Developmental Psychology

Children's Developing Ability to Make Sense of Evaluative Judgments Based on a Speaker's Evaluative History

F. Ece Özkan and Samuel Ronfard

Online First Publication, April 16, 2026. https://dx.doi.org/10.1037/dev0002190

AMERICAN PSYCHOLOGICAL ASSOCIATION

Developmental Psychology

ISSN: 0012-1649

https://doi.org/10.1037/dev0002190

Children's Developing Ability to Make Sense of Evaluative Judgments Based on a Speaker's Evaluative History

F. Ece Özkan and Samuel Ronfard

Department of Psychological and Brain Sciences, University of Toronto Mississauga

When someone says something is "okay," it is not always clear what they mean. It could mean they like it. It could mean they do not like it. To decide which interpretation is correct, we can use that speaker's past evaluative comments to determine their evaluative baseline and assess whether "okay" falls below or above it. Across two studies ($N = 366$), we examined whether children (aged 5–8) and adults do so. The always-likes character had a history of making positive evaluations. The never-likes character had a history of making negative evaluations. In Study 1, 7-year-olds and adults (but not 5-year-olds) more frequently inferred that the never-likes character liked the item compared to the always-likes character. In Study 2, after being explicitly told that "it's okay" could mean that the speaker liked or did not like the item, both 5- to 8-year-old children and adults more frequently inferred that the never-likes character liked the item compared to the always-likes character. With increasing age, participants' tendency to interpret "it's okay" positively decreased across conditions. These results contribute to our understanding of how children and adults make sense of ambiguous statements in evaluative contexts.

Public Significance Statement

When someone says something is "ok," it could mean they like it, or it could mean they do not like it. To decide what they meant, we can think about what they have said in the past—do they typically like or dislike similar things? By age 5, children can use this strategy, but only when explicitly told that "it's ok" is ambiguous; otherwise, they default to interpreting it as liking. By 7, children no longer need this scaffolding.

Keywords: mental state inferences, pragmatic reasoning, evaluative information, preferences, ambiguity

Supplemental materials: https://doi.org/10.1037/dev0002190.supp

Imagine that you are inviting someone over for dinner. You tell them that you are considering making lasagna. They respond, "Lasagna's ok." Should you make lasagna for them? This depends on whether you think they like lasagna or not. To figure out whether they actually like lasagna, you can use multiple cues, for example, how they said it (prosody), how they acted when they said it (their gestures), the context's cultural norms (i.e., whether "ok" is typically used positively or negatively), and what you know about the speaker based on your past interactions with them. In the present study, we focus on the use of past interactions with a speaker to interpret a speaker's true preferences. We test young children and adults to understand the development of this ability.

Let us return to our lasagna example to illustrate how past interactions with a speaker might help interpret an ambiguous evaluative statement. If a person generally makes positive comments about food when dining out, for example, "good" or "great," you might infer that they *do not* like lasagna and are trying to be polite. However, if this person generally makes negative evaluations, for example, "bad" or "not good," you might infer that they really *do* like lasagna. As these examples demonstrate, evaluative utterances

Lauren Howard served as action editor.

F. Ece Özkan https://orcid.org/0000-0002-1636-9609

The authors declare that they have no conflicts of interest to disclose. This research was supported by a Social Sciences and Humanities Research Council of Canada Insight Grant awarded to Samuel Ronfard (Grant 435-2021-0170).

The authors thank all participants and the parents of the child participants. The authors thank all members of the Childhood Learning and Development Lab at the University of Toronto Mississauga and the research assistants for their help with data collection and recruitment. Special thanks go to Corrin Doucette for designing the visuals used in the study and to Esther Portillo-Cisneros, Allison Gutierrez, Rachel Lipson, and Kyli Kindree for voicing the prompts.

F. Ece Özkan played a lead role in data curation, formal analysis, investigation, methodology, project administration, visualization, and writing–original draft and an equal role in conceptualization and writing–review and editing. Samuel Ronfard played a lead role in funding acquisition, resources, and supervision, a supporting role in formal analysis, methodology, and project administration, and an equal role in conceptualization and writing–review and editing.

Correspondence concerning this article should be addressed to F. Ece Özkan, Department of Psychological and Brain Sciences, University of Toronto Mississauga, 3359 Mississauga Road, Communication, Culture, and Technology Building, 4025, Mississauga, ON L5L 1C6, Canada. Email: fece.ozkan@mail.utoronto.ca

are sometimes ambiguous, but this ambiguity can be resolved by using a speaker's evaluative baseline (their typical way of evaluating similar items). As a result, it is possible to infer opposite preferences based on the same utterance if speakers have different evaluative baselines. Across two studies, we focused on children's and adults' ability to use a person's past evaluative judgments to calibrate their inferences about that person's true preferences.

Such reasoning calls upon our theory of mind—our ability to infer others' beliefs, intentions, and desires from their actions, as well as our ability to infer how mental states shape actions (see Wellman, 2014). Children's theory of mind develops rapidly between 2 and 6. During this time, children progressively understand that people can have competing desires and beliefs, false beliefs, and display different emotions than the ones they are feeling (Wellman, 2014; Wellman & Liu, 2004). By age 4, children make sophisticated inferences about speakers' mental states from their actions, emotional expressions, and testimony. For instance, 4-year-olds who hear "My friend has glasses" infer that the speaker means someone with *only* glasses, not someone with glasses and a hat—demonstrating pragmatic inference about intended reference (Stiller et al., 2015). By age 5, children integrate multiple cues when reasoning about preferences and desires. They infer an agent prefers A over B only when both are equally accessible; if A is easy to reach but B is difficult, they withhold this inference, recognizing that action choices reflect effort costs as well as preferences (Jara-Ettinger et al., 2015). Similarly, 5-year-olds use emotional reactions to infer desires: If someone looks happy after receiving an apple (when the alternative was a banana), children conclude that the person wanted the apple (Wu & Schulz, 2018). These studies demonstrate children's early sensitivity to contextual factors in mental state inference.

Recent theoretical and empirical work has also shown that listeners use their understanding of speakers' beliefs and intentions as well as common ground to constrain their interpretation of what they are told (Bohn & Köymen, 2018; Bohn, Tessler, et al., 2021; Shafto et al., 2014; Sullivan & Barner, 2016; see Vasil, 2023). This allows listeners to recover a speaker's intended meaning via pragmatic reasoning (Degen, 2023; Goodman & Frank, 2016; see also Grice, 1975). Children aged 2–3 can resolve referential ambiguity by using what they know about a speaker's knowledge, for example, whether the speaker is familiar with an item or not (Bleijlevens et al., 2023; Bohn, Tessler, et al., 2021; see also Bohn, Le et al., 2021). In the context of social learning, these inferences also allow children to evaluate pedagogical information (Gweon, 2021; Landrum et al., 2015; Shafto et al., 2012; Sobel & Kushnir, 2013). For instance, 5-year-old children can infer whether an informant is underinformative or overinformative based on the information an informant provided to a naïve or knowledgeable learner (Gweon et al., 2014; Gweon et al., 2018).

In sum, past work demonstrates that across contexts, young children can deploy sophisticated assumptions about agents' rationality to constrain their inferences about what they mean, like, and want. However, there are cases like the lasagna example, where these assumptions are not enough to make sense of what someone means. Such situations are common in daily life, and thus, resolving those ambiguities is important for navigating the world. In such cases, the meaning of "ok" has to be contextually determined by adapting to the statistics of specific speakers. A relevant example comes from our inferences about a speaker's certainty upon hearing a modality marker like "probably." We calibrate using the speaker's past statements. Past work with adults shows that listeners adapt their expectations of whether a speaker will say "might" or "probably" to qualify their statements based on whether the speaker showed a confident profile in the past—saying "probably" when the probability of an outcome was 60%—or a cautious one—saying "might" when the probability of an outcome was 60% (Schuster & Degen, 2020). In the case of the lasagna example, listeners also need to use their knowledge of the speaker's past evaluative judgments to identify the speaker's evaluative baseline and use this to make sense of the speaker's mental state (their desire) and therefore recover the meaning of "it's okay."

Past work on evaluative information, for example, people's subjective judgments about phenomena (Marble & Boseovski, 2020), shows that 4- to 8-year-olds prefer endorsing someone who provides positive evaluations (Boseovski et al., 2017). But children do not show blind trust in positive evaluations. By age 4, children prefer getting feedback for their work from an informant who was selective by providing positive feedback on good work but negative feedback on bad work over an informant who made positive comments for both good and bad work (Asaba et al., 2018). This type of overpraising makes the speaker uninformative or underinformative because the speaker's evaluations do not match the nature of the work, for example, whether it is a good or bad painting. Thus, Asaba et al. (2018) demonstrated that children can make inferences about whether a person's judgment is informative by noticing whether it covaries with real-world differences in painting quality. However, in cases like the lasagna example, there is no reference point for judging the informant's evaluation of the food as "ok." The food's rating of "ok" is neither accurate nor inaccurate.

Consider another example: When someone evaluates a new product called *torpotam*, listeners with no independent knowledge of *torpotam* must infer the speaker's intended meaning by considering the speaker's evaluative history and standards. In the present study, we seek to capture a common real-world scenario: interpreting others' evaluations of entities solely based on their previous discourse. To do so, in contrast to past work, we used a design in which participants could not use an objective criterion to interpret a speaker's comments or to make inferences about their selectivity and mental states. Our design isolated speaker-calibrated inference by using items that participants had no prior experience with. Participants had to use their knowledge of a speaker's past judgments to calibrate the meaning of an ambiguous evaluative expression.

Across two studies, we examined whether 5- to 8-year-old children and adults can use a speaker's evaluative history to infer that speaker's true preferences based on the ambiguous evaluative statement "It's okay." We hypothesized that children as young as 5 may be able to do so. However, it is also possible that the ability to interpret ambiguous evaluative statements by using speakers' past statements develops later. Past research showed that preschoolers demonstrate a positivity bias when reasoning about other people. For example, they prefer endorsing the claims of an informant who makes positive rather than negative evaluations, make positive judgments more easily than negative judgments when evaluating others, and overgeneralize others' positive characteristics (Boseovski & Lee, 2006; Boseovski et al., 2017; Marble & Boseovski, 2020; see Boseovski, 2010). Thus, 5-year-olds may be biased toward a positive interpretation of ambiguous comments even when the speaker's past comments suggest otherwise.

Five-year-olds may also struggle to interpret the same evaluative expressions differently based on differences in speakers' evaluative histories because doing so might require understanding that people can say one thing but mean something else (nonliteral speech) and understanding that people can differ in their evaluative judgments because they construe the same phenomena differently (interpretive theory of mind). These two insights develop after age 5. Past work showed that 5-year-olds struggle with understanding irony, differentiating literal from intended meaning (Filippova & Astington, 2008). Understanding that different people can have different interpretations regarding the same phenomena develops around age 7 (Carpendale & Chandler, 1996; Pillow, 2012; see also Osterhaus & Koerber, 2023).

Study 1

In Study 1, we manipulated speakers' evaluative history across three within-subjects conditions. In the *always-likes* condition, the character had a history of making positive evaluations about food items in boxes. In the *never-likes* condition, the character had a history of making negative evaluations about food items in the past. In the *baseline* condition, the participants did not know the character's history. All characters said "It's okay" for the food in the target box. The participants rated the character's liking of the target and stated whether the character would want another target box. If children and adults consider the characters' evaluative history, we expected that the participants would (a) give higher ratings for the never-likes speaker's liking of the target food compared to the always-likes and the baseline and (b) be more likely to state that the character would want another target box in the never-likes condition compared to the always-likes and baseline conditions. If they do not consider speakers' past evaluations, we hypothesized that they would interpret "It's okay" similarly across three conditions.

Method

Participants

Our sample included 55 5-year-old (M = 5.44 [years], range = 5.02–6.02; according to parent reports, 28 were assigned female at birth and 27 were assigned male at birth), 54 7-year-old (M = 7.44 [years], range = 7.01–7.93; 27 were assigned female at birth, 27 were assigned male at birth), and 54 adult (M = 35.54 [years], range = 20–70; 30 identified their sex as female and 24 as male) English speakers based in Canada (N_{total} = 163). Among children whose parents indicated their background (101 out of 109), 41.6% were White (Caucasian), 20.8% were indicated as "mixed," 13.9% were South Asian, 8.9% were Chinese, 4% were Black, 3% were Arab/West Asian, 2% were indicated as Jewish, 1% were Japanese, 1% were Latin American, 1% were Filipino, 1% were Indo Caribbean, and 2% selected "other." Among adults, 57.4% were White, 31.5% were Asian, 3.7% were Black, and 7.4% indicated "other." We determined the sample size prior to data collection based on a simulation-based power analysis (see preregistration). We excluded one additional 5-year-old's data due to parental interference and one 7-year-old's data due to connection issues. We excluded three adults' data due to their long completion time. We recruited children through the university database and adults via Prolific.

Materials

The warm-up included images of three squares in varying sizes and colors. In the testing trials, the participants were shown videos, including the images of the characters with prerecorded voices, seven unique boxes on the tables, and animations (three experimental characters and one storyteller describing the task). Accompanying characters' evaluations (e.g., "It's good/bad/okay"), participants saw thumbs-up (green), thumbs-down (red), and thumbs in the middle (yellow); see Figure 1. These thumbs remained above each box to act as a reminder of the speaker's past statements when participants answered the test questions. The characters' faces were covered with hats, so the participants could only focus on the verbal expressions. We kept the voice recordings as neutral as possible in terms of perceived valence, as an experimental control, and counterbalanced the prerecorded voices of the speakers across conditions. We used a 5-point hand gesture scale indicating the liking: *not at all*, *a little*, *some*, *a lot*, and *very very much* (see also Evans & Jirout, 2023). The participants saw the stimuli on Qualtrics.

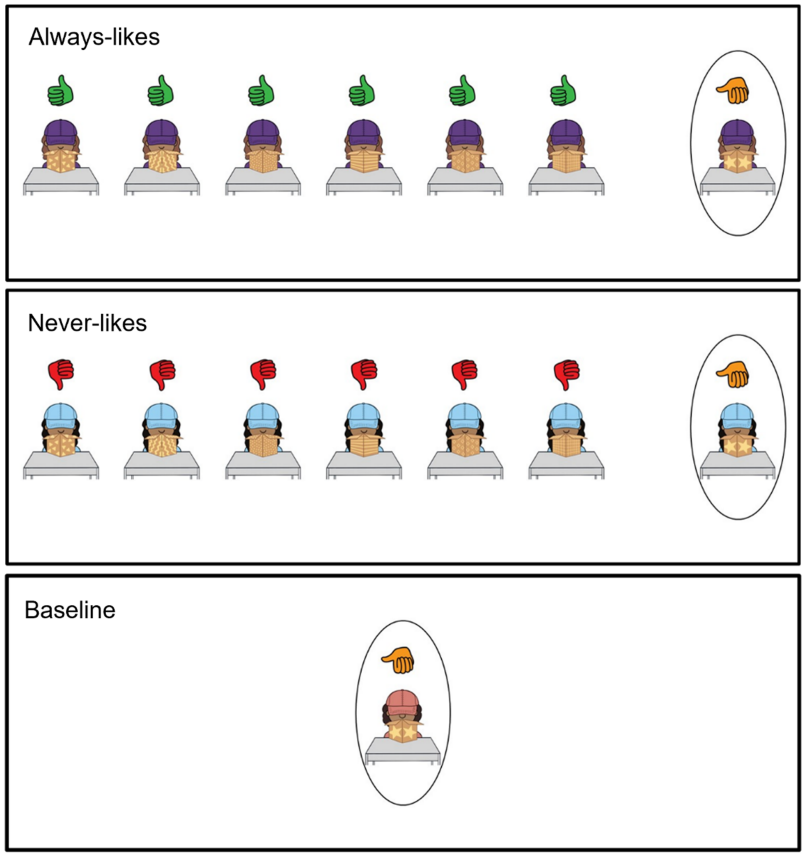
Procedure

Adults completed the experiment on their own. Children attended one-to-one Zoom meetings with the experimenter (E). E shared her screen with the participant. After the screen setup, the study started with a warm-up trial. The participants saw three squares (small, medium, and big) on the screen and were asked which was the biggest, which was the smallest, and which was in the middle. They were also asked which one they liked the most, which they liked the least, and which they liked in the middle, to familiarize them with the preference questions they saw in the experimental trials. After the warm-up, E described the liking scale to the participants. Then, the experimental trials began. The storyteller introduced the character of the trial and said, "This is [character's name]. Now, she will receive some boxes. In each box, there will be a different food. We will not see the food inside the boxes, but she will eat them and tell us what she thinks." Then, the participants watched the animations of the character receiving a box on a table and made a comment along with the storyteller's description: storyteller: "She received a box, ate the food inside, and said that," the character: "It's good/bad/okay." We had three within-subjects conditions. In the *always-likes*[1] condition, the character said, "It's good" for each of the six boxes she received one by one. In the *never-likes* condition, the character said, "It's bad" for the six boxes. Critically, for the seventh box, that is, the target—the box with the stars—in both conditions, the characters said, "It's okay." In the *baseline* condition, the character received only the target box and said, "It's okay" (see Figure 1). The patterns on the boxes were different for each box. The never-likes and the always-likes characters received the same patterned boxes in the same order to make it clear that the characters are receiving similar food items but making different evaluations (except for the target box). At the end of each condition, E asked, "How much do you think she liked the box with the stars? *Not at all* (0), *a little* (1), *some* (2), *a lot* (3), or *very very much* (4)?" Then, E asked, "Do you think she would want another box with the stars? Yes, or no?" At the end of the experiment, E asked the participants to rank the characters based on how much

[1] The condition names were slightly different in the preregistration: "Green" corresponds to the "always-likes," and "red" corresponds to the "never-likes" conditions.

Figure 1

Summary Depiction of the Three Conditions Each Participant Went Through: Always-Likes, Never-Likes, and Baseline



Note. The characters received one food at a time and indicated their liking with the thumbs up/down/middle accompanied by a prerecorded voice ("good/bad/okay"). The order of the conditions was randomized across the participants. See the online article for the color version of this figure.

they liked the food in the final box: "Who do you think liked it the most, who do you think liked it the least, and who do you think liked it in the middle?" As children had difficulty in responding to this question and responded randomly, we moved the analyses of this question to the Supplemental Materials.

Analytic Plan

As we had repeated observations per participant, to understand age and condition differences, we used linear mixed models for the continuous rating question and generalized linear mixed models with a binomial distribution for the binary question on characters' liking, as preregistered. Our full models included the fixed predictors of age (categorical: 5-year-olds, 7-year-olds, and adults), condition (always-likes, never-likes, and baseline), and their interaction; control predictors of trial number (scaled) and sex; and the random intercept for participant ID. We compared this model with a null model that included only the control predictors and the random intercept. All of the full models indicated below improved the fit on the null model ($ps < .001$). For the model comparisons, we used the "drop1" function in R, which allows us to assess the significance of each term by assessing the effect of removing each item from the model.

Transparency and Openness

Study 1's design, hypotheses, and analysis plan were preregistered at the Open Science Framework https://osf.io/btzkv/overview?view_only=aa31acc286ad4d72bfa3f449c811ec75. Our data set, scripts for statistical analyses, and example materials are accessible at https://osf.io/qynu9/overview?view_only=500d0ab685484f5d93a73b221346951b (Ozkan & Ronfard 2026b). We used R (R Development Core Team 4.3.0, R Core Team, 2023) for all analyses across two studies. We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study, and we follow Journal Article Reporting Standards (Appelbaum et al., 2018). We also explain our analytic strategy in the Analytic Plan section.

Results

How Much Do You Think She Liked It?

The response variable was participants' ratings for the characters' liking, ranging from 0 (*not at all*) to 4 (*very very much*). The interaction between condition and age category did not improve the model significantly, as demonstrated by the model comparison we

conducted using the *drop1* function, $F(4, 319) = 0.81$, $p = .52$. The reduced model with the main effects of condition and age category, along with other control predictors and the random intercept, revealed that only the effect of age category was significant,[2] $F(2, 159) = 21.35$, $p < .001$ (see Figure 2). Pairwise comparisons with Tukey adjustment showed that regardless of the conditions, 5-year-olds gave higher ratings compared to 7-year-olds (estimated $M_{\text{difference}} = 0.31$, $SE = 0.11$), $t(159) = 2.95$, $p = .01$, 95% CI [0.06, 0.56], Cohen's $d = 0.45$, and compared to adults (estimated $M_{\text{difference}} = 0.69$, $SE = 0.11$), $t(159) = 6.53$, $p < .001$, 95% CI [0.44, 0.94], Cohen's $d = 1.0$. Seven-year-olds' ratings were also higher than those of adults (estimated $M_{\text{difference}} = 0.38$, $SE = 0.11$), $t(159) = 3.56$, $p = .001$, 95% CI [0.13, 0.63], Cohen's $d = 0.55$.

Do You Think She Would Want Another Box With the Stars?

The response variable was the participants' reports regarding whether the character would want another target box (1) or not (0). We have one missing trial from a 5-year-old. The full model yielded a significant interaction effect of condition and age, $\chi^2(4) = 20.93$, $p < .001$; see Figure 3. Next, we conducted separate analyses for each age group as preregistered. Five-year-olds' responses did not differ across conditions ($p = .275$). However, the condition effect was significant in 7-year-olds ($p = .03$) and adults ($p < .001$). Pairwise comparisons with Tukey adjustment showed that 7-year-olds were less likely to state that the always-likes character would want another target box compared to the never-likes character ($OR = 0.32$, $SE = 0.14$, $z = -2.59$, $p = .026$, 95% CI [0.11, 0.90]). Adults showed the same pattern[3] ($OR = 0.06$, $SE = 0.04$, $z = -4.38$, $p < .001$, 95% CI [0.01, 0.26]). Adults were also more likely to infer that the never-likes character would want the target box compared to the baseline character ($OR = 8.0$, $SE = 4.02$, $z = 4.14$, $p < .001$, 95% CI [2.46, 26.0]).

We also conducted binomial tests to compare participants' responses to the chance level (0.5), that is, whether their interpretation of the statement was more likely to indicate liking (more than 0.5) or not-liking (less than 0.5). Five-year-olds inferred that the baseline character would want another target box significantly more than expected by chance (37/54 trials, $p = .009$). Seven-year-olds' responses did not significantly differ from chance in any of the conditions ($ps \geq .076$). Adults were significantly less likely than chance to infer that the baseline character would want another target box (7/54 trials, $p < .001$). Adults also thought that the always-likes character did not want another target box (3/54 trials, $p < .001$).

Discussion

Study 1 showed that participants' ratings did not differ significantly across conditions. However, in the binary question, 7-year-olds and adults were significantly more likely to state that the never-likes character would want another target box compared to the always-likes character. It is possible that the quantitative inference, assessing the *magnitude* of one's liking, was harder than making a binary evaluation about one's wanting and future behavior without additional contextual cues like prosody or facial expressions.

Five-year-olds' responses did not differ across conditions, suggesting that they did not make speaker-specific inferences about the speakers' liking and wanting. Why? Five-year-olds' positive interpretations of "it's okay" on both questions could have biased our results. It is possible that 5-year-olds identified the expression as clearly positive rather than ambiguous. Consequently, they did not use the information about the speaker's past evaluative comments despite being able to do so. To test this, in Study 2, we highlighted the ambiguity of the expression by stating that "it's okay" could mean that the character liked it or did not like it.

In Study 2, we did not use the Likert-scale question because it did not yield any condition differences in the first study. Also, instead of asking, "Do you think she would want another box with the stars?" in Study 2, we asked about the character's liking directly: "Do you think she liked it or she did not like it?" This simpler phrasing (relative to Study 1) does not require participants to make the additional inference that "if she liked it, she would want another one."

Study 2

Method

Participants

Our sample included 203 participants. None of the participants from Study 1 participated in Study 2. We tested four age groups: 53 5-year-olds ($M = 5.46$ [in years], range = 5.03–5.98; according to parent reports, 24 were assigned female and 29 as male at birth), 50 6-year-olds ($M = 6.57$, range = 6.03–7.11; 25 were assigned female at birth, 25 were assigned male at birth), 50 7-year-olds ($M = 7.61$, range = 7.04–7.98; 28 were assigned female at birth, 22 were assigned male at birth), and 50 adults ($M = 37.24$, range = 18–73; 26 identified their sex as female and 24 as male). We determined the sample size prior to the data collection via a simulation-based power analysis based on the first study's results (see preregistration for the details). The participants were English speakers based in Canada. Among children whose parents indicated their background (143 out of 153), 32.9% were South Asian, 28% were White (Caucasian), 12.6% indicated as mixed, 9.8% were Chinese, 4.2% were Latin American, 4.2% were Black, 1.4% were Japanese, 1.4% were Arab/West Asian, 0.7% were Filipino, 0.7% were Middle Eastern, 0.7% were Korean, and 3.5% indicated "other." Among adults who indicated their background (49 out of 50), 49% were White, 33% were Asian, 10% were Black, 6% of them indicated "mixed," and 2% indicated "other." We excluded an additional three children's data due to technical issues (i.e., connection and camera/mic problems) and two adults' data due to their completion time.

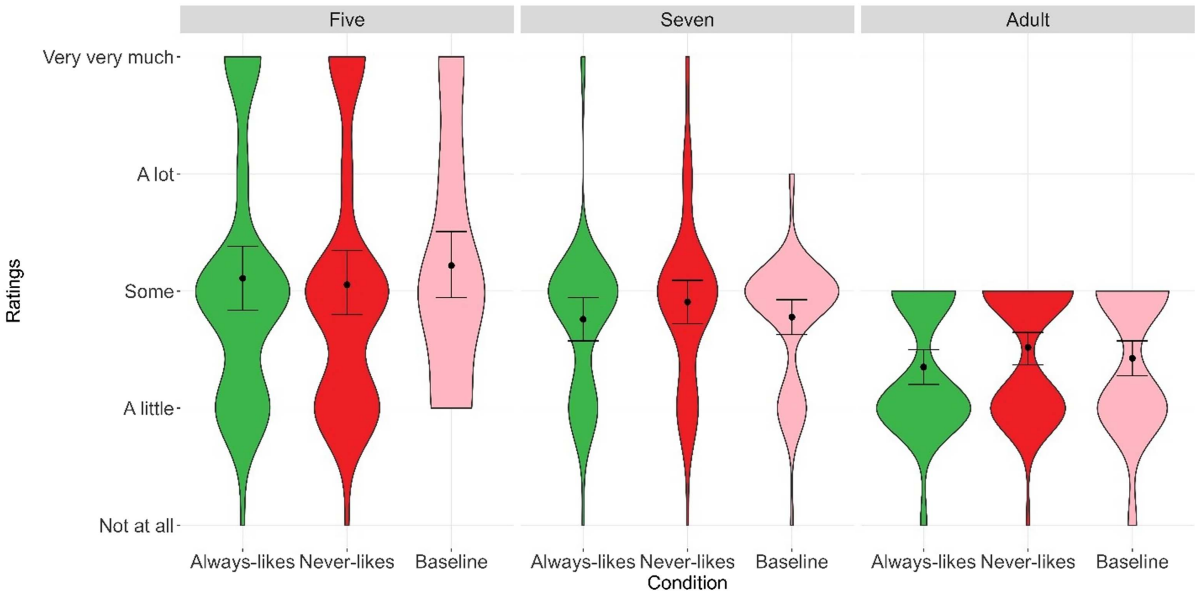
Procedure

The procedure was identical to Study 1 except that (a) participants went through two conditions only: always-likes and never-likes, and (b) at the end of each trial, E only asked a binary liking question, highlighting the ambiguity of the expression:

[2] We also fitted an ordinal model (a cumulative link mixed model) to check the robustness of our results in the scale question. Similar to the linear mixed model, this model only yielded a significant effect of age category ($p < .001$). The only difference in the pairwise comparison results was that the difference between 5-year-olds and 7-year-olds was not significant in the cumulative link mixed model ($p = .18$).

[3] For the adults' model, we removed the random intercept due to singular fit and conducted a binomial regression (a generalized linear model) with the same remaining predictors.

Figure 2
Condition Means for the Rating Question Across Age Groups



Note. 0 = *not at all*, 1 = *a little*, 2 = *some*, 3 = *a lot*, 4 = *very very much*. The error bars represent the 95% confidence intervals. See the online article for the color version of this figure.

> Remember the box with the stars. [character's name] said, "It's okay," when she ate the food in that box. I am not sure what she meant when she said that. It could mean that she liked it, or it could mean that she did not like it. Do you think she liked it, or she did not like it?

We removed the baseline condition in Study 2 for two reasons. First, by explicitly telling participants that "it's okay" can be interpreted as positive or negative, we control for participants' default interpretation of the meaning of that phrase. The control condition of Study 1 was designed to capture that default interpretation and thus provide a contrast to the other two conditions. This is no longer needed with our rephrasing. Second, because we tell participants that "it's okay" can be interpreted in two ways, a control condition in which participants have no evaluative history to draw from would be expected to yield 50% of responses interpreting "it's okay" as liking and 50% as disliking. This would be expected because participants would have no other information other than our statement to determine the meaning.

Analytic Plan

Our preregistered generalized linear mixed model yielded a singular fit due to very low variance (near 0) in the random-effect structure, and we had only two observations per participant. Thus, to test condition and age differences in participants' responses, we had to remove the random effect of participant ID and instead used generalized linear models with a binary distribution. We compared the full models with the null models. The null models included only the control predictors of participants' sex and trial number. The full models reported below significantly improved the fit on the null models ($ps < .001$).

Transparency and Openness

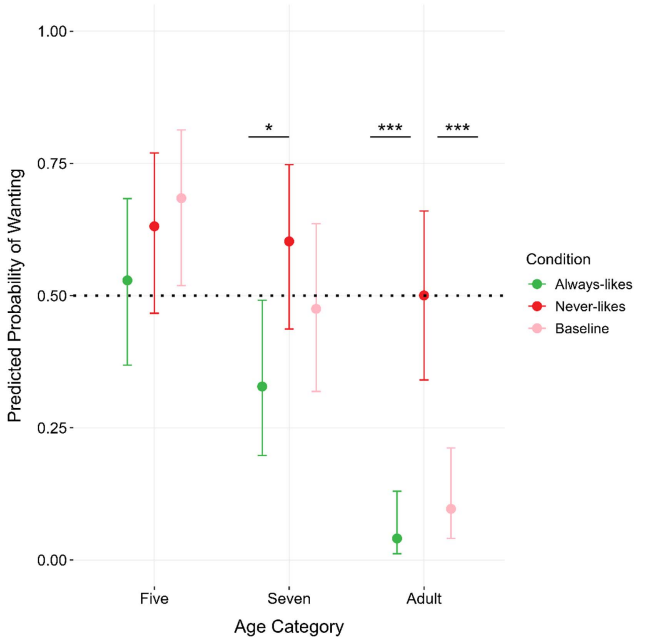
Study 2's design, hypotheses, and analysis plan were preregistered at the Open Science Framework https://osf.io/unfsm/overview?view_only=ef1a4e835822470c8b229f89c25297dd. Our data set and scripts for statistical analyses are accessible at https://osf.io/gtrhz/overview?view_only=26eaa66a8a7d41ad8588ff143ce0bf54 (Ozkan & Ronfard 2026a). We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study, and we follow Journal Article Reporting Standards (Appelbaum et al., 2018).

Results

The response variable was whether participants stated that the character liked the target food or did not like it. Our full model, including the interaction between condition and age category (5, 6, 7, and adults) along with the control predictors of sex and trial number, did not yield a significant interaction effect, $\chi^2(3) = 3.05$, $p = .38$. The reduced model yielded significant main effects of condition, $\chi^2(1) = 55.77$, $p < .001$, and age category, $\chi^2(3) = 10.64$, $p = .014$. Participants were more likely to report that the character liked the food in the never-likes condition compared to the always-likes condition ($OR = 4.93$, $SE = 1.1$, $z = 7.15$, $p < .001$, 95% CI [3.18, 7.63]); see Figure 4. Tukey-adjusted pairwise comparisons between age groups showed that 5-year-olds were more likely to think that the characters liked the food compared to 7-year-olds ($OR = 2.28$, $SE = 0.72$, $z = 2.63$, $p = .043$, 95% CI [1.02, 5.12]) and compared to adults ($OR = 2.51$, $SE = 0.79$, $z = 2.93$, $p = .018$, 95% CI [1.12, 5.63]). Other comparisons were not significant. Binomial tests for each age group showed that while children's responses were not significantly different from the chance level (0.5) in the

Figure 3

Predicted Probability of Liking by Age and Condition (Study 1)



Note. The figure shows the predicted probability of participants responding to the wanting question as "Yes, she would want another target box" across conditions and age groups. Dots represent the means. The error bars indicate the 95% confidence intervals. See the online article for the color version of this figure.
* $p < .05$. *** $p < .001$.

always-likes condition ($ps \geq .06$), adults were significantly less likely than chance to think that the character liked the food (12/50 trials, $p < .001$). However, 5-year-olds (44/53 trials), 6-year-olds (37/50), 7-year-olds (34/50), and adults (38/50) stated that the character liked the food in the never-likes condition above the chance level ($ps \leq .015$). Additionally, as preregistered, we fitted a model examining the effect of age continuously in children's data, after centering age. This model yielded significant main effects of condition, $\chi^2(1) = 31{,}07$, $p < .001$, and age, $\chi^2(1) = 5{,}59$, $p = .018$. With increasing age, children were less likely to think that the characters liked the food ($OR = 0.73$, $SE = 0.14$, $z = -2.34$, $p = .019$, 95% CI [0.55, 0.95]). The condition effect mimics the previous analysis.

Discussion

The results of Study 2 showed that children aged 5–7 and adults were more likely to state that the never-likes character liked the target item compared to the always-likes character. By implication, when told that "it's okay" is ambiguous because it can be interpreted in two ways, 5-year-olds make different inferences about its meaning based on differences in speakers' evaluative histories. This suggests that 5-year-olds in Study 1 interpreted "it's okay" similarly across conditions because, unlike older participants, they did not detect its ambiguity and did not require clarification. We also replicated the finding from Study 1 that, with increasing age, participants were less likely to interpret "it's okay" as positive.

General Discussion

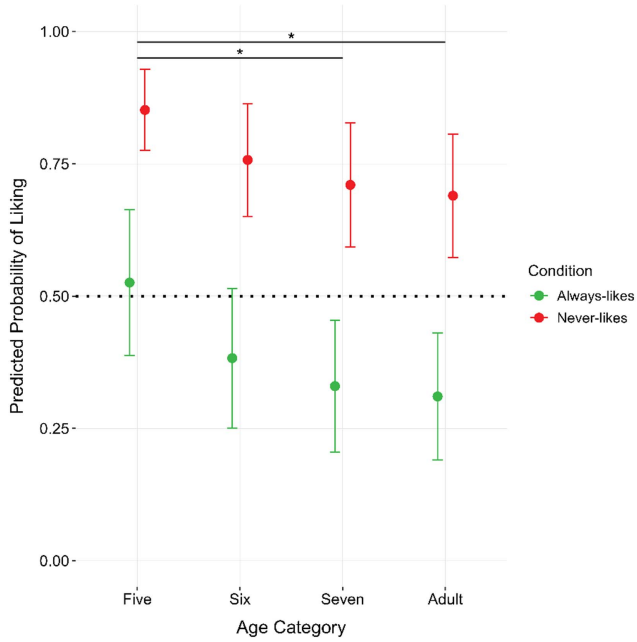
Accurately interpreting what people mean when they make evaluative judgments is essential for navigating daily interactions. In a workplace where colleagues rarely give praise, hearing that your work is "okay" signals genuine approval; in a more effusive environment, the same word suggests mediocrity. Similarly, when choosing a movie with a friend, understanding their evaluative patterns—what they have praised or criticized before—helps you gauge their preferences. These interpretive skills are fundamental to pragmatic reasoning in daily communications and support learning from others.

Across two studies, we examined how children and adults use speakers' past comments to make sense of those speakers' evaluative statements. Study 1 showed that when two speakers both described a target food as "okay," 7-year-olds and adults inferred that the speaker with a negative evaluative history (never-likes) liked it more than the typically positive speaker (always-likes). Five-year-old children attributed similar preferences to those two speakers. A possible explanation is that 5-year-olds interpreted "it's okay" as meaning that the speaker liked the food. If 5-year-old children did not think "it's okay" was ambiguous, they did not need to draw on the speaker's past evaluative comments to interpret that statement.

We tested this possibility in Study 2. We explicitly told participants that "it's okay" could mean the character liked or disliked the item. The results of Study 2 showed that children aged 5–7 and adults were more likely to state that the never-likes character liked the target item more than the always-likes character. Thus, when told that the expression was ambiguous, children as young as 5 made

Figure 4

Predicted Probability of Liking by Age and Condition (Study 2)



Note. The figure shows the predicted probabilities of participants' judgments regarding whether the character liked (1) or did not like (0) the target item across conditions and age groups. Error bars represent the 95% confidence intervals. See the online article for the color version of this figure.
* $p < .05$.

different inferences about the meaning of an ambiguous evaluative statement depending on the evaluative histories of the speakers.

Note that our task required more sophisticated reasoning than simply predicting whether a character who always/never likes things will like the next item. Such a straightforward prediction would yield: If they liked everything before, they will like this too. Instead, participants faced an inverse inference problem: Given an ambiguous evaluation ("it's okay"), infer the speaker's true preference by considering their evaluative history. When the speaker with a negative evaluative history says "it's okay," this is relatively positive for them. Thus, participants needed to recognize that the same words carry different meanings depending on the speaker's past comments. In sum, Study 2 demonstrates that even 5-year-olds can calibrate ambiguous evaluations against speaker-specific baselines to infer mental states.

This finding aligns with past work showing that although young children use ambiguous referents to talk about objects, they clarify ambiguity when they are provided with feedback, for example, "Which one? Do you need the girl *eating* or the girl *reading*?" (Matthews et al., 2012), or when they are provided with a motivation to reduce the ambiguity, such as using requestive speech to ask for an object from someone (Bahtiyar & Küntay, 2009). Past work also suggests children's early competence in making speaker-specific inferences to resolve referential ambiguity. For example, Bohn, Le, et al. (2021) showed that after hearing a speaker who repeatedly referred to items from a specific category, for example, vehicles, 4-year-old children inferred that when that particular speaker, but not another speaker, said, "Can you touch *it*," that speaker meant touching a vehicle.

Building on these results, our study suggests that while preschool children can use a speaker's history to make sense of ambiguous statements, they struggle to do so with ambiguous *evaluative* judgments. Our results suggest that this difficulty might be due to children's tendency to interpret a potentially ambiguous statement like "it's okay" as positive—a finding consistent with past work on childhood positivity bias (Boseovski, 2010; see also Boseovski et al., 2017; Marble & Boseovski, 2020). Indeed, 5-year-old children judged that the always-likes and the never-likes speakers liked the food in the target box more often than 7-year-old children and adults (see the Supplemental Materials).

However, such positivity bias may not be the only reason 5-year-olds struggled to identify "it's okay" as ambiguous. Their difficulty may also stem from two developing capacities: understanding that words can convey nonliteral meanings—an awareness that increases between ages 5 and 7 (Filippova & Astington, 2008)—and recognizing that people can interpret the same thing differently (Carpendale & Chandler, 1996; Pillow, 2012; see also Osterhaus & Koerber, 2023). Both insights enable children to recognize that identical evaluative statements can carry different meanings depending on the speaker's perspective and communicative intent. For instance, the age-related decrease in interpreting "it's okay" positively may reflect growing awareness that speakers sometimes use mild language to politely soften negative evaluations, rather than to express genuine approval.

Past work shows that how people express politeness differs across languages and cultures (Hickey & Stewart, 2005; see also Yoon et al., 2020, for the motivational factors). In our study, we examined the speaker-specific inferences by manipulating the individual's past evaluations; however, another dimension of examining resolving the ambiguity of evaluative expressions is to focus on broader contextual dynamics, for example, group norms and cross-cultural examinations. Future work might examine whether and how different cultural practices affect people's inferences about other people's beliefs and desires based on their evaluative testimony. We expect that in environments where it is common to be overpolite and overpraising, people might interpret ambiguous comments like "it's ok" negatively. In contrast, in environments where people make more neutral or negative comments, people might attribute a more positive meaning to ambiguous evaluative testimony.

Our work also relates to the social learning literature. Extending the past work that showed that children use a generic informant's past accuracy and dispositions to make epistemic and social inferences (Bhatti et al., 2024; Gweon, 2021; Gweon et al., 2014; Koenig & Harris, 2005; Landrum et al., 2015; Pasquini et al., 2007; Ronfard & Lane, 2018, 2019; Sobel & Kushnir, 2013; Tong et al., 2020), we showed that young children also use an informant's past subjective judgments to disambiguate the meaning of an ambiguous evaluative statement for that particular informant. In addition to revealing others' mental states, in some contexts, these inferences also provide information about the object of evaluation. For example, if your friend with a negative history of evaluations liked a particular movie, you might infer that this is a movie worth watching. Conversely, you might hesitate about watching a movie that your *always-likes* friend suggests. Further research might explore the development of such inferences.

In our study, we recorded the characters' utterances of "It's okay/bad/good" in a neutral tone to control for any effect of prosody. We also covered the characters' faces with a hat and used static images. However, in daily conversations, cues such as prosody and gestures (Hinnell & Parrill, 2020; Snedeker & Trueswell, 2003; Wharton, 2012), as well as facial expressions, provide rich cues for resolving ambiguity. Further research might examine how children and adults use these cues, along with speakers' past utterances, to resolve ambiguous evaluative statements.

In our study, speakers' past statements included only positive or negative evaluations, other than the "okay" evaluation they provided at the end. Future research might examine listeners' probabilistic inferences about the speaker's ambiguous evaluations by introducing greater variability in the speaker's past statements, for example, *sometimes* offering positive or negative evaluations (see Bhatti et al., 2024).

The baseline condition in Study 1 was designed to mimic a situation in which we do not know a speaker's history. For this reason, the baseline informant received only the target box and provided an ambiguous evaluation of the food in the target box. It is possible that when we do not know the history of a speaker versus when we know the history of a speaker, it makes a difference in how confident we are in our interpretation, which would not be reflected in our measures regarding whether we interpret the utterance negatively or positively. Further research might also examine how speakers' history affects listeners' resolution of ambiguity in relation to how certain they feel about their interpretation.

Although we focused on the ambiguous expression "OK" in the present study, in daily life, we encounter many instances in which people's positive evaluative comments are also interpreted relative to their evaluative baseline. For example, an expression such as "good" might reflect less liking for a person who typically uses more intense positive adjectives, for example, "great" or "excellent," compared to

someone who uses less intense adjectives, for example, "okay." Thus, the present study contributes to our understanding of the development of interpreting evaluative statements by making speaker-specific inferences.

Conclusion

Our study showed that recovering the true mental states of speakers based on their ambiguous evaluations is complex, and with increasing age, children get better at making such inferences by using a specific speaker's past: While 7-year-olds made those inferences naturally, 5-year-olds needed to be informed about the ambiguity of the statement. This development appears to be related to younger children's tendency to attribute a positive meaning to an ambiguous evaluative statement, which might eliminate the need to consider speakers' past.

References

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA publications and communications board task force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Asaba, M., Hembacher, E., Qiu, S., Anderson, B., Frank, M., & Gweon, H. (2018). Young children use statistical evidence to infer the informativeness of praise. *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 40). Cognitive Science Society. https://escholarship.org/uc/item/3sb9g402

Bahtiyar, S., & Küntay, A. C. (2009). Integration of communicative partner's visual perspective in patterns of referential requests. *Journal of Child Language*, *36*(3), 529–555. https://doi.org/10.1017/S0305000908009094

Bhatti, D., Lane, J. D., & Ronfard, S. (2024). Updating trust: How children combine trait information with prior accuracy as they interact with an informant. *Developmental Psychology*, *60*(6), 1145–1160. https://doi.org/10.1037/dev0001731

Bleijlevens, N., Contier, F., & Behne, T. (2023). Pragmatics aid referent disambiguation and word learning in young children and adults. *Developmental Science*, *26*(4), Article e13363. https://doi.org/10.1111/desc.13363

Bohn, M., & Köymen, B. (2018). Common ground and development. *Child Development Perspectives*, *12*(2), 104–108. https://doi.org/10.1111/cdep.12269

Bohn, M., Le, K. N., Peloquin, B., Köymen, B., & Frank, M. C. (2021). Children's interpretation of ambiguous pronouns based on prior discourse. *Developmental Science*, *24*(3), Article e13049. https://doi.org/10.1111/desc.13049

Bohn, M., Tessler, M. H., Merrick, M., & Frank, M. C. (2021). How young children integrate information sources to infer the meaning of words. *Nature Human Behaviour*, *5*(8), 1046–1054. https://doi.org/10.1038/s41562-021-01145-1

Boseovski, J. J. (2010). Evidence for "rose-colored glasses": An examination of the positivity bias in young children's personality judgments. *Child Development Perspectives*, *4*(3), 212–218. https://doi.org/10.1111/j.1750-8606.2010.00149.x

Boseovski, J. J., & Lee, K. (2006). Children's use of frequency information for trait categorization and behavioral prediction. *Developmental Psychology*, *42*(3), 500–513. https://doi.org/10.1037/0012-1649.42.3.500

Boseovski, J. J., Marble, K. E., & Hughes, C. (2017). Role of expertise, consensus, and informational valence in children's performance judgments. *Social Development*, *26*(3), 445–465. https://doi.org/10.1111/sode.12205

Carpendale, J. I., & Chandler, M. J. (1996). On the distinction between false belief understanding and subscribing to an interpretive theory of mind. *Child Development*, *67*(4), 1686–1706. https://doi.org/10.1111/j.1467-8624.1996.tb01821.x

Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, *9*, 519–540. https://doi.org/10.1146/annurev-linguistics-031220-010811

Evans, N. S., & Jirout, J. J. (2023). Investigating the relation between curiosity and creativity. *Journal of Creativity*, *33*(1), Article 100038. https://doi.org/10.1016/j.yjoc.2022.100038

Filippova, E., & Astington, J. W. (2008). Further development in social reasoning revealed in discourse irony understanding. *Child Development*, *79*(1), 126–138. https://doi.org/10.1111/j.1467-8624.2007.01115.x

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Grice, H. P. (1975). Logic and conversation. In P. Cole, & J. Morgan (Eds.), *Syntax and semantics* (Vol. 3 pp. 41–58). Academic Press. https://doi.org/10.1163/9789004368811_003

Gweon, H. (2021). Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences*, *25*(10), 896–910. https://doi.org/10.1016/j.tics.2021.07.008

Gweon, H., Pelton, H., Konopka, J. A., & Schulz, L. E. (2014). Sins of omission: Children selectively explore when teachers are under-informative. *Cognition*, *132*(3), 335–341. https://doi.org/10.1016/j.cognition.2014.04.013

Gweon, H., Shafto, P., & Schulz, L. (2018). Development of children's sensitivity to over informativeness in learning and teaching. *Developmental Psychology*, *54*(11), 2113–2125. https://doi.org/10.1037/dev0000580

Hickey, L., & Stewart, M. (Eds.). (2005). *Politeness in Europe* (Vol. 127). Multilingual Matters. https://doi.org/10.14198/raei.2006.19.23-3

Hinnell, J., & Parrill, F. (2020). Gesture influences resolution of ambiguous statements of neutral and moral preferences. *Frontiers in Psychology*, *11*, Article 587129. https://doi.org/10.3389/fpsyg.2020.587129

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015). Children's understanding of the costs and rewards underlying rational action. *Cognition*, *140*, 14–23. https://doi.org/10.1016/j.cognition.2015.03.006

Koenig, M. A., & Harris, P. L. (2005). Preschoolers mistrust ignorant and inaccurate speakers. *Child Development*, *76*(6), 1261–1277. https://doi.org/10.1111/j.1467-8624.2005.00849.x

Landrum, A. R., Eaves, B. S., & Shafto, P. (2015). Learning to trust and trusting to learn: A theoretical framework. *Trends in Cognitive Sciences*, *19*(3), 109–111. https://doi.org/10.1016/j.tics.2014.12.007

Marble, K. E., & Boseovski, J. J. (2020). Content counts: A trait and moral reasoning framework for children's selective social learning. *Advances in Child Development and Behavior*, *58*, 95–136. https://doi.org/10.1016/bs.acdb.2020.01.004

Matthews, D., Butcher, J., Lieven, E., & Tomasello, M. (2012). Two- and four-year-olds learn to adapt referring expressions to context: Effects of distracters and feedback on referential communication. *Topics in Cognitive Science*, *4*(2), 184–210. https://doi.org/10.1111/j.1756-8765.2012.01181.x

Osterhaus, C., & Koerber, S. (2023). The complex associations between scientific reasoning and advanced theory of mind. *Child Development*, *94*(1), e18–e42. https://doi.org/10.1111/cdev.13860

Ozkan, F., & Ronfard, S. (2026a, March 9). *Evaluative Information Study 2: Children's developing ability to use a speaker's past comments to make speaker-specific inferences*. https://osf.io/gtrhz

Ozkan, F., & Ronfard, S. (2026b, March 9). *How do children and adults use a person's evaluative history to constrain their inferences about that person's evaluations?* https://osf.io/qynu9

Pasquini, E. S., Corriveau, K. H., Koenig, M., & Harris, P. L. (2007). Preschoolers monitor the relative accuracy of informants. *Developmental Psychology*, *43*(5), Article 1216. https://doi.org/10.1037/0012-1649.43.5.1216

Pillow, B. H. (2012). Conceptual knowledge about cognitive activities. *Children's discovery of the active mind* (pp. 13–44). Springer. https://doi.org/10.1007/978-1-4614-2248-8_2

R Core Team. (2023). *R (4.3.1): A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org

Ronfard, S., & Lane, J. D. (2018). Preschoolers continually adjust their epistemic trust based on an informant's ongoing accuracy. *Child Development*, *89*(2), 414–429. https://doi.org/10.1111/cdev.12720

Ronfard, S., & Lane, J. D. (2019). Children's and adults' epistemic trust in and impressions of inaccurate informants. *Journal of Experimental Child Psychology*, *188*, Article 104662. https://doi.org/10.1016/j.jecp.2019.104662

Schuster, S., & Degen, J. (2020). I know what you're probably going to say: Listener adaptation to variable use of uncertainty expressions. *Cognition*, *203*, Article 104285. https://doi.org/10.1016/j.cognition.2020.104285

Shafto, P., Goodman, N. D., & Frank, M. C. (2012). Learning from others: The consequences of psychological reasoning for human learning. *Perspectives on Psychological Science*, *7*(4), 341–351. https://doi.org/10.1177/1745691612448481

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89. https://doi.org/10.1016/j.cogpsych.2013.12.004

Snedeker, J., & Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language*, *48*(1), 103–130. https://doi.org/10.1016/S0749-596X(02)00519-3

Sobel, D. M., & Kushnir, T. (2013). Knowledge matters: How children evaluate the reliability of testimony as a process of rational inference. *Psychological Review*, *120*(4), 779–797. https://doi.org/10.1037/a0034191

Stiller, A. J., Goodman, N. D., & Frank, M. C. (2015). Ad-hoc implicature in preschool children. *Language Learning and Development*, *11*(2), 176–190. https://doi.org/10.1080/15475441.2014.927328

Sullivan, J., & Barner, D. (2016). Discourse bootstrapping: Preschoolers use linguistic discourse to learn new words. *Developmental Science*, *19*(1), 63–75. https://doi.org/10.1111/desc.12289

Tong, Y., Wang, F., & Danovitch, J. (2020). The role of epistemic and social characteristics in children's selective trust: Three meta-analyses. *Developmental Science*, *23*(2), Article e12895. https://doi.org/10.1111/desc.12895

Vasil, J. (2023). A new look at young children's referential informativeness. *Perspectives on Psychological Science*, *18*(3), 624–648. https://doi.org/10.1177/17456916221112072

Wellman, H. M. (2014). *Making minds: How theory of mind develops*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199334919.001.0001

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, *75*(2), 523–541. https://doi.org/10.1111/j.1467-8624.2004.00691.x

Wharton, T. (2012). Pragmatics and prosody. In K. Allan & K. M. Jaszczolt (Eds.), *The Cambridge handbook of pragmatics* (pp. 567–584). Cambridge University Press. https://doi.org/10.1017/CBO9781139022453.031

Wu, Y., & Schulz, L. E. (2018). Inferring beliefs and desires from emotional reactions to anticipated and observed events. *Child Development*, *89*(2), 649–662. https://doi.org/10.1111/cdev.12759

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2020). Polite speech emerges from competing social goals. *Open Mind*, *4*, 71–87. https://doi.org/10.1162/opmi_a_00035

Received July 23, 2025
Revision received December 6, 2025
Accepted March 7, 2026 ■