method
lmuMunich	31	30	23	22.53	In-lab
mecdmpihcbs	19	19	10	27.79	In-lab
socialcogUmiami	16	15	9	19.27	In-lab
sociocognitivelab	17	17	11	32.12	In-lab
tauccd	15	12	6	24.50	In-lab
Total	803	703	484	24.75	

Table 2

Lab and Participant information.

Lab	N collected	N included	Sex (N Female)	Mean Age (months)	Method
CogConcordia	21	8	4	22.92	web-based
CorbitLab	11	10	5	22.77	In-lab
DevlabAU	18	17	8	19.00	In-lab
PKUSu	50	32	13	20.84	In-lab
SkidLSDLab	8	2	0	20.11	In-lab
ToMcdlSalzburg	17	12	6	22.20	In-lab
UIUCinfantlab	18	15	9	21.96	In-lab
babyLeidenEdu	18	12	8	22.59	In-lab
babylabAmsterdam	28	12	6	23.19	In-lab
babylabBrookes	17	12	7	22.15	In-lab
babylabChicago	17	13	4	20.10	In-lab
babylabINCC	16	9	6	23.40	In-lab
babylabNijmegen	19	10	3	23.52	In-lab
babylabOxford	25	19	8	23.42	In-lab
babylabPrinceton	17	11	7	22.15	In-lab
babylabTrento	18	17	10	22.72	In-lab
babylabUmassb	7	6	2	20.35	In-lab
babylingOslo	17	14	7	21.99	In-lab
babyuniHeidelberg	16	12	4	22.69	In-lab
beinghumanWroclaw	24	14	7	23.77	web-based
careylabHarvard	17	12	5	21.99	In-lab
cecBYU	16	14	4	22.39	In-lab
childdevlabAshoka	16	10	6	22.44	In-lab
gaugGöttingen	28	15	9	23.06	In-lab
gertlabLancaster	21	17	8	23.03	In-lab
infantcogUBC	26	19	8	24.39	In-lab
irlConcordia	19	12	5	22.47	In-lab
kidsdevUniofNewcastle	16	14	9	22.36	In-lab
kokuhamburg	19	14	7	25.99	In-lab

Table 2 continued

Lab	N collected	N included	Sex (N Female)	Mean Age (months)	Method
labUNAM	18	12	7	22.68	In-lab
lmuMunich	48	24	16	22.68	In-lab
mecdmpihcbs	25	12	8	23.58	In-lab
mpievaCCP	22	18	10	23.33	In-lab
saxelab	31	15	2	23.13	web-based
socallabUCSD	47	15	4	22.09	web-based
tauccd	15	12	8	22.99	In-lab
unicph	43	29	16	21.50	In-lab
Total	809	521	256	22.48	

Table 3

Results of the Bayesian mixed effects models for the familiarization trials.

model	term	estimate	std.error	conf.low	conf.high
PTL toddlers	Intercept	0.12	0.02	0.09	0.15
PTL toddlers	Trial number	-0.05	0.01	-0.06	-0.03
PTL adults	Intercept	0.26	0.02	0.23	0.29
PTL adults	Trial number	0.10	0.01	0.09	0.11
First Look toddlers	Intercept	0.44	0.09	0.27	0.62
First Look toddlers	Trial number	-0.22	0.05	-0.32	-0.12
First Look adults	Intercept	1.03	0.09	0.86	1.20
First Look adults	Trial number	0.38	0.04	0.30	0.47

Table 4

Results of the Bayesian mixed effects models for the test trials.

model	term	estimate	std.error	conf.low	conf.high
PTL toddlers	Intercept	0.61	0.02	0.58	0.65
PTL toddlers	Condition	0.10	0.03	0.03	0.16
PTL toddlers	Age	0.01	0.01	-0.01	0.02
PTL toddlers	Condition x Age	-0.01	0.02	-0.04	0.02
PTL adults	Intercept	0.48	0.02	0.45	0.51
PTL adults	Condition	-0.20	0.03	-0.26	-0.15
First Look toddlers	Intercept	0.52	0.11	0.32	0.72
First Look toddlers	Condition	0.53	0.21	0.11	0.93
First Look toddlers	Age	0.06	0.04	-0.03	0.14
First Look toddlers	Condition x Age	-0.13	0.09	-0.31	0.05
First Look adults	Intercept	0.05	0.09	-0.13	0.22
First Look adults	Condition	-0.89	0.17	-1.21	-0.56