

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 example,

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

99 studies have shown that toddlers spontaneously adjust their behavior to the experimenter's

100 beliefs (D. Buttelmann, Carpenter, & Tomasello, 2009; Király, Oláh, Csibra, & Kovács,

101 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, Bayliss, Becker, &
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, Buttelmann, &
122 Kovács, 2020; Scott & Baillargeon, 2017). This growing body of literature has led to a
123 theoretical transformation of the field. In particular, findings with young infants have
124 paved the way for novel accounts of the development and cognitive foundations of ToM.

125 The previous consensus was that full-fledged ToM emerges only at around age 4,
126 potentially as the result of developing executive functions, complex language skills and
127 other factors (e.g., Perner, 1991; Wellman & Cross, 2001). In contrast, the newer accounts
128 proposed that some basic forms of ToM may be phylogenetically more ancient and may
129 develop much earlier in ontogeny (e.g., Baillargeon, Scott, & He, 2010; Carruthers, 2013;
130 Kovács, 2016; Leslie, 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 et al., 2012; Southgate et al., 2007;
148 Surian & Geraci, 2012; and Low & Watts, 2013), as well as studies using other paradigms,
149 and classified them as a successful, partial, or non-replication, depending on whether all,
150 some, or none of the original main effects were found. Although no formal analysis of effect
151 size was carried out, overall, non-replications and partial replications outnumbered
152 successful replications, regardless of the method used. In addition to the failure to replicate
153 spontaneous anticipation of agents' behaviors based on their beliefs, many of the
154 replication studies revealed an even more fundamental problem of spontaneous AL
155 procedures: a failure to adequately anticipate an agent's action in the absence of a belief.
156 That is, researchers did not find evidence for spontaneous anticipation of agents' behaviors
157 based on their goals, even in the initial familiarization trials of the experiments, where the
158 agent's beliefs do not play any role yet (e.g., Kampis et al., 2020; Kulke, Reiß, et al., 2018;

159 Schuwerk et al., 2018). The familiarization trials are designed to convey the goal of the
160 agent, as well as the general timing and structure of events, to set up participants'
161 expectations in the test trials where the agent's epistemic state is then manipulated.
162 Typically, the last familiarization trial can also be used to probe participants' spontaneous
163 action anticipation; and test trials can only be meaningfully interpreted if there is evidence
164 of above-chance anticipation in the familiarization trials. In several AL studies many
165 participants had to be excluded from the main analyses for failing to demonstrate robust
166 action anticipation during the familiarization trials (e.g., Kampis et al., 2020; Kulke, Reiß,
167 et al., 2018; Schuwerk et al., 2018; Southgate et al., 2007). This raises the possibility that
168 these paradigms may not be suitable for reliably eliciting spontaneous action prediction in
169 the first place (for discussion see Baillargeon et al., 2018). In sum, in light of the complex
170 and mixed state of the evidence, it currently remains unclear whether infants, toddlers, and
171 adults engage in spontaneous ToM. This calls for systematic, large-scale, *a priori* designed
172 multi-lab study that stringently tests for the robustness, reliability, and replicability of
173 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, Hertwig, & Kahneman, 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, MacLean, & Santos, 2020; Phillips et al., 2021)? 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, Call, & Tomasello, 2008; Martin & Santos, 2016; Perner, 1991; Phillips et al., 2021). 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, Everett, Croft, and Flavell (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, Wimmer, & Perner, 1986; Moll & Tomasello, 2006; O'Neill, 1996; F. Buttelmann & Kovács, 2019; F. Buttelmann, Suhrke, & Buttelmann, 2015; Kampis et al., 2020; though for some findings indicating Level II perspective-taking at an early age see Scott & Baillargeon, 2009; Scott, Richman, & Baillargeon, 2015), and that

238 non-human primates seem to master knowledge-ignorance tasks while not demonstrating
239 any more complex, meta-representational form of ToM (e.g., Hare, Call, & Tomasello, 2001;
240 Kaminski et al., 2008; Karg, Schmelz, Call, & Tomasello, 2015). The knowledge-ignorance
241 distinction thus appears to be an ideal candidate for assessing epistemic status-based
242 action anticipation in a wide range of populations. To date, however, no study has probed
243 whether or how children's (and adults') spontaneous action anticipation, as indicated by
244 AL, is sensitive to ascriptions of knowledge vs. ignorance. Most studies that have addressed
245 ToM with AL measures have targeted the more sophisticated true/false belief contrast. As
246 reviewed above, the results of those studies yield a mixed picture regarding replicability of
247 the findings. It has been argued that tasks that reliably replicate are ones which can be
248 solved with the more basic knowledge-ignorance distinction, whereas tasks that do not
249 replicate require more sophisticated belief-ascertainment (Powell et al., 2018)¹, suggesting that
250 only some but not all findings might not be replicable. Based on these considerations, the
251 present study tests whether toddlers and adults engage in knowledge- and ignorance-based
252 AL to probe the most basic form of spontaneous, epistemic state-based action anticipation.

253 **Design and Predictions of the Present Study**

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

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, Wenzel, Proft, Rakoczy, & Liszkowski, 2019; Kulke, Reiß, et al., 2018; Powell et al., 2018). Such a pattern remains ambiguous since it may merely reflect a knowledge-ignorance distinction.

approaching a target (familiarization trials). AL is measured during familiarization trials to probe whether participants understood the agent's goal and spontaneously anticipate their actions. Subsequently, during test trials the agent's visual access is manipulated, leading them to be either *knowledgeable* or *ignorant* about the location of the target.

Participants' AL will be measured during test trials to determine whether or not they take into account the agent's epistemic access and adjust their action anticipation accordingly.

Participants' looking patterns will be recorded using either lab-based corneal reflection eye-tracking or online recording of gaze patterns. We chose to provide the online testing option to increase the flexibility for data collection given the disruption caused by the Covid-19 pandemic. This option will also provide the opportunity to potentially compare in-lab and online testing procedures (Sheskin et al., 2020). Novel animated stimuli were collectively developed within the MB2 consortium on the basis of previous work (e.g., Clements & Perner, 1994) and based on input from collaborators with experience with both successful and failed replication studies (e.g., Grosse Wiesmann et al., 2017; Surian & Geraci, 2012). These animated 3D scenes feature a dynamic interaction aimed to optimally engage participants' attention: a chasing scenario involving two agents, a *chaser* and a *chasee* (see Figures 1 and 2). As part of the chase, the chasee enters from the top of an upside-down Y-shaped tunnel with two boxes at its exits. The tunnel is opaque so participants cannot see the chasee after it enters the tunnel, but can hear noises that indicate movement. The chasee eventually exits from one of the arms of the Y, and goes into the box on that side. The chaser observes the chasee exit the tunnel and go into a box, and then follows it through the tunnel. During familiarization trials, the chaser always exits the tunnel on the same side as the chasee, and approaches the box where the chasee is currently located. Thus, if participants engage in spontaneous action anticipation during familiarization trials, they should reliably anticipate during the period when the chaser is in the tunnel that it will emerge at the exit that leads to the box containing the chasee.

During test trials, the chasee always first hides in one of the boxes but shortly thereafter

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

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).

304 period. In two pilot studies (see Methods section), we addressed the foundational question
305 of the current study: whether these stimuli reveal spontaneous goal-directed action
306 anticipation as measured by AL in the above-described familiarization trials (i.e., without a
307 change of location by the chasee or manipulation of the chaser's epistemic state). We found
308 that our paradigm indeed elicited action anticipation and exclusion rates due to lack of
309 anticipation were significantly lower relative to previous (original and replication) AL
310 studies. Both toddlers and adults showed reliable anticipation of the chaser's exit at the
311 chasee's location, indicating that in contrast with many previous AL studies the current
312 paradigm successfully elicits spontaneous goal-based action anticipation. Based on these
313 pilot data we concluded that the paradigm is suitable for examining the second and critical
314 question: whether toddlers and adults, in their spontaneous goal-based action anticipation,
315 take into account the agent's epistemic state. We predict that if participants track the
316 chaser's perceptual access and resulting epistemic state (knowledge/ignorance) and
317 anticipate their actions accordingly, they should look more in anticipation to the exit at the
318 chasee's location than the other exit in the *knowledge* condition, but should not do so (or
319 to a lesser degree; see below) in the *ignorance* condition. We anticipate three potential
320 factors that could influence participant's gaze patterns: Keeping track of the chaser's
321 epistemic status in the *ignorance* condition might either lead to no expectations as to
322 where the chaser will look (resulting in chance level looking between the two exits) or (if
323 participants follow an "ignorance leads to mistakes"-rule, see e.g., Ruffman, 1996) to an
324 expectation that the chaser will go to the wrong location [longer looking to the exit with
325 the empty box; e.g., Fabricius, Boyer, Weimer, and Carroll (2010)]. Either way,
326 participants may still show a 'pull of the real' even in the *ignorance* condition, i.e., reveal a
327 default tendency to look to the side where the chasee is located. But if they truly keep
328 track of the epistemic status of the chaser (*knowledge* vs. *ignorance*), they should show this
329 tendency to look to the side where the chasee really is in the *ignorance* condition to a lesser
330 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 based on their goals in this paradigm, as they did in our pilot study? Second, can we find evidence that they take into account the agent's epistemic access (knowledge vs. ignorance) and adjust their action anticipation accordingly? In addressing these questions, the present study will significantly contribute to our knowledge on spontaneous ToM. It will inform us whether the present paradigm and stimuli can elicit spontaneous goal-based and mental-state-based action anticipation in adults and toddlers, based on a large sample of about 800 participants in total from over 20 labs. In the long run, the present study will lay the foundation for future work to address broader questions of what *kind* of epistemic states toddlers and adults spontaneously attribute to others in their action anticipation and what cognitive mechanisms allow them to do so.

342

Methods

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

349

Stimuli

Figures 1 and 2 provide an overview of the paradigm. For the stimuli, 3D animations were created depicting a chasing scenario between two agents (chaser and chasee) who start in the upper part of the scene. At the very top of the scene a door leads to outside the visible scene. Below this area, a horizontal fence separates the space, and thus the lower part of the space can be reached by the Y-shaped tunnel only. Additional information on

355 the general scene setup, events, and timings in the familiarization and the test trials, as
 356 well as trial randomization can be found in the Supplemental Material.

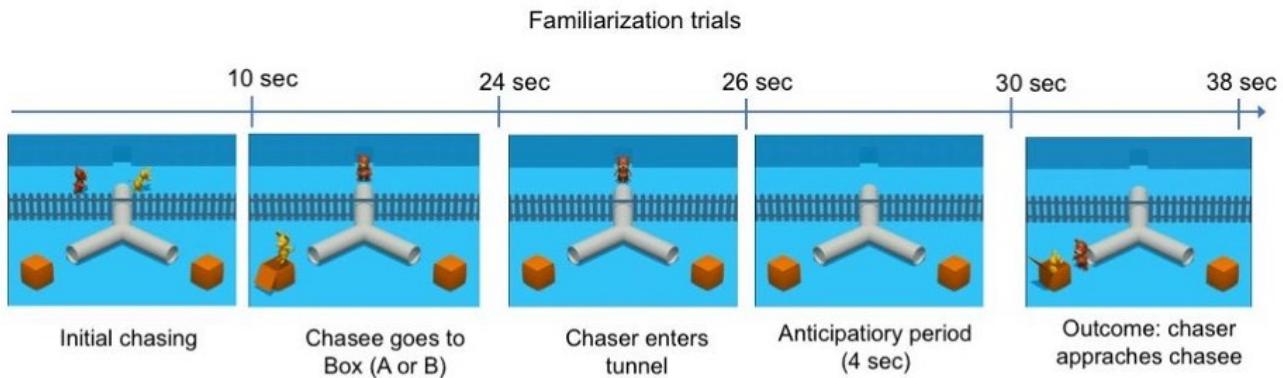


Figure 1. Timeline of the familiarization trials.

357 **Familiarization Trials.** All participants will view four familiarization trials (for an
 358 overview of key events see Figure 1). During familiarization trials, after a brief chasing
 359 introduction, the chasee enters an upside-down Y-shaped tunnel with a box at both of its
 360 exits. The chasee then leaves the tunnel through one of the exits and hides in the box on
 361 the corresponding side. Subsequently, the chaser enters the tunnel (to follow the chasee),
 362 and participants' AL to the tunnel exits is measured before the chaser exits on the side the
 363 chasee is hiding, as an index of their goal-based action anticipation. In these familiarization
 364 trials, if participants engage in spontaneous action anticipation, they should reliably
 365 anticipate that the chaser should emerge at the tunnel exit that leads to the box where the
 366 chasee is. After leaving the tunnel, the chaser approaches the box in which the chasee is
 367 hiding and knocks on it. Then, the chasee jumps out of the box and the two briefly interact.

368 **Familiarization Phase Pilot Studies.** In a pilot study with 18- to
 369 27-month-olds ($n = 65$) and adults ($n = 42$), seven labs used in-lab corneal reflection
 370 eye-tracking to collect data on gaze behavior in the familiarization phase. A key
 371 desideratum of our paradigm is that it should produce sufficient AL, as a low rate of AL in
 372 previous studies has led to high exclusion rates. The goals of the pilot study were to 1)
 373 estimate the level of correct goal-based action predictions in the familiarization phase, 2)

374 determine the optimal number of familiarization trials, 3) check for issues with perceptual
375 properties of stimuli (e.g., distracting visual saliences), and 4) test the general procedure
376 including preprocessing and analyzing raw gaze data from different eye-tracking systems.
377 We found that the familiarization stimuli elicited a relatively high proportion of
378 goal-directed action anticipations, but we were concerned about the effects of some minor
379 properties of the stimulus (in particular, a small rectangular window in the tunnel tube
380 that allowed participants to see the agents at one point on their path to the tunnel exits).
381 In a second pilot study with 18- to 27-month-olds ($n = 12$, three participating labs), slight
382 changes of stimulus features (the removal of the window in the tube; temporal changes of
383 auditory anticipation cue) did not cause major changes in the AL rates. Sixty-eight percent
384 of toddlers' first looks in the first pilot, 69% of toddlers' first looks in the second pilot, and
385 69% of adults' first looks were toward the correct area of interest (AOI) during the
386 anticipatory period. The average proportion of looking towards the correct AOI during the
387 anticipatory period was 70.7% ($CI_{95\%} = 67.6\% - 73.8\%$) in toddlers in the first pilot, 70.5%
388 ($CI_{95\%} = 62.8\% - 78.2\%$) in the second pilot for toddlers, and 75.3% ($CI_{95\%} = 71.0\% -$
389 79.5%) in adults. In Bayesian analyses, we found strong evidence that toddlers and adults
390 looked more towards the target than towards the distractor during the anticipation period.
391 Based on conceptual and practical methodological considerations while also considering
392 previous studies, we decided to include four trials in the final experiment. The pilot data
393 results of the toddlers supported this decision insofar as we observed a looking bias towards
394 the correct location already in trials 1-4, without additional benefit of trials 5-8. Further,
395 prototypical analysis pipelines were established for combining raw gaze data from different
396 eye-trackers. In short, we developed a way to resample gaze data from different
397 eye-trackers to be at a common Hz rate and to define proportionally correct AOIs for
398 different screen dimensions with the goal to merge all raw data into one data set for
399 inferential statistics. The established analysis procedure is described further in the Data
400 Preprocessing section below. In sum, we concluded that this paradigm sufficiently elicits

401 goal-directed action predictions, an important prerequisite for drawing any conclusion on
402 AL behavior in the test trials of this study. A detailed description of the two pilot studies
403 can be found in the Supplemental Material.

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

428 case of the chaser's absence in the *knowledge* condition, there were two main options, both
429 with inevitable trade-offs. First, we could have increased the duration of the chaser's
430 absence in the *knowledge* condition to match the duration of the chaser's absence in both
431 conditions. Yet, this would potentially disrupt the flow of events, such as keeping track of
432 the chasee's actions and the general scene dynamics, since nothing would happen for a
433 substantial amount of time. Second, the chaser can be absent for a shorter time in the
434 *knowledge* than in the *ignorance* condition, in which case the flow of events – the chasee's
435 actions and the general scene dynamics – remains natural. We chose the second option
436 because we reasoned that the artificial break in the *knowledge* condition could disrupt the
437 participant's tracking of the chaser's epistemic state, thus being a confound that would be
438 more detrimental than the difference in the duration of absence. Further, the current
439 contrast has the advantage that the chasee's sequence and timing of actions are identical in
440 both conditions, thus minimizing the difference between conditions. Finally, with the
441 current design, the duration of the chaser's absence will be closely matched in the later
442 planned false belief - true belief contrast, because in the future false belief condition, the
443 chaser has to be absent for fewer events (because the chaser witnesses the first hiding
444 events after the chasee reappeared at the other side of the tunnel).

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

450 **Lab Participation Details**

451 **Time-Frame.** The contributing labs will start data collection as soon as they are
452 able to once our Registered Report receives an in-principle acceptance. The study will be
453 submitted for Stage 2 review within one year after in-principle acceptance (i.e., post-Stage

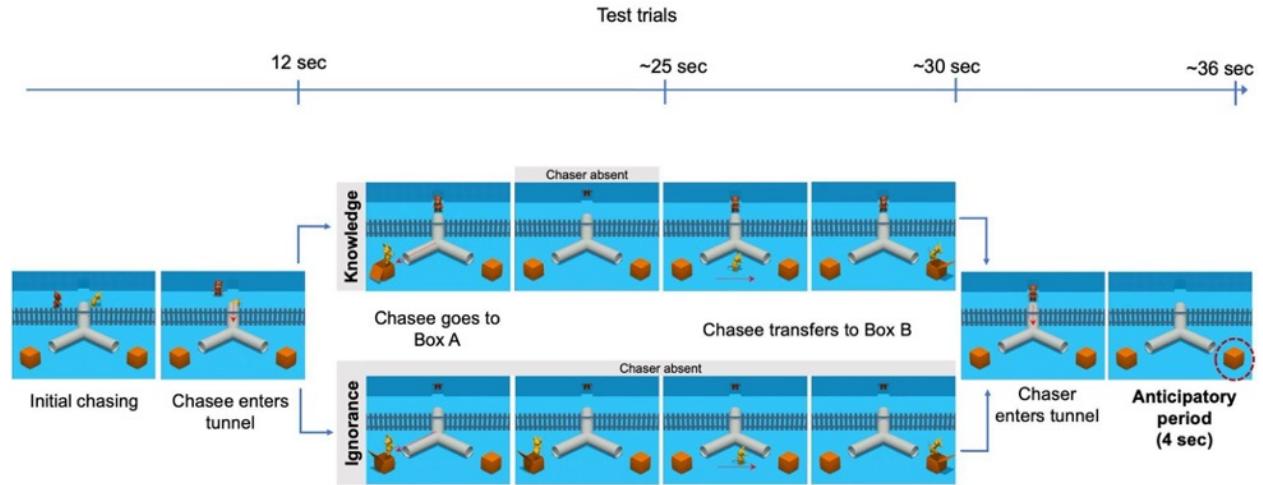


Figure 2. Schematic overview of stimuli and conditions of the test trials.

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.

454 1 review). We anticipate that this time window gives the individual labs enough flexibility
455 to contribute the committed sample sizes; however, if this timeline needs adjusting due to
456 the Covid-19 pandemic this decision will be made prior to any data analysis.

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

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

471 **Participants.** In a preliminary expression of interest, 26 labs signed up to
472 contribute a minimal sample size of 16 toddlers and/or adults. Based on this information,
473 we expect to recruit a total sample of 520 toddlers (ages 18-27 months) and 408 adults
474 (ages 18-55 years). To avoid an unbalanced age distribution in the toddlers sample, labs
475 will sign up for testing at least one of two age bins (bin 1: 18-22 months, bin 2: 23-27
476 months), and will be asked to ensure approximately equal distribution of participants' age
477 in their collected sample if possible. They will be asked to try to ensure that the mean age
478 of their sample lies in the middle of the range of the chosen bin and that participant ages
479 are distributed across their whole bin. Both for adults and toddlers, basic demographic
480 data will be collected on a voluntary basis with a brief questionnaire (see Supplemental

481 Material for details). The requested demographic information that is not used in the
482 registered confirmatory and/or exploratory analyses of this study will be collected for
483 further potential follow-up analyses in spin-off projects within the MB framework. After
484 completing the task, adult participants will be asked to fill a funneled debriefing
485 questionnaire. This questionnaire asks what the participant thinks the purpose of the
486 experiment was, whether the participant had any particular goal or strategy while watching
487 the videos, and whether the participant consciously tracked the chaser's epistemic state.
488 Additionally, we collect details regarding each testing session (see Supplemental Material).

489

490

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

499 Apparatus and Procedure

500 **Eye-tracking Methods.** We expect that participating labs will use one of three
501 types of eye-tracker brands to track the participant's gaze patterns: Tobii, EyeLink, or
502 SMI. Thus, apparatus setup will slightly vary in individual labs (e.g., different sampling
503 rates and distances at which the participants are seated in front of the monitor).

504 Participating labs will report their eye-tracker specifications and study procedure alongside

505 the collected data. To minimize variation between labs, all labs using the same type of
506 eye-tracker will use the same presentation study file specific to that eye-tracker type. The
507 Supplemental Material will provide an overview of employed eye-trackers, stimulus
508 presentation softwares, sampling rates and screen dimensions.

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

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

528 **General Lab Practices**

529 To ensure standardization of procedure, materials for testing practices and
530 instructions will be prepared and distributed to the participating labs. Each lab will be

531 responsible for maintaining these practices and report all relevant details on testing
532 sessions (for details see the Supplemental Material).

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

540 Design Analysis

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

557 we did not determine our sample size based on power analysis, here we provide a
558 simulation-based design analysis to demonstrate the range of BFs we might expect to see,
559 given a plausible range of effect sizes and parameters. We focus this analysis on our key
560 analysis of the test trials (as specified below), namely the difference in AL on the first test
561 trial that participants saw. We describe below the simulation for the child sample, but
562 based on our specifications, we expect that a design analysis for adult data would produce
563 similar results. We first ran a simulation for the first look analysis. In each iteration of our
564 simulation, we used a set of parameters to simulate an experiment, using a first look
565 (described below) as the key measure. For the key effect size parameter for condition
566 (*knowledge* vs. *ignorance*), we sampled a range of effect sizes in logit space spanning from
567 small to large effects (Cohen's $d = 0.20 - 0.80$; log odds from 0.36 - 1.45). For each
568 experiment, the betas for age and the age x condition interaction were sampled uniformly
569 between -0.20 and 0.20. The age of each participant was sampled uniformly between 18
570 and 27 months and then centered. The intercept was sampled from a normal distribution
571 (1, 0.25), corresponding to an average looking proportion of 0.73. Lab intercepts and the
572 lab slope by condition were set to 0.1, and other lab random effects were set to 0 as we do
573 not expect them to be meaningfully non-zero. These values were chosen based on pilot
574 data (average looking proportion), but also to have a large range of possible outcomes (lab
575 intercept, age and age x condition interaction). We are confident that the results would be
576 robust to different choices. We then used these simulated data to simulate an experiment
577 with 22 labs and 440 toddlers and computed the resulting BFs, as specified in the analysis
578 plan below. We adopted all of the priors specified in the results section below³. We ran 349
579 simulations and, in 72% of them, the BF showed strong evidence in favor of the full model

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.

580 (BF > 10); in 6% the BF showed substantial evidence (10 > BF > 3); it was inconclusive
581 14% of the time (1/10 > BF > 3), and in 8% of cases the null model was substantially
582 favored (see Figure 3). In none of the simulations the BF was < 1/10. Thus, under the
583 parameters chosen here for our simulations, it is likely that the planned experiment is of
584 sufficient size to detect the expected effect. We also ran a design analysis for the
585 proportional looking analysis. We used the same experimental parameters (number of labs,
586 participants, ages, etc.). For generating simulated data, we drew the condition effect from
587 a uniform distribution between .05 and .20 (in proportion space). The age and
588 age:condition effects were drawn from uniform distributions between -.05 and .05. Sigma,
589 the overall noise in the experiment, was drawn from a uniform distribution between .05 and
590 .1. The intercept was drawn from a normal distribution with mean .65 and a standard
591 deviation of .05. The by-lab standard deviation for the intercept and condition slope was
592 set to .01. Priors were as described in the main text. We ran 119 simulations, and in all
593 119 we obtained a BF greater than 10, suggesting that, under our assumptions, the study
594 is well-powered.

595 Data Preprocessing

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

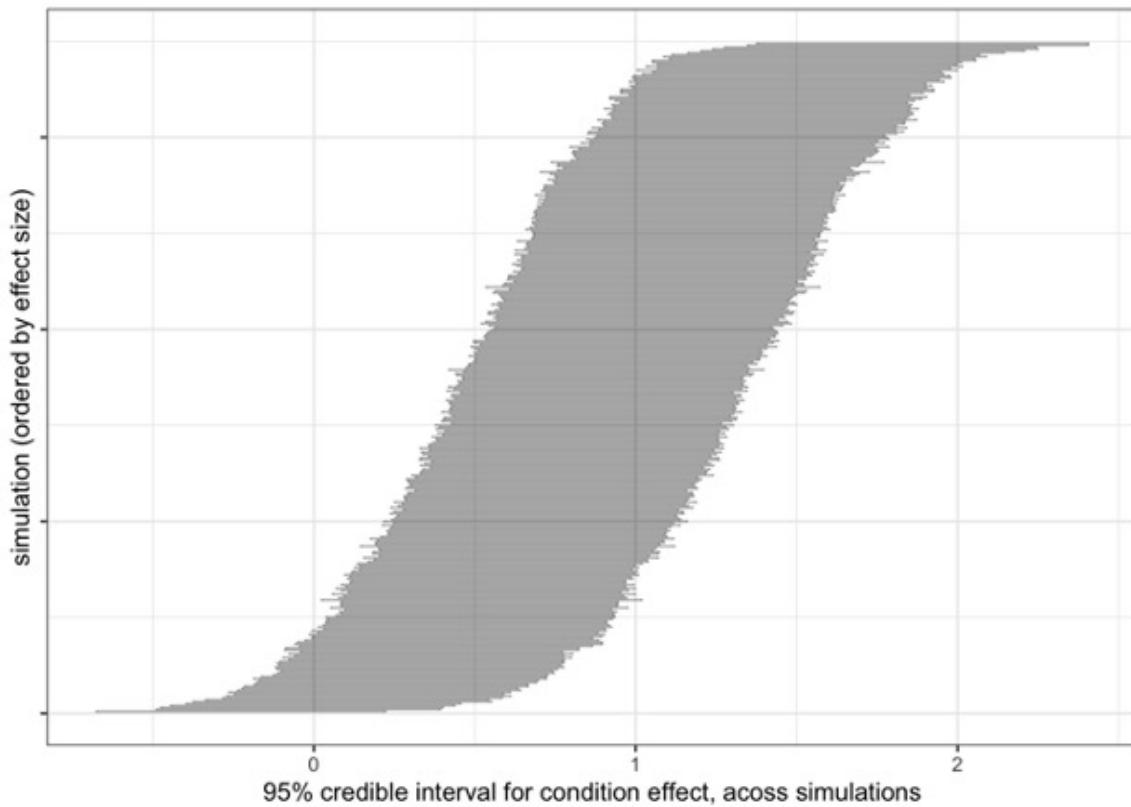


Figure 3. Effect sizes of simulated experiments.

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.

period). Applying this criterion would have caused us to exclude 4% of the trials in our pilot data, which inspection of our pilot data suggested was an appropriate trade-off between not excluding too much usable data and not analyzing trials which were uninformative. For each monitor size, we will determine the specific AOIs and compute whether the specific x- and y-position for each participant, trial, and time point fall within their screen resolution-specific AOIs. Our goal is to determine whether participants are anticipating the emergence of the chaser from one of the two tunnel exits. Thus, we defined AOIs on the stimulus by creating a rectangular region around the tunnel exit that is D units from the top, bottom, left, and right of the boundary of the tunnel exit, where D is the diameter of the tunnel exits. We then expanded the sides of the AOI rectangles by 25%

in all directions to account for tracker calibration error. Our rationale was that, if we made the AOI too small, we might fail to capture anticipations by participants with poor calibrations. In contrast, if we made the regions too large, we might capture some fixations by participants looking at the box where the chasee actually is. On the other hand, these chasee looks would not be expected to vary between conditions and so would only affect our baseline level of looking. Thus, the chosen AOIs aim at maximizing our ability to capture between-condition differences. For an illustration of the tunnel exit AOIs see Figure 4. We are not analyzing looks to the boxes, since they can less unambiguously be interpreted as epistemic state-based action predictions and because we observed few anticipatory looks to the boxes in the pilot studies. For more detailed information about the AOI definition process see the description of the pilot study results in the Supplemental Material.

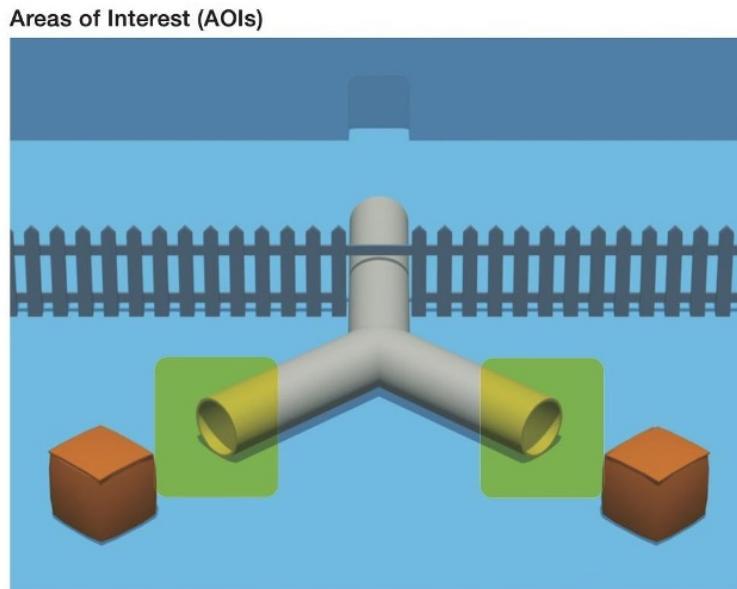


Figure 4. Illustration of Areas of Interest (AOIs) for gaze data analysis during the anticipatory period.

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

627 **Manual Coding.** For data gathered without an eye-tracker (e.g., videos of
628 participants gathered from online administration), precise estimation of looks to specific
629 AOIs will not be possible. Instead, videos will be coded for whether participants are looking
630 to the left or the right side of the screen (or “other/off screen”). In our main analysis,
631 during the critical anticipatory window, we will treat these looks identically to looks to the
632 corresponding AOI. See exploratory analyses for analysis of data collected online.

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

640 **Dependent Variables.** We define two primary dependent variables: 1. First look.
641 First saccades will be determined as the first change in gaze occurring within the
642 anticipatory time window that is directed towards one of the AOIs. The first look is then
643 the binary variable denoting the target of this first saccade (i.e., either the correct or
644 incorrect AOI) and is defined as the first AOI where participants fixated at for at least 150
645 ms, as in rayner2009eye. The rationale for this definition was that, if participants are
646 looking at a location within the tunnel exit AOIs before the anticipation period, they
647 might have been looking there for other reasons than action prediction. We therefore count
648 only looks that start within the anticipation period because they more unambiguously
649 reflect action predictions. This further prevents us from running into a situation where we
650 would include a lot of fixations on regions other than the tunnel exit AOIs because
651 participants are looking somewhere else before the anticipation period begins. 2.
652 Proportion DLS [also referred to as total relative looking time; Senju et al. (2009)]. We
653 compute the proportion looking (p) to the correct AOI during the full 4000 ms anticipatory

654 window (correct looking time / (correct looking time + incorrect looking time)), excluding
655 looks outside of either AOI.

656

Results

657 Confirmatory Analyses

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

of this type (Bürkner, 2017). Because the BF is sensitive to the choice of prior, we also ran a secondary analysis with a less informative prior: fixed effect coefficients chosen from a normal distribution with mean 0 and SD of 3, and random effect standard deviations drawn from a normal prior with a mean of 0 and SD of 0.5. With respect to the specification of random effects, we followed the approach advocated by Barr (2013), that is, specifying the maximal random effect structure justified by our design. Since we are interested in lab-level variation, we will fit random effect coefficients for fixed effects of interest within labs (e.g., condition within lab). Further, where there were participant-level repeated measure data (e.g., familiarization trials), we fitted random effects of participants. For the proportional looking score analysis, we used a uniform prior on the intercept between -0.5 and 0.5 (corresponding to proportional looking scores between 0 and 1: the full possible range). For the priors on the fixed effect coefficients, we used a normal prior with a mean of 0 and an SD of 0.1. Because these regressions are in proportion space, 0.10 corresponds to a change in proportion of 10%. For the random effect priors, we used a normal distribution with mean 0 and standard deviation .05. The LKJ prior was specified as above.

694

695

Familiarization Trials. Figure 5 shows the proportion of total relative looking time (non-logit transformed) and proportion of first looks for toddlers and adults plotted across familiarization trials and test trials. Our first set of analyses examined data from the four familiarization trials and asked whether participants anticipated the chaser's reappearance at one of the tunnel exits. In our first analysis, we were interested in whether participants engage in AL during the familiarization trials. To quantify the level of familiarization, we fitted Bayesian mixed effect models predicting target looks based on trial number (1-4) with random effects for lab and participants and random slopes for trial

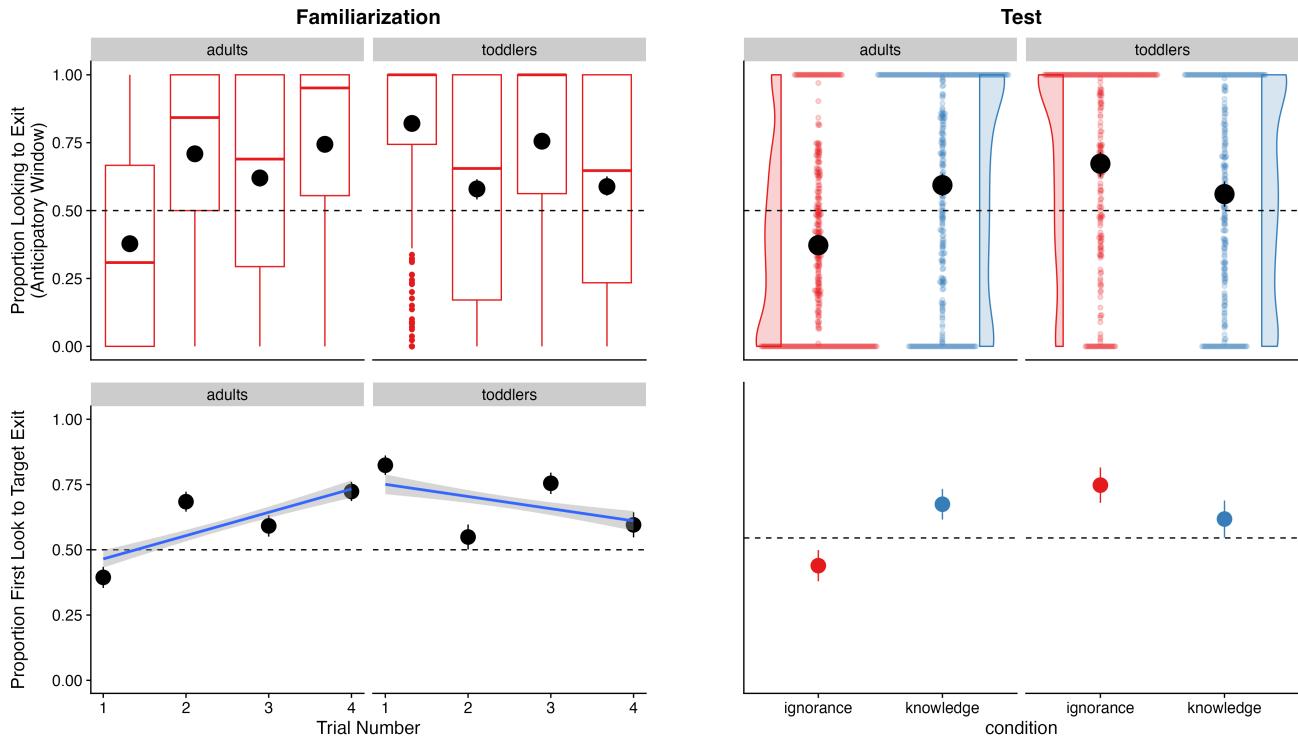


Figure 5. Proportional target looking and proportion of first looks for toddlers and adults during familiarization and test.

number for each. In R formula notation (which we adopt here because of its relative concision compared with standard mathematical notation), our base model was as follows: $measure \sim 1 + trial_number + (trial_number|lab) + (trial_number|participant)$ We fitted a total of four instances of this model, one for each age group (toddlers vs. adults) and dependent measure (proportion looking score vs. first look). First look models were fitted using a logistic link function. The proportion looking score models were Gaussian. Our key question of interest was whether overall anticipation is higher than chance levels on the familiarization trial immediately before the test trials, in service of evaluating the evidence that participants are attentive and making predictive looks immediately prior to test. To evaluate this question across the four models, we coded trial number so that the last trial before the test trials (trial 4) was set to the intercept, allowing the model intercept to encode an estimate of the proportion of correct anticipation immediately

716 before test. We then fitted a simpler model for comparison
717 $measure \sim 0 + trial_number + (trial_number|lab) + (trial_number|participant)$, which
718 included no intercept term. We then computed the BF comparing this model to the full
719 model. This BF quantified the evidence for an anticipation effect for each group and
720 measure.

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

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

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

735 ***Proportion of first looks.***

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

742 looks, with the negative coefficient indicating a decrease in target looks across the
 743 familiarization trials.

744 *Adults.* Comparing the Bayesian mixed effects model of adults predicting proportion
 745 of first looks based on trial number (1-4), with random effects for lab and participants and
 746 random slopes for trial number for each with the simpler model without an intercept, we
 747 computed a Bayes factor of $BF > 1000$, strongly favoring the full model over the null
 748 model. There was again support for an effect of trial number on proportion of first looks,
 749 with the positive coefficient indicating an increase in proportion of first target looks across
 750 the familiarization trials.

751 **Test Trials.** We focused our confirmatory analysis on the first test trial (see
 752 Exploratory Analysis section for an analysis of both trials). Our primary question of
 753 interest was whether AL differs between conditions (knowledge vs. ignorance, coded as
 754 $-.5/.5$) and by age (in months, centered). For child participants, we fitted models with the
 755 specification:

756 $measure\ 1 + condition + age + condition : age + (1 + condition + age + condition : age | lab)$.

757 For adult participants, we fitted models with the specification

758 $measure\ 1 + condition + (1 + condition | lab)$. Again, we fitted models with a logistic link
 759 for first look analyses and with a standard linear link for DLS. In each case, our key BF
 760 was a comparison of this model with a simpler “null” model that did not include the fixed
 761 effect of condition but still included other terms. We take a $BF > 3$ in favor of a particular
 762 model as substantial evidence and a $BF > 10$ in favor of strong evidence. A $BF < 1/3$ is
 763 taken as substantial evidence in favor of the simpler model, and a $BF < 1/10$ as strong
 764 evidence in favor of the simpler model. For the model of data from toddlers, we
 765 additionally were interested in whether the model shows changes in AL with age. We
 766 assessed evidence for this by computing BFs related to the comparison with a model that
 767 did not include an interaction between age and condition as fixed effects

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

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

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

787 *Toddlers.* As first model, we used a Bayesian mixed effects models to predict
788 toddlers' PTL based on condition, age, and the interaction of condition and age, while
789 accounting for variability across labs. The Bayes factor comparing this model to the simpler
790 null model without the interaction of condition was estimated to be BF = 22.4, favoring
791 the full model over the null model. Table 4 shows the statistics for regression coefficients of
792 the full model. These results suggest a significant effect of condition on PTL, with the
793 positive coefficient indicating higher PTL for ignorance trials compared to knowledge trials.

794 Additionally, we assessed whether toddlers' AL changed with age. Comparing our full

795 model, which included an interaction between age and condition, with a simpler model
796 without this interaction yielded a Bayes factor, $BF = 0.4$, providing modest support for the
797 simpler model. This result suggests that the interaction between age and condition might
798 not be a necessary predictor, as it doesn't provide substantial additional explanatory
799 power. This implies that age-related changes in AL are likely consistent across conditions,
800 rather than differing between them.

801 *Adults.* Next, we used a Bayesian mixed effects model to predict PTL based on
802 condition for adults, again with random effects for lab. The Bayes factor comparing this
803 model to the simpler null model without the main effect of condition was estimated to be
804 $BF > 1000$, strongly favoring the full model over the null model. These results suggest a
805 significant main effect of condition on PTL, with the negative coefficient indicating a
806 higher number of target looks for knowledge than for ignorance trials.

807 ***Proportion of first looks.***

808 *Toddlers.* Investigating proportion of first looks for toddlers, we again used a
809 Bayesian mixed effects model to predict target looks based on condition, with random
810 effects for lab. The Bayes factor comparing the full model to the simpler model was
811 estimated to be $BF = 2.5$, providing no substantial evidence in favor of the full model over
812 the null model.

813 Again, we examined whether age influenced the difference in action anticipation
814 between knowledge and ignorance trials. To do this, we compared the full model, which
815 included an interaction between age and condition, with a simpler model without this
816 interaction. The computed Bayes factor, $BF = 0.0$, strongly supports the simpler model,
817 suggesting that the interaction term does not substantially improve the model's fit. This
818 implies that age does not appear to significantly affect the difference in action anticipation
819 between the two trial types.

820 *Adults.* We compared a Bayesian mixed-effects model predicting the proportion of
821 first looks based on condition, including random effects for lab to a simpler model without
822 the main effect of condition. The analysis yielded a Bayes factor of $BF > 1000$, providing
823 strong evidence in favor of the full model over the null model. Results indicated that first
824 looks to the target were significantly more frequent in the knowledge condition compared
825 to the ignorance condition.

826 **Exploratory Analyses**

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

- 830 1. Spill-over: we will analyze within-participants data from the second test trial that
831 participants saw, using exploratory models to assess whether (1) findings are
832 consistent when both trials are included (overall condition effect), (2) whether effects
833 are magnified or diminished on the second trial (order main effect), and (3) whether
834 there is evidence of “spillover” - dependency in anticipation on the second trial
835 depending on what the first trial is (condition x order interaction effect).
- 836 2. We will explore whether condition differences vary for participants who show higher
837 rates of anticipation during the four familiarization trials. For example, we might
838 group participants according to whether they did or did not show correct AL at the
839 end of the familiarization phase, defined as overall longer looking at the correct AOI
840 than the incorrect AOI on average in trials 3 and 4 of the familiarization phase.
- 841 3. In analyses introducing model terms for certain measurement characteristics (e.g.,
842 types of eye-tracker manufacturers, screen dimensions), we will quantify potential
843 variability between different in-lab data acquisition methods (cf., ManyBabies

Consortium, 2020). If we have a sufficiently large sample of participants tested with online sources (e.g., contributions of at least 32 participants), we will conduct a separate analysis with a model term for online participants that estimates whether condition effects are different in this population. We will further report whether exclusion rates are different for this population.

4. If we observe substantial looking (defined *post hoc* by evaluating scatter plot videos of gaze data) to the boxes as well as the tunnel exit AOIs, we will conduct an exploratory analysis using tighter AOIs around tunnel exits and boxes, asking whether box and tunnel looking vary separately by age or by condition. In particular, we expect that the difference in AL between the two conditions will be bigger for the tunnel exits than for the box (as looks to the correct box might indicate looks to the target, which is in the same box for both conditions, rather than action anticipation).

5. To examine whether participants monitor both the bear and the mouse during the mouse's location change, and how this may influence AL in the test phase, we define new time windows of interest (TOIs) corresponding to the mouse's location change in each condition and areas of interest (AOIs) for both the mouse and bear. We hypothesize that participants who attend to both AOIs will exhibit greater AL compared to those who predominantly track the mouse during its location change. Specifically, we will analyze the frequency of gaze shifts between the mouse and bear, as well as the duration of gaze directed toward each AOI during the mouse's location change.

Spill-over. Analyzing condition-effects of within-participants data for both test trials, we fitted a Bayesian mixed-effects model with the dependent variable of PTL and main effects of condition and age and their interaction for toddlers. Comparing this full model to a null model that did not include the fixed effect of condition, we obtained a Bayes Factor of $BF = 0.1$, providing strong evidence in favor of the null model.

870 For adults, we also fitted a Bayesian mixed-effects model to predict their PTL for
871 both test trials with the main effect of condition and random effects for participant and
872 lab. Again, the data provided very strong evidence for the inclusion of the main effect of
873 condition with a Bayes Factor of $BF > 1000$. The effect of condition was negative and
874 credible, suggesting that PTL was significantly lower in the ignorance condition compared
875 to the knowledge condition.

876 In order to investigate whether there's an interaction of condition and test trial
877 number, we fitted Bayesian mixed-effects model to predict PTL with fixed effects for
878 condition, test trial number, and their interaction, along with random intercepts and slopes
879 for these variables across labs, for toddlers and adults separately. For toddlers, the Bayes
880 factor, $BF = 0.4$, modestly favored the simpler null model without the interaction term,
881 indicating that the interaction between condition and test trial number does not add
882 substantial explanatory power to the model. These results suggest that neither condition
883 nor its interaction with test trial number significantly impacts PTL in this sample. For
884 adults, the Bayes Factor, $BF = 21.7$, provided strong evidence for including the interaction
885 of condition and test trial number as fixed effect. These results indicate that while PTL
886 increased over trials, this effect was moderated by condition, with the ignorance condition
887 showing a slower rate of increase compared to the knowledge condition.

888 **Relationship between familiarization and test.** To investigate whether
889 participants who show anticipatory looking during the familiarization phase also exhibit
890 anticipatory looking during the test phase, we explored three different measures. First, we
891 assessed anticipatory looking in participants who successfully anticipated during the final
892 familiarization trial, defined as those whose first fixation was on the target. Second, we
893 examined anticipatory looking in participants who consistently demonstrated anticipatory
894 behavior across all familiarization trials, operationalized as having a PTL greater than 0.5
895 in each trial. Finally, we computed correlations to explore whether performance in the
896 familiarization phase was related to performance in the test trials.

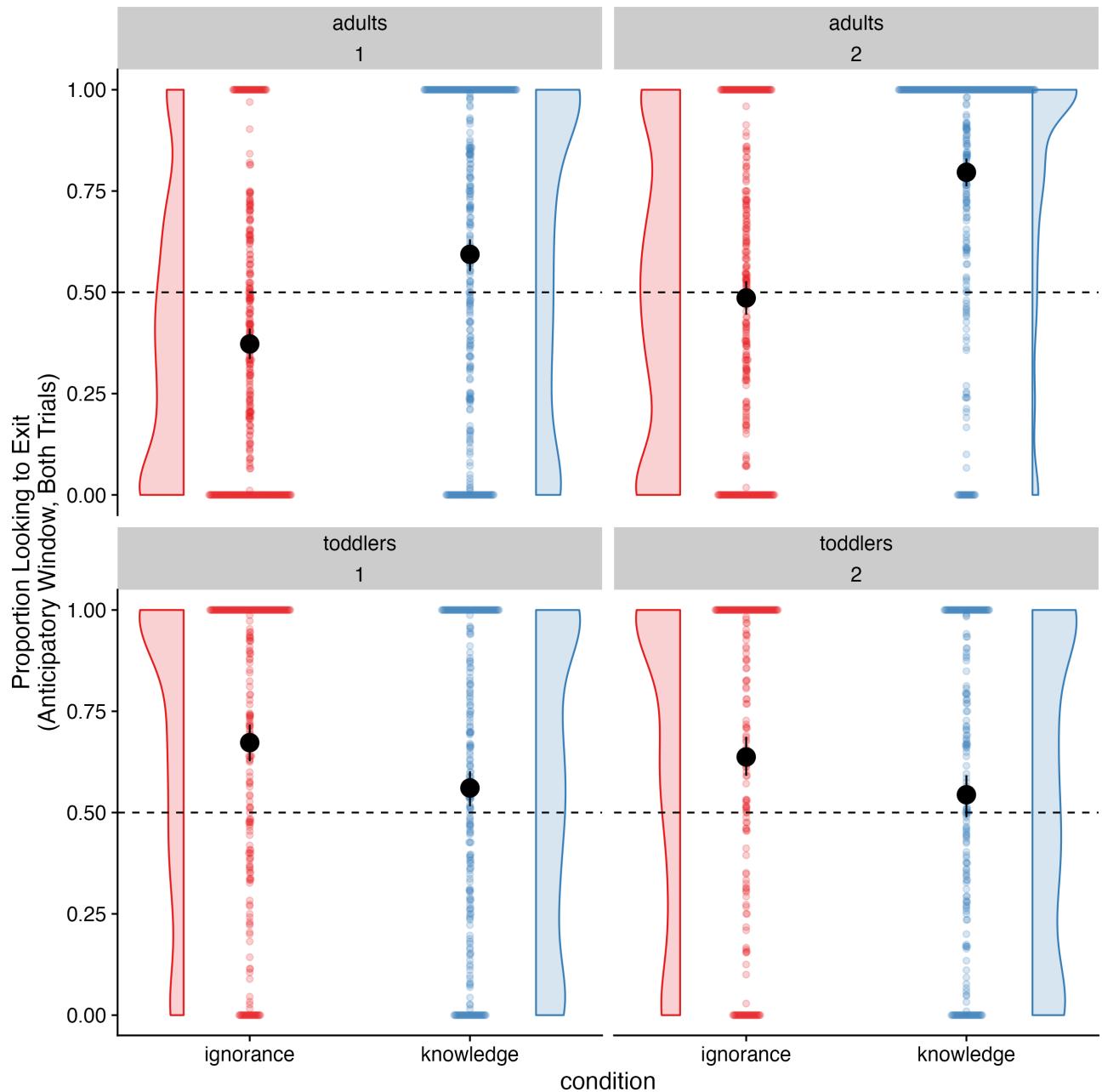


Figure 6. Proportional exit looking for the first and second test trial for toddlers and adults in the ignorance and knowledge condition.

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

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

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

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

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

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

903 simpler null model without the main effect of condition was inconclusive, $BF = 0.7$. For

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

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

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

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

908 for ignorance trials.

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

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

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

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

913 familiarization trials. Comparing the full model to the null model of toddlers revealed a

914 Bayes Factor of $BF = 16.9$, providing evidence in favor of the full model that included the

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

916 over the null model was $BF > 1000$, indicating that the inclusion of condition in the full

917 model substantially improved the explanation of the observed data.

918 ***Correlation between familiarization and test.*** We also examined the

919 correlation between familiarization and test performance across the two age cohorts and

920 conditions (see Figure 7). While no significant correlations were found for adults in either

921 condition, toddlers in the knowledge condition exhibited a significant positive correlation of

922 anticipatory looking in familiarization and test, $r=0.15$, $t(254)=2.35$, $p=0.02$.

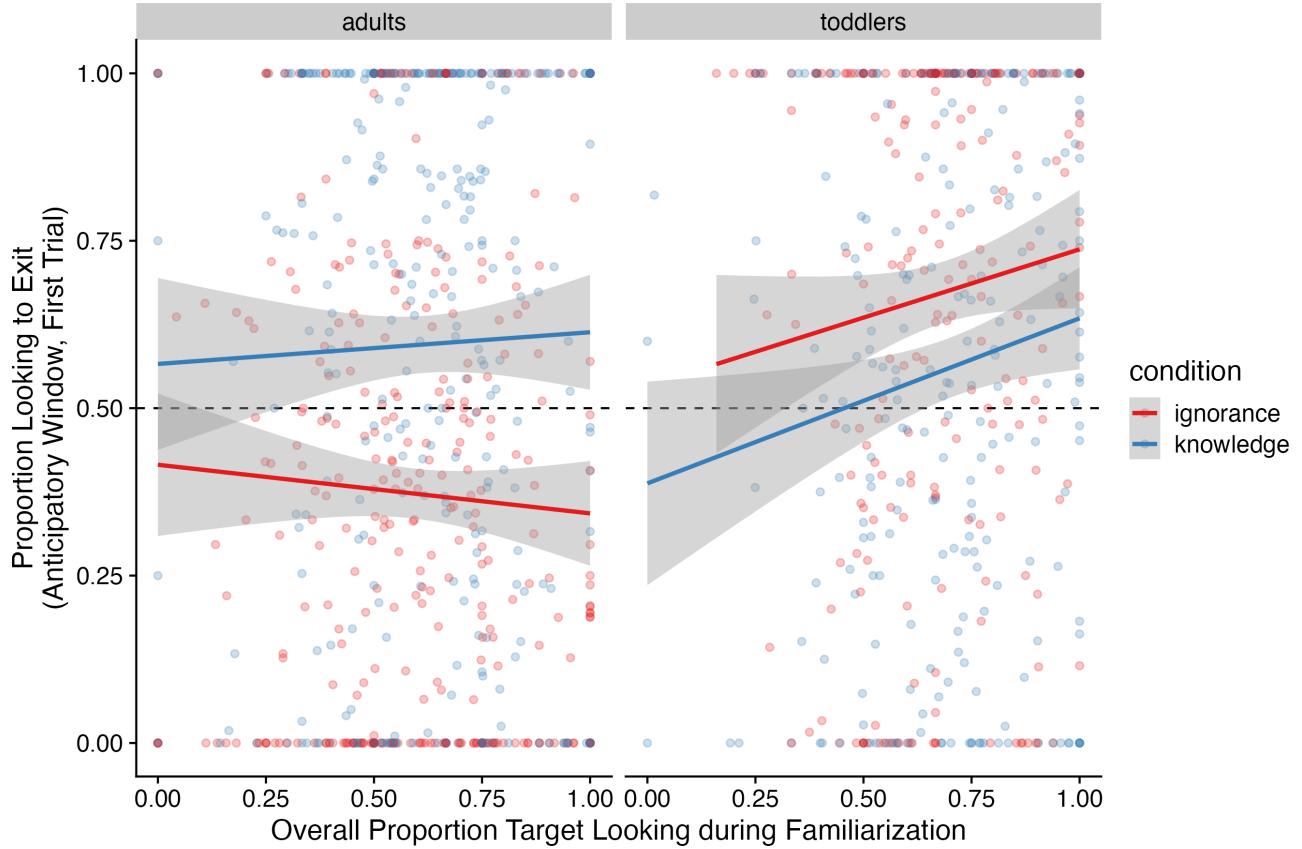


Figure 7. Relationship of anticipatory looking during familiarization and test for both age cohorts and conditions.

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

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 was minimal and also uncertain, suggesting no strong evidence that the effect of condition varied by method. The estimated Bayes factor comparing the full model to the null model was approximately $BF = 0.7$, which indicates that the data slightly favors the null model over the full model. This suggests that the predictors included in the full model do not substantially improve the explanation of the observed data compared to the null model.

933 For adults, the main effect of method was slightly negative but uncertain, suggesting

934 that the method had little to no clear effect on the outcome. The interaction between

935 condition and method was negative but with a wide credible interval crossing zero,

936 indicating uncertainty about whether the effect of condition varied by method. The

937 estimated Bayes factor in favor of the full model over the null model was $BF = 3.0$. This

938 Bayes factor indicates that the evidence in favor of the model is modest but not strong.

939 While the model is more likely than the null model to explain the observed data, the

940 support is relatively weak, suggesting that the predictors in the model provide only a small

941 improvement in explaining the data compared to the null model.

942 In sum, the analysis suggests that the method used (web-based vs. in-lab) does not

943 have a strong impact on AL, as the effect of method and its interaction with condition were

944 small and uncertain. Additionally, the results should be interpreted with caution due to

945 the relatively small sample size for web-based data compared to in-lab data collection,

946 which may limit the robustness of the findings.

947 **Box and tunnel looking vary separately by age or by condition.** We will

948 conduct an exploratory analysis using tighter AOIs around tunnel exits and boxes, asking

949 whether box and tunnel looking vary separately by age or by condition. In particular, we

950 expect that the difference in AL between the two conditions will be bigger for the tunnel

951 exits than for the box (as looks to the correct box might indicate looks to the target, which

952 is in the same box for both conditions, rather than action anticipation).

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

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

955 ***location change of the mouse.*** We fitted a Bayesian mixed-effects model to examine

956 the relationship between the number of shifts between mouse and bear and age cohort

957 during location change of the mouse, while accounting for random effects by lab. The effect

958 of condition was negative and approached significance, suggesting a potential reduction in

the number of shifts for the ignorance condition compared to the knowledge condition. The main effect of age cohort was positive and credible, Estimate=0.56, indicating that the the number of shifts was higher for adults than for toddlers. Importantly, the interaction between condition and age cohort was negative and credible, indicating that the negative effect of condition was more pronounced in the adult cohort (see Figure 8). Comparing this model to a simpler model without the interaction of condition and age cohort, a Bayes Factor of $BF > 1000$ was computed. This provides strong evidence in favor of including the interaction of condition and age cohort in the model.

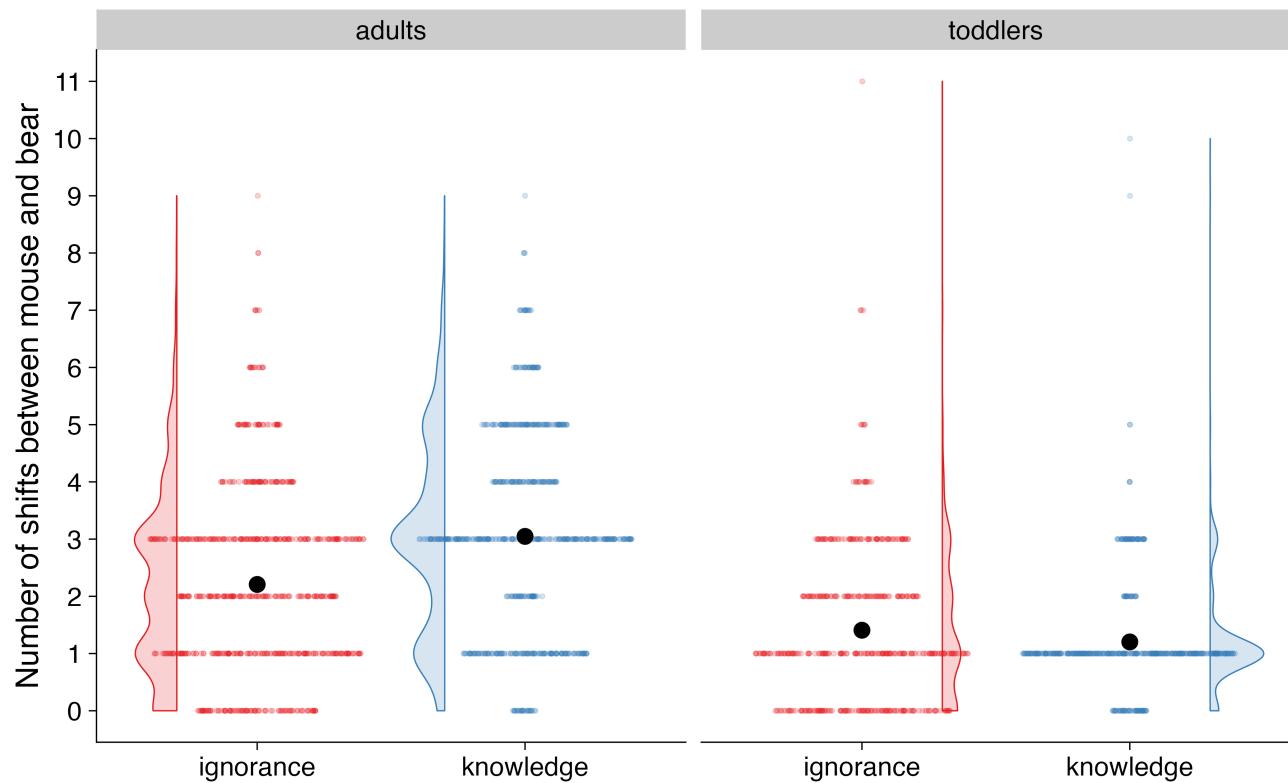


Figure 8. Number of shifts between mouse and bear during location change of mouse in the test phase for toddlers and adults in the ignorance and knowledge condition.

967 *AL as a function of number of gaze shifts between mouse and bear*

968 ***during location change.*** In order to examine the effect of condition and the number of
969 shifts between mouse and bear during location change of the mouse on anticipatory

looking, we fitted Bayesian mixed-effects models for each age cohort separately. The dependent variable was PTL in the anticipation period. The fixed effects included the main effects of condition, the number of shifts, and their interaction. We also included random intercepts and slopes for number of shifts within each participant and within each lab, allowing us to account for the hierarchical structure of the data and potential variability between labs and participants.

For toddlers, comparing this model to a simpler model without the interaction of condition and number of shifts, a Bayes Factor of $BF > 1000$ was computed, indicating that the data strongly favors the null model over the full model. This suggests that the predictors number of shifts and the interaction with condition included in the full model do not improve the explanation of the observed data compared to the null model.

For adults, the number of shifts showed a small but credible positive effect, suggesting that more shifts were associated with an increase in PTL. The interaction between condition and the number of shifts was negative and credible, indicating that the effect of condition became more negative as the number of shifts increased. The estimated Bayes factor comparing the full model to the null model was approximately $BF > 1000$, providing strong evidence in favor of the full model over the null model.

AL as a function of differential fixation times of bear and mouse during location change of the mouse. In order to examine the effect of condition and the difference in looking times for mouse and bear during location change of the mouse on anticipatory looking, we fitted a Bayesian mixed-effects model for both age cohorts separately. The dependent variable was the proportion of target looking. The fixed effects included the main effects of condition, the difference in fixation times of mouse and bear, and their interaction. We also included random intercepts and slopes for differences in fixation times of mouse and bear within each participant and within each lab, allowing us to account for the hierarchical structure of the data and potential variability between labs and participants.

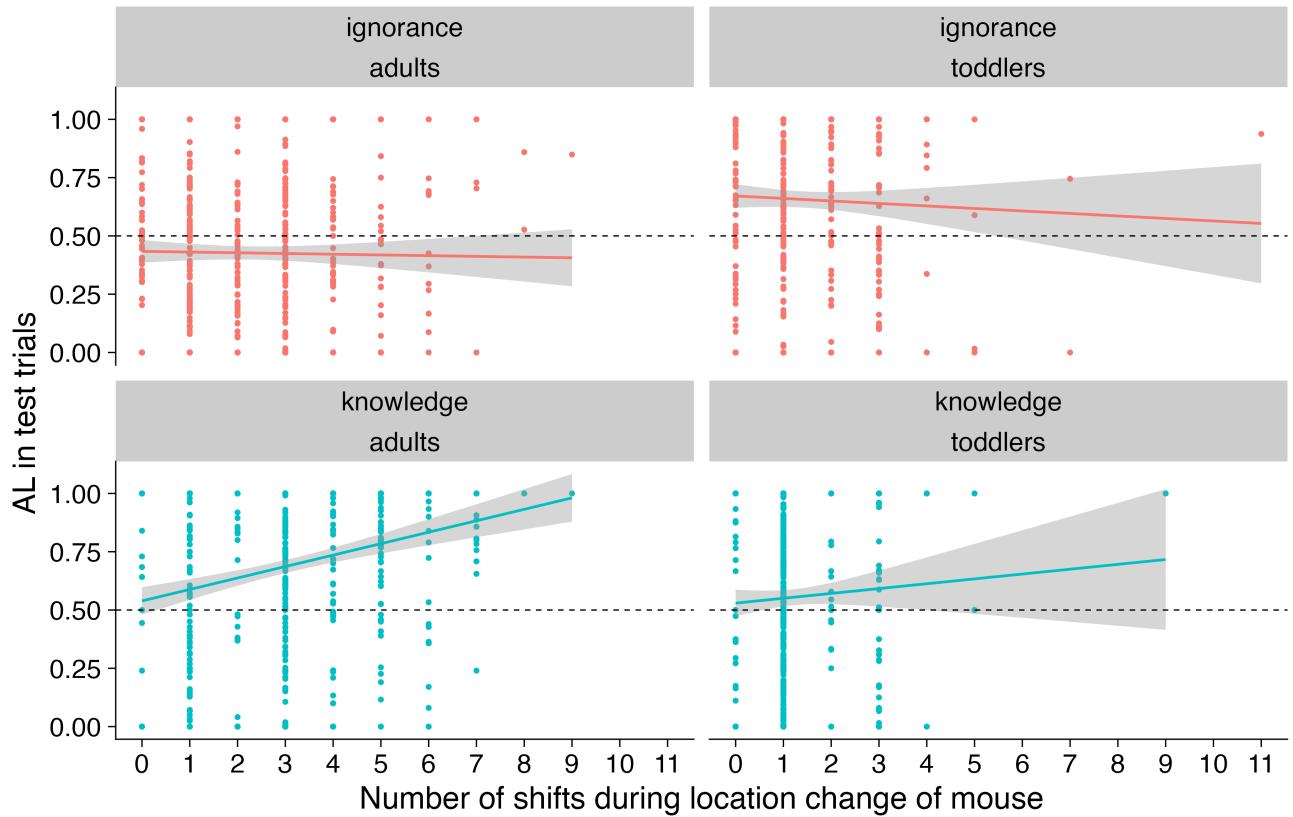


Figure 9. AL as a function of number of shifts between mouse and bear during location change of mouse in the test phase for toddlers and adults.

997 For toddlers, the fixed effect of difference in mouse-bear looking on anticipatory

998 looking was essentially zero, Estimate = 0, as was the interaction between condition and

999 diff_mouse_bear, Estimate = 0. These results indicate that neither the main effect of

1000 fixation time differences nor its interaction with condition meaningfully predicted

1001 anticipatory looking.

1002 Comparing this model to a simpler model without the interaction of condition and

1003 difference in mouse-bear looking, a Bayes Factor of $BF > 1000$ was computed. This

1004 provides overwhelming evidence that the inclusion of condition and the difference in

1005 fixation times improves model fit. However, the individual contributions of the difference in

1006 mouse-bear looking and its interaction with condition appear negligible, as indicated by

1007 their near-zero estimates and confidence intervals.

1008 For adults, the regression results indicated the following: The coefficient for
1009 diff_mouse_bear was essentially zero, Estimate = 0, suggesting no notable effect of the
1010 difference in fixation times on anticipatory looking. Additionally, the interaction between
1011 condition and the difference in mouse-bear looking was non-significant, Estimate = 0,
1012 indicating that the relationship between condition and anticipatory looking did not vary
1013 meaningfully with changes in fixation time difference.

1014 When comparing this full model with a simpler model that excluded the interaction
1015 between condition and the difference in mouse-bear looking, we obtained a Bayes factor of
1016 $BF > 1000$. This provides very strong evidence in favor of the simpler model, suggesting
1017 that the interaction between condition and fixation time difference does not substantially
1018 improve the model's fit. Overall, these results imply that while condition affects
1019 anticipatory looking, the difference in fixation times between the mouse and bear, and its
1020 interaction with condition, do not meaningfully contribute to the prediction of target
1021 looking behavior.

1022 General Discussion

1023 The current large-scale, multi-lab study set out to examine whether toddlers and
1024 adults engage in spontaneous ToM. In particular, we used an anticipatory looking
1025 paradigm to explore whether 18- to 27-month-old toddlers and adults distinguish between
1026 two basic forms of epistemic states: knowledge and ignorance. Our call for participation
1027 resulted in contributions from 47 labs, representing a total of 809 toddlers from xyz
1028 countries and 805 adults from xyz countries, of which 1224 were included in the final
1029 sample used for analysis (see Table 1). We begin our discussion by summarizing the
1030 principal results of the study with respect to confirmatory analysis and then discuss
1031 limitations of the study as well as future directions.

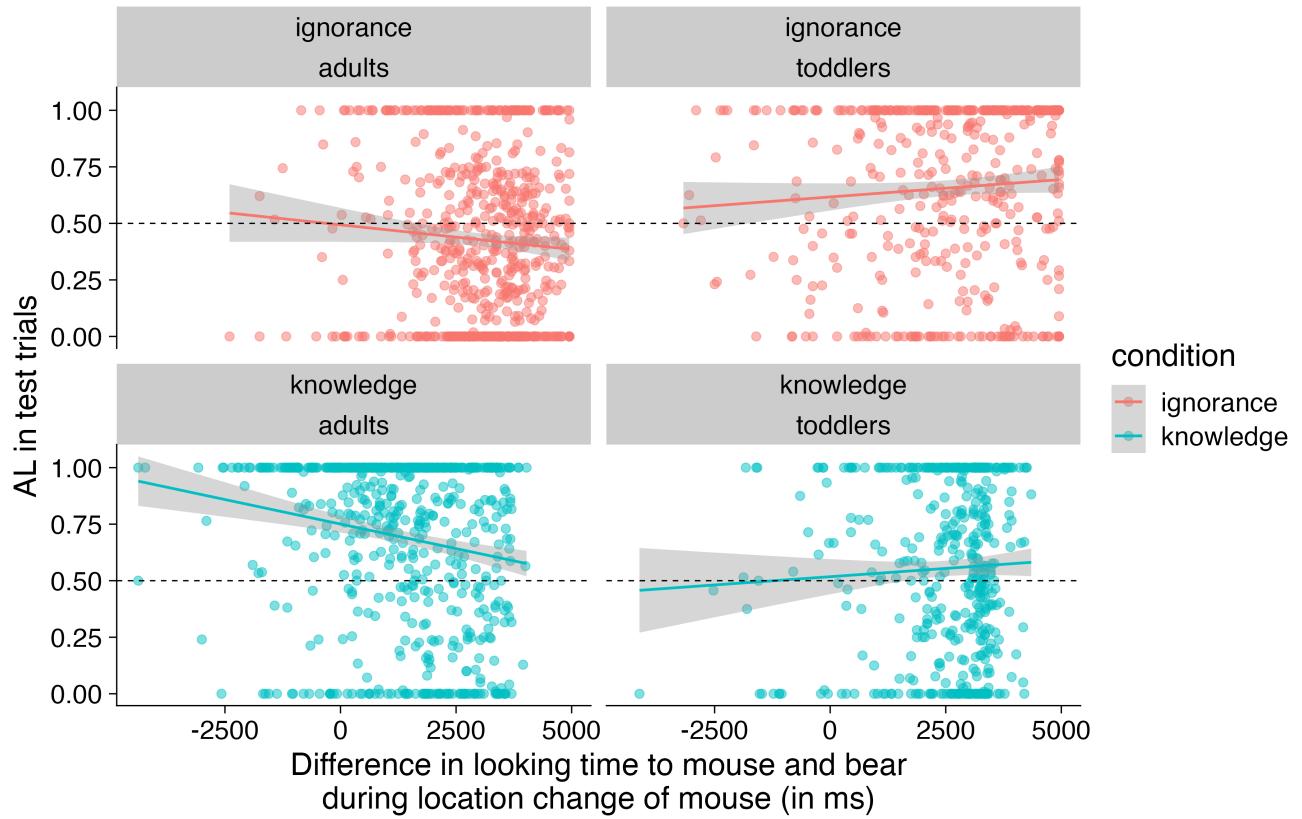


Figure 10. AL as a function of the difference in looking time to mouse and bear during location change of mouse (in ms) for each age cohort and each condition.

1032 Conclusion

1033

References

- 1034 Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and
1035 belief-like states? *Psychological Review*, 116(4), 953.
- 1036 Baillargeon, R., Buttelmann, D., & Southgate, V. (2018). Invited commentary:
1037 Interpreting failed replications of early false-belief findings: Methodological and
1038 theoretical considerations. *Cognitive Development*, 46, 112–124.
- 1039 Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants.
1040 *Trends in Cognitive Sciences*, 14(3), 110–118.
- 1041 Barone, P., Corradi, G., & Gomila, A. (2019). Infants' performance in
1042 spontaneous-response false belief tasks: A review and meta-analysis. *Infant Behavior
and Development*, 57, 101350.
- 1043 Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects
1044 models. Frontiers Media SA.
- 1045 Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan.
1046 *Journal of Statistical Software*, 80, 1–28.
- 1047 Burnside, K., Ruel, A., Azar, N., & Poulin-Dubois, D. (2018). Implicit false belief across
1048 the lifespan: Non-replication of an anticipatory looking task. *Cognitive Development*,
1049 46, 4–11.
- 1050 Buttelmann, D., Carpenter, M., & Tomasello, M. (2009). Eighteen-month-old infants show
1051 false belief understanding in an active helping paradigm. *Cognition*, 112(2), 337–342.
- 1052 Buttelmann, F., & Kovács, Á. M. (2019). 14-month-olds anticipate others' actions based
1053 on their belief about an object's identity. *Infancy*, 24(5), 738–751.
- 1054 Buttelmann, F., Suhrke, J., & Buttelmann, D. (2015). What you get is what you believe:
1055 Eighteen-month-olds demonstrate belief understanding in an unexpected-identity task.
1056 *Journal of Experimental Child Psychology*, 131, 94–103.
- 1057 Carruthers, P. (2013). Mindreading in infancy. *Mind & Language*, 28(2), 141–172.
- 1058 Clements, W. A., & Perner, J. (1994). Implicit understanding of belief. *Cognitive*

- 1060 *Development*, 9(4), 377–395.
- 1061 Csibra, G., & Gergely, G. (2007). “Obsessed with goals”: Functions and mechanisms of
1062 teleological interpretation of actions in humans. *Acta Psychologica*, 124(1), 60–78.
- 1063 Dennett, D. C. (1989). *The intentional stance*. MIT press.
- 1064 Dörrenberg, S., Rakoczy, H., & Liszkowski, U. (2018). How (not) to measure infant theory
1065 of mind: Testing the replicability and validity of four non-verbal measures. *Cognitive*
1066 *Development*, 46, 12–30.
- 1067 Dörrenberg, S., Wenzel, L., Proft, M., Rakoczy, H., & Liszkowski, U. (2019). Reliability
1068 and generalizability of an acted-out false belief task in 3-year-olds. *Infant Behavior and*
1069 *Development*, 54, 13–21.
- 1070 Elsner, B., & Adam, M. (2021). Infants’ goal prediction for simple action events: The role
1071 of experience and agency cues. *Topics in Cognitive Science*, 13(1), 45–62.
- 1072 Fabricius, W. V., Boyer, T. W., Weimer, A. A., & Carroll, K. (2010). True or false: Do
1073 5-year-olds understand belief? *Developmental Psychology*, 46(6), 1402.
- 1074 Flavell, J. H. (1988). *The development of children’s knowledge about the mind: From*
1075 *cognitive connections to mental representations*.
- 1076 Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children’s
1077 knowledge about visual perception: Further evidence for the level 1–level 2 distinction.
1078 *Developmental Psychology*, 17(1), 99.
- 1079 Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., ...
1080 Yurovsky, D. (2017). A collaborative approach to infant research: Promoting
1081 reproducibility, best practices, and theory-building. *Infancy*, 22(4), 421–435.
1082 <https://doi.org/10.1111/infa.12182>
- 1083 Frank, M. C., Vul, E., & Saxe, R. (2012). Measuring the development of social attention
1084 using free-viewing. *Infancy*, 17(4), 355–375.
- 1085 Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron*, 50(4), 531–534.
- 1086 Ganglmayer, K., Attig, M., Daum, M. M., & Paulus, M. (2019). Infants’ perception of

- 1087 goal-directed actions: A multi-lab replication reveals that infants anticipate paths and
1088 not goals. *Infant Behavior and Development*, 57, 101340.
- 1089 Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naive theory of
1090 rational action. *Trends in Cognitive Sciences*, 7(7), 287–292.
- 1091 Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12
1092 months of age. *Cognition*, 56(2), 165–193.
- 1093 Gliga, T., Jones, E. J., Bedford, R., Charman, T., & Johnson, M. H. (2014). From early
1094 markers to neuro-developmental mechanisms of autism. *Developmental Review*, 34(3),
1095 189–207.
- 1096 Gronau, Q. F., Sarafoglou, A., Matzke, D., Ly, A., Boehm, U., Marsman, M., ...
1097 Steingroever, H. (2017). A tutorial on bridge sampling. *Journal of Mathematical
1098 Psychology*, 81, 80–97.
- 1099 Grosse Wiesmann, C., Friederici, A. D., Singer, T., & Steinbeis, N. (2017). Implicit and
1100 explicit false belief development in preschool children. *Developmental Science*, 20(5),
1101 e12445.
- 1102 Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know?
1103 *Animal Behaviour*, 61(1), 139–151.
- 1104 Hayashi, T., Akikawa, R., Kawasaki, K., Egawa, J., Minamimoto, T., Kobayashi, K., et
1105 al.others. (2020). Macaques exhibit implicit gaze bias anticipating others'
1106 false-belief-driven actions via medial prefrontal cortex. *Cell Reports*, 30(13), 4433–4444.
- 1107 Heyes, C. (2014). Submentalizing: I am not really reading your mind. *Perspectives on
1108 Psychological Science*, 9(2), 131–143.
- 1109 Hogrefe, G.-J., Wimmer, H., & Perner, J. (1986). Ignorance versus false belief: A
1110 developmental lag in attribution of epistemic states. *Child Development*, 567–582.
- 1111 Horschler, D. J., MacLean, E. L., & Santos, L. R. (2020). Do non-human primates really
1112 represent others' beliefs? *Trends in Cognitive Sciences*, 24(8), 594–605.
- 1113 Kaminski, J., Call, J., & Tomasello, M. (2008). Chimpanzees know what others know, but

- 1114 not what they believe. *Cognition*, 109(2), 224–234.
- 1115 Kampis, D., Buttelmann, F., & Kovács, Á. M. (2020). *Developing a theory of mind: Are*
1116 *infants sensitive to how other people represent the world?*
- 1117 Kampis, D., Karman, P., Csibra, G., Southgate, V., & Hernik, M. (2021). A two-lab direct
1118 replication attempt of southgate, senju and csibra (2007). *Royal Society Open Science*,
1119 8(8), 210190.
- 1120 Kano, F., Krupenye, C., Hirata, S., Tomonaga, M., & Call, J. (2019). Great apes use
1121 self-experience to anticipate an agent's action in a false-belief test. *Proceedings of the*
1122 *National Academy of Sciences*, 116(42), 20904–20909.
- 1123 Karg, K., Schmelz, M., Call, J., & Tomasello, M. (2015). The goggles experiment: Can
1124 chimpanzees use self-experience to infer what a competitor can see? *Animal Behaviour*,
1125 105, 211–221.
- 1126 Király, I., Oláh, K., Csibra, G., & Kovács, Á. M. (2018). Retrospective attribution of false
1127 beliefs in 3-year-old children. *Proceedings of the National Academy of Sciences*, 115(45),
1128 11477–11482.
- 1129 Knudsen, B., & Liszkowski, U. (2012). 18-month-olds predict specific action mistakes
1130 through attribution of false belief, not ignorance, and intervene accordingly. *Infancy*,
1131 17(6), 672–691.
- 1132 Kovács, Á. M. (2016). Belief files in theory of mind reasoning. *Review of Philosophy and*
1133 *Psychology*, 7, 509–527.
- 1134 Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: Susceptibility to
1135 others' beliefs in human infants and adults. *Science*, 330(6012), 1830–1834.
- 1136 Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate
1137 that other individuals will act according to false beliefs. *Science*, 354(6308), 110–114.
- 1138 Kulke, L., Duhn, B. von, Schneider, D., & Rakoczy, H. (2018). Is implicit theory of mind a
1139 real and robust phenomenon? Results from a systematic replication study.
1140 *Psychological Science*, 29(6), 888–900.

- 1141 Kulke, L., & Hinrichs, M. A. B. (2021). Implicit theory of mind under realistic social
1142 circumstances measured with mobile eye-tracking. *Scientific Reports*, 11(1), 1215.
- 1143 Kulke, L., Johannsen, J., & Rakoczy, H. (2019). Why can some implicit theory of mind
1144 tasks be replicated and others cannot? A test of mentalizing versus submentalizing
1145 accounts. *PloS One*, 14(3), e0213772.
- 1146 Kulke, L., & Rakoczy, H. (2018). Implicit theory of mind—an overview of current
1147 replications and non-replications. *Data in Brief*, 16, 101–104.
- 1148 Kulke, L., & Rakoczy, H. (2019). Testing the role of verbal narration in implicit theory of
1149 mind tasks. *Journal of Cognition and Development*, 20(1), 1–14.
- 1150 Kulke, L., Reiß, M., Krist, H., & Rakoczy, H. (2018). How robust are anticipatory looking
1151 measures of theory of mind? Replication attempts across the life span. *Cognitive
1152 Development*, 46, 97–111.
- 1153 Kulke, L., Wübker, M., & Rakoczy, H. (2019). Is implicit theory of mind real but hard to
1154 detect? Testing adults with different stimulus materials. *Royal Society Open Science*,
1155 6(7), 190068.
- 1156 Leslie, A. M. (2005). Developmental parallels in understanding minds and bodies. *Trends
1157 in Cognitive Sciences*, 9(10), 459–462.
- 1158 Liszkowski, U., Carpenter, M., & Tomasello, M. (2007). Pointing out new news, old news,
1159 and absent referents at 12 months of age. *Developmental Science*, 10(2), F1–F7.
- 1160 Low, J., & Watts, J. (2013). Attributing false beliefs about object identity reveals a
1161 signature blind spot in humans' efficient mind-reading system. *Psychological Science*,
1162 24(3), 305–311.
- 1163 Luo, Y., & Baillargeon, R. (2007). Do 12.5-month-old infants consider what objects others
1164 can see when interpreting their actions? *Cognition*, 105(3), 489–512.
- 1165 Luo, Y., & Baillargeon, R. (2010). Toward a mentalistic account of early psychological
1166 reasoning. *Current Directions in Psychological Science*, 19(5), 301–307.
- 1167 ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research

- 1168 using the infant-directed-speech preference. *Advances in Methods and Practices in*
1169 *Psychological Science*, 3(1), 24–52.
- 1170 Martin, A., & Santos, L. R. (2016). What cognitive representations support primate theory
1171 of mind? *Trends in Cognitive Sciences*, 20(5), 375–382.
- 1172 Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate
1173 conjunction effects? An exercise in adversarial collaboration. *Psychological Science*,
1174 12(4), 269–275.
- 1175 Meristo, M., Morgan, G., Geraci, A., Iozzi, L., Hjelmquist, E., Surian, L., & Siegal, M.
1176 (2012). Belief attribution in deaf and hearing infants. *Developmental Science*, 15(5),
1177 633–640.
- 1178 Moll, H., & Tomasello, M. (2006). Level 1 perspective-taking at 24 months of age. *British*
1179 *Journal of Developmental Psychology*, 24(3), 603–613.
- 1180 O'Neill, D. K. (1996). Two-year-old children's sensitivity to a parent's knowledge state
1181 when making requests. *Child Development*, 67(2), 659–677.
- 1182 Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs?
1183 *Science*, 308(5719), 255–258.
- 1184 Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016).
1185 *Proceedings of the 25th international joint conference on artificial intelligence (IJCAI)*.
- 1186 Perner, J. (1991). *Understanding the representational mind*. The MIT Press.
- 1187 Perner, J., & Ruffman, T. (2005). Infants' insight into the mind: How deep? *Science*,
1188 308(5719), 214–216.
- 1189 Phillips, J., Buckwalter, W., Cushman, F., Friedman, O., Martin, A., Turri, J., ... Knobe,
1190 J. (2021). Knowledge before belief. *Behavioral and Brain Sciences*, 44, e140.
- 1191 Poulin-Dubois, D., Rakoczy, H., Burnside, K., Crivello, C., Dörrenberg, S., Edwards, K., et
1192 al.others. (2018). Do infants understand false beliefs? We don't know yet—a
1193 commentary on baillargeon, buttelmann and southgate's commentary. *Cognitive*
1194 *Development*, 48, 302–315.

- 1195 Powell, L. J., Hobbs, K., Bardis, A., Carey, S., & Saxe, R. (2018). Replications of implicit
1196 theory of mind tasks with varying representational demands. *Cognitive Development*,
1197 46, 40–50.
- 1198 Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind?
1199 *Behavioral and Brain Sciences*, 1(4), 515–526.
- 1200 Prielwasser, B., Fowles, F., Schweller, K., & Perner, J. (2020). Mistaken max befriends
1201 duplo girl: No difference between a standard and an acted-out false belief task. *Journal*
1202 *of Experimental Child Psychology*, 191, 104756.
- 1203 Prielwasser, B., Rafetseder, E., Gargitter, C., & Perner, J. (2018). Helping as an early
1204 indicator of a theory of mind: Mentalism or teleology? *Cognitive Development*, 46,
1205 69–78.
- 1206 Ruffman, T. (1996). Do children understand the mind by means of simulation or a theory?
1207 Evidence from their understanding of inference. *Mind & Language*, 11(4), 388–414.
- 1208 Schneider, D., Bayliss, A. P., Becker, S. I., & Dux, P. E. (2012). Eye movements reveal
1209 sustained implicit processing of others' mental states. *Journal of Experimental*
1210 *Psychology: General*, 141(3), 433.
- 1211 Schneider, D., Slaughter, V. P., Bayliss, A. P., & Dux, P. E. (2013). A temporally
1212 sustained implicit theory of mind deficit in autism spectrum disorders. *Cognition*,
1213 129(2), 410–417.
- 1214 Schneider, D., Slaughter, V. P., & Dux, P. E. (2017). Current evidence for automatic
1215 theory of mind processing in adults. *Cognition*, 162, 27–31.
- 1216 Schuwerk, T., Prielwasser, B., Sodian, B., & Perner, J. (2018). The robustness and
1217 generalizability of findings on spontaneous false belief sensitivity: A replication
1218 attempt. *Royal Society Open Science*, 5(5), 172273.
- 1219 Scott, R. M., & Baillargeon, R. (2009). Which penguin is this? Attributing false beliefs
1220 about object identity at 18 months. *Child Development*, 80(4), 1172–1196.
- 1221 Scott, R. M., & Baillargeon, R. (2017). Early false-belief understanding. *Trends in*

- 1222 *Cognitive Sciences*, 21(4), 237–249.
- 1223 Scott, R. M., Richman, J. C., & Baillargeon, R. (2015). Infants understand deceptive
1224 intentions to implant false beliefs about identity: New evidence for early mentalistic
1225 reasoning. *Cognitive Psychology*, 82, 32–56.
- 1226 Senju, A., Southgate, V., Miura, Y., Matsui, T., Hasegawa, T., Tojo, Y., ... Csibra, G.
1227 (2010). Absence of spontaneous action anticipation by false belief attribution in children
1228 with autism spectrum disorder. *Development and Psychopathology*, 22(2), 353–360.
- 1229 Senju, A., Southgate, V., Snape, C., Leonard, M., & Csibra, G. (2011). Do 18-month-olds
1230 really attribute mental states to others? A critical test. *Psychological Science*, 22(7),
1231 878–880.
- 1232 Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: An absence of
1233 spontaneous theory of mind in asperger syndrome. *Science*, 325(5942), 883–885.
- 1234 Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., et al.others.
1235 (2020). Online developmental science to foster innovation, access, and impact. *Trends
1236 in Cognitive Sciences*, 24(9), 675–678.
- 1237 Southgate, V., Johnson, M. H., Karoui, I. E., & Csibra, G. (2010). Motor system activation
1238 reveals infants' on-line prediction of others' goals. *Psychological Science*, 21(3), 355–359.
- 1239 Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of
1240 false belief by 2-year-olds. *Psychological Science*, 18(7), 587–592.
- 1241 Southgate, V., & Vernetti, A. (2014). Belief-based action prediction in preverbal infants.
1242 *Cognition*, 130(1), 1–10.
- 1243 Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants.
1244 *Psychological Science*, 18(7), 580–586.
- 1245 Surian, L., & Franchin, L. (2020). On the domain specificity of the mechanisms
1246 underpinning spontaneous anticipatory looks in false-belief tasks. *Developmental
1247 Science*, 23(6), e12955.
- 1248 Surian, L., & Geraci, A. (2012). Where will the triangle look for it? Attributing false

- 1249 beliefs to a geometric shape at 17 months. *British Journal of Developmental*
1250 *Psychology*, 30(1), 30–44.
- 1251 Thoermer, C., Sodian, B., Vuori, M., Perst, H., & Kristen, S. (2012). Continuity from an
1252 implicit to an explicit understanding of false belief from infancy to preschool age.
1253 *British Journal of Developmental Psychology*, 30(1), 172–187.
- 1254 Träuble, B., Marinović, V., & Pauen, S. (2010). Early theory of mind competencies: Do
1255 infants understand others' beliefs? *Infancy*, 15(4), 434–444.
- 1256 Wellman, H. M., & Cross, D. (2001). Theory of mind and conceptual change. *Child*
1257 *Development*, 72(3), 702–707.
- 1258 Wiesmann, C. G., Friederici, A. D., Disla, D., Steinbeis, N., & Singer, T. (2018).
1259 Longitudinal evidence for 4-year-olds' but not 2-and 3-year-olds' false belief-related
1260 action anticipation. *Cognitive Development*, 46, 58–68.
- 1261 Woodward, A. L., & Sommerville, J. A. (2000). Twelve-month-old infants interpret action
1262 in context. *Psychological Science*, 11(1), 73–77.
- 1263 Yang, Q., Bucci, M. P., & Kapoula, Z. (2002). The latency of saccades, vergence, and
1264 combined eye movements in children and in adults. *Investigative Ophthalmology &*
1265 *Visual Science*, 43(9), 2939–2949.

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 | conf.low | conf.high |
|---------------------|------------------|----------|----------|-----------|
| PTL toddlers | Intercept Effect | 0.12 | 0.09 | 0.15 |
| PTL adults | Intercept Effect | 0.26 | 0.23 | 0.29 |
| First Look toddlers | Intercept Effect | 0.44 | 0.27 | 0.61 |
| First Look adults | Intercept Effect | 0.91 | 0.78 | 1.05 |

Table 4

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

| model | term | estimate | conf.low | conf.high |
|---------------------|------------------|----------|----------|-----------|
| PTL toddlers | Condition Effect | 0.10 | 0.03 | 0.17 |
| PTL adults | Condition Effect | -0.20 | -0.26 | -0.15 |
| First Look toddlers | Condition Effect | 0.53 | 0.13 | 0.93 |
| First Look adults | Condition Effect | -0.89 | -1.21 | -0.56 |