# Title

Thought and Language: Effects of Groupmindedness on Young Children's Interpretation of Exclusive *We*

## Authors

Jared Vasil[1*], Dayna Price[1], & Michael Tomasello[1,2]

## Affiliations

[1]Department of Psychology and Neuroscience, Duke University, Durham, NC, USA; [2]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

* Corresponding author, jared.vasil@duke.edu

## Abstract

The current study investigated whether age-related changes in conceptualization of social groups influences interpretation of the pronoun *we*. Sixty-four 2- and 4-year-olds (*N* = 29 female, 49 White-identifying) viewed scenarios in which it was ambiguous how many puppets performed an activity together. When asked who performed the activity, a speaker puppet responded, "We did!" In one condition, the speaker was near one and distant from another puppet, implying a dyadic interpretation of *we*. In another condition, the speaker was distant from both, thus pulling for a group interpretation. In the former condition, 2- and 4-year-olds favored the dyadic interpretation. In the latter condition, only 4-year-olds favored the group interpretation. Age-related conceptual development "expands" the set of conceivable plural person referents.

## Introduction

Before their third birthday, children participate in dyadic activities with partners. In these activities, children conceive of themselves and partners as a joint agent "we" in which roles are coordinated to attain shared goals (Tomasello, 2019). Then, at around their third birthday, thinking and behavior undergo a shift with the emergence of groupminded thinking and its behavioral concomitants, like joint commitment (Gräfenhain et al., 2009) and norm enforcement (Schmidt et al., 2012). This groupminded turn causes children to conceive of themselves and specific others as part of a larger group "we." Does the emergence of groupmindedness influence language use?

Language functions to align partners' attention, chiefly through the use of *referring expressions*. Referring expressions are appropriately produced by speakers as young as 2 years old to invite joint attention towards *intended referents* (Vasil, 2023). In turn, listeners as young as 1 year old can appropriately interpret referring expressions (Liszkowski, 2018). Importantly, conceptual skills undergird every conceptualizable – and, thus, interpretable – referent (Langacker, 1987). In ontogeny, the implication is that relevant conceptual undergirding must be established before it can be exploited as a tool for referential interpretation (Bruner, 1983).

As one example, the emergence of groupminded thinking at 3 years of age may influence children's interpretation of *we*. Specifically, 2-year-olds may interpret *we* narrowly to include only the speaker and herself. Two-year-olds would thereby make 2-person "dyadic interpretations" of *we*. In contrast, 4-year-olds may leverage their groupminded conceptual structure to interpret *we* more flexibly. Specifically, 4-year-olds may more readily make 3-person "group interpretations" than 2-year-olds, while retaining the ability to form dyadic interpretations, as appropriate.

This developmental hypothesis gains initial credence from prior research that suggests that 2-year-olds can appropriately comprehend (Bohn et al., 2020; Girouard et al., 1997) and produce (reviewed in Vasil, 2023) a range of personal pronouns (i.e., words like *I*, *me*, *you*, *us*, etc.). Moreover, the ability to coordinate two and three visual perspectives precedes 2-year-olds' acquisition of singular first- and second-person pronouns (e.g., *I* and *you*, respectively) and singular third-person pronouns (e.g., *he*), respectively (Ricard et al., 1999). Thus, by the time they turn 3 years old (i.e., around the age that they undergo the groupminded turn), children appropriately use singular personal pronouns and their use of these forms is related to their visual perspective taking skills. However, almost nothing is known about the acquisition of first-person plural pronouns (e.g., *we*) or whether the groupminded turn influences their use.

Two studies provide insight into children's use of first-person plural forms and another into whether groupmindedness influences use of deictic demonstratives. Examining the corpora of 479 English-speaking 2- to 5-year-olds, Vasil et al. (2023) found that first-person plurals occurred infrequently in children's speech. Nonetheless, the authors' dataset included several thousand first-person plural tokens spontaneously uttered by 2-year-olds (see also Ibbotson et al., 2018). While it is unclear whether first-person plural forms were used appropriately by 2-year-olds in Vasil et al. (2023), those authors found that (i) 2-year-olds use those forms; and (ii) children produced proportionally more first-person plural pronouns after age 3 than before. This latter finding was interpreted as suggesting that the groupminded turn influences the use of first-person plural pronouns. Unfortunately, the observational design of Vasil et al. (2023) precluded causal inference. In related research, Vasil and Tomasello (2022) found that linguistically framing a joint activity with *we* increases 3-year-olds' joint commitment during the activity, compared to framing with *you*. Unfortunately, one cannot be certain about the role of groupmindedness because Vasil and Tomasello's (2022) sample did not bridge the theoretically critical age of 3 years. Another relevant study was conducted by Liebal et al. (2013). In their research, an adult ambiguously requested that 3-year-olds retrieve a toy (using the German equivalent of *that*), one of which was a novel toy and the other a culturally shared entity (e.g., a Santa Claus doll). When speakers appeared familiar with the toy, children interpreted *that* as referring to the shared entity, the one that "we" share in our culture. Unfortunately, one cannot be certain about influences of groupmindedness because Liebal et al.'s (2013) sample did not bridge the theoretically critical age of 3 years. Moreover, Liebal et al. (2013) investigated effects of groupmindedness indirectly, namely, by studying its effects on children's ability to use cultural knowledge to guide referential interpretation. While cultural knowledge presupposes group conceptualization skills (Tomasello, 2019), it also presupposes contingent cultural knowledge (e.g., about Christmas). Stronger evidence would more directly tap group conceptualization skills. Altogether, these studies may suggest that the emergence of groupmindedness influences language use. However, limitations of each study force caution about inferring influences, if any, of groupmindedness on interpretation.

This study remedied the above limitations. We conducted an experimental investigation of young children's interpretation of *we* before and after the theoretically critical age of 3 years. Moreover, when investigating whether the emergence of groupmindedness influences language use, examining children's interpretation of *we* is superior to examining their contingent cultural knowledge. This is because *we* is conventionally associated with a solely first-person plural semantics. In contrast, contingent cultural knowledge is conventionally associated with a first-person plural semantics and further frame-relevant conceptual structure (Fillmore, 1975).

The hypothesis was preregistered that, in appropriately ambiguous contexts, only children who are 3 years of age and older will appropriately make group-level interpretations ([*anonymized*] https://aspredicted.org/blind.php?x=RVL_8ZQ). All data and code are freely available ([*anonymized*] https://osf.io/9hygn/?view_only=c096f9dc90af46a18aba2d76475cf5b3.).

## Method

**Participants**. There were 75 participants. The final sample included 64 participants, 32 2-year-olds (*M* = 2.55 years, *range* = 2.30 – 2.75, 13 female) and 32 4-year-olds (*M* = 4.57 years, *range* = 4.23-4.75, 16 female). Thus, there were 11 participants who were excluded (failed the warmup, *N* = 8; no test trial responses, *N* = 2; outside preregistered age range, *N* = 1). All exclusions were 2-year-olds, save for one 3-year-old who was outside the preregistered age range. Caregivers who earlier indicated interest in participation were randomly emailed. Caregivers were predominantly White-identifying (*N* = 50), Black or African American-identifying (*N* = 5), or Biracial – Black/White-identifying (*N* = 3); with an annual reported income of predominantly ">130,000" (*N* = 31), "100,000-129,999" (*N* = 16), or "60,000-79,999" (*N* = 8). Participants were sampled from *XXX* to *XXX*. Caregivers received a $10 Amazon gift card; participants, a certificate. Study design and procedure were approved by the *XXX* Institutional Review Board (protocol *XXX*).

**Design**. A within-subjects, repeated measures design with four conditions. There were two comprehension and two production conditions. Participants saw each condition twice (i.e., 8 trials per participant). Trial type (comprehension or production first) and condition order (within trial type) were counterbalanced between participants. One experimenter (E) ran 58 and another E 6 participants. Although children produce personal pronouns appropriately (Vasil, 2023), participants rarely produced *we* (Supplementary Figure 1). Thus, production is ignored in this article. Participants' infrequent production of *we* accords with their infrequent production of *we* in other experimental (Orvell et al., 2018) and naturalistic contexts (Vasil et al., 2023).

**Materials**. Three hand puppets (a dog, a cow, and Eeyore), a rectangular box (18 x 10 x 8 in.), a light blue cloth, wooden blocks, a puzzle, and three public domain clips of transitive actions. The cloth was draped around the box to resemble a table.

**Procedure**. Participants were tested via Zoom. The procedure was initially to be administered in lab. However, COVID-19 forced redesigning for Zoom. E greeted caregivers and participants before initiating a Zoom setup phase.

***Zoom setup phase***. E used a PowerPoint to ensure that caregivers set their Zoom session to full screen with self-view hidden and appropriate volume. Caregivers were asked not to talk during the procedure, except to refocus participants' attention.

***Warmup phase***. During two to three warmup trials, participants watched clips of animate actor-inanimate patient actions (e.g., a cow pushing a cart). Next, E asked participants who performed the action (e.g., "who pushed the cart?"). Simultaneously, participants saw a decision screen that displayed three solid-colored squares with a different animal depicted in each. As aid, atop the decision screen was a video still that displayed the patient (e.g., the cart). After E's question, E immediately enumerated the answer choices onscreen from left to right (e.g., "was it cow in the blue square? chicken in the green square?" etc.). Participants indicated their answer. E repeated the question up to two times There was a correct choice in warmup trials. Participants had to correctly respond once to proceed. Participants skipped the third trial if they answered correctly on the first two trials. Forty-four participants were correct on the first two trials. Seven participants were correct on only one trial, although 3 of these children were distracted on at least one trial.

***Test phase***. After the warmup, E introduced three puppets. Participants saw the puppets sitting around a rectangular table (Figure 1A). Two puppets (P1/Eeyore and P2/Cow) were seated together on one side of the table while one puppet (P3/Doggie) was seated on another side. Atop the table was an incomplete toy (blocks or a puzzle). E noted that the toy was incomplete. Then, the screen faded to black (Figure 1B). The completed toy reappeared (Figure 1C). Next, E asked the speaker puppet who built the toy. The speaker puppet responded "We did!"

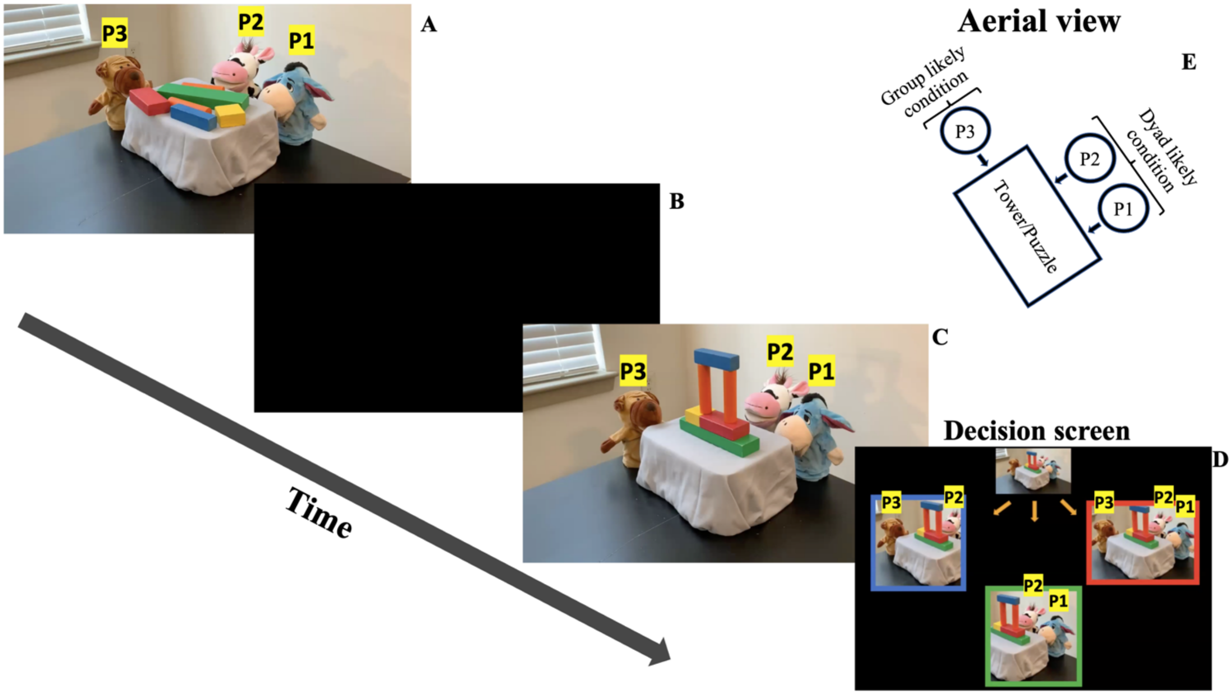

**Figure 1**. The diagonal represents test trial chronology. Highlighted inset text displayed for reader convenience. Aerial view displays conditions.

The manipulation targeted the position of the speaker puppet relative to the other puppets (Figure 1E). P3 was speaker puppet in the group likely condition. P1 or P2 was speaker puppet in the dyad likely condition. For example, in the group likely condition, E asked "Doggie, who built the tower?" and Doggie responded, "We did!" Then, E told participants that E was unsure who built the toy. E asked participants to indicate who built the toy. E then enumerated the choices onscreen from left to right (e.g., "was it Eeyore and Cow in the green square? Cow and Doggie in the green square?" etc.). While E asked this, participants saw a decision screen (Figure 1D). Participants chose one of three interpretations: P1-P2 (dyadic), P2-P3 (dyadic), or P1-P2-P3 (group). As memory aid, atop decision screens was a still of the completed toy and puppets. Also, E reiterated the speaker puppet's name before asking their question. The next trial began after participants indicated their interpretation. There were 4 test trials (two trials per condition).

In the dyad likely condition, participants were predicted to favor dyadic interpretations. In the group likely condition, only 4-year-olds were predicted to switch to group interpretations. The dyad likely condition speaker puppet (P1 or P2) and the position of decision screen choices (left, center, right) were counterbalanced between participants. Tower preceded puzzle trials.

**Coding**. We coded interpretations of *we*. Participants indicated interpretations by pointing to one of the pictures (E sometimes asked caregivers to clarify points) or by referring to the puppets or their square's color. Only initial responses were coded. Child-level data was excluded following warmup failures (*N* = 8) or no or unintelligible responses to both trials of a condition (*N* = 2). Trial-level data was excluded following no or unintelligible responses to one trial a condition (*N* = 7 trials from 6 participants, all 2-year-olds). In total, there were 256 trials (64 participants x 4 trials), resulting in 256 – 7 = 249 analyzed trials. A naïve coder coded all data. Reliability on a random 25% of all 256 trials by a knowledgeable coder was excellent, κ = 0.977, agreement = 98.4%.

**Data analysis plan.** Analyses used Bayesian hierarchical modeling (Gelman et al., 2013) via brms (Bürkner, 2017) in R (R Core Team, 2018).

***Complex analyses***. These involved a set of relatively complex models. Specifically, models contained fixed effects of age, experimenter, gender, trial order (comprehension or production first), and speaker puppet order (P1/P3, P2/P3, P3/P1, or P3/P2) and had identical random effects and binomially distributed outcomes (dyadic or group interpretation, logit link). A null model assumed zero condition effect, a reduced model an additive condition effect, and a full model an interactive condition $x$ age effect. The full model formula was

group interpretation | total trials ~ age group * condition + gender + experimenter + trial order + speaker puppet order + (condition | participant id)

Weakly informative normal priors were placed on fixed effects (see code). Age group and condition were treatment coded (reference: 2-year-olds, dyad likely); others were sum coded. Models were selected via model stacking based on approximate leave-one-out cross validation (all pareto-k less than 0.7; Yao et al., 2018). Robustness of model weights was adequate across repeated comparisons with independent sets of posterior samples. All R-hats equaled 1.00 with zero divergences. Effective sample sizes and prior and posterior predictive checks were adequate.

***Simple analyses***. These analyses involved simpler models of conditional means and age associations. These were like *t*-tests with appropriate random effects and likelihood (see below).

***Posterior parameter estimates***. We report the 95% highest posterior density interval (HDI) and posterior probability greater than $x$ for parameter $\beta_i$, $\Pr(\beta_i > x|\boldsymbol{D})$. The former includes the 95% most likely values of $\beta_i|\boldsymbol{D}$; the latter quantifies confidence that $\beta_i > x|\boldsymbol{D}$. We say that there is "strong evidence" that $\beta_i$ is greater than $x$ if $1.000 \geq \Pr(\beta_i > x|\boldsymbol{D}) > .950$, "moderate evidence" if $.950 \geq \Pr(\beta_i > x|\boldsymbol{D}) > .900$, and "weak evidence" if $.900 \geq \Pr(\beta_i > x|\boldsymbol{D}) \geq .500$. On the other hand, we say that there is "strong evidence" that $\beta_i$ is less than $x$ if $.000 \leq \Pr(\beta_i > x|\boldsymbol{D}) < .050$, "moderate evidence" if $.050 \leq \Pr(\beta_i > x|\boldsymbol{D}) < .100$, and "weak evidence" if $.100 \leq \Pr(\beta_i > x|\boldsymbol{D}) < .500$. Note that these evidence statements are not interpreted like *p*-values. They are not statements about the plausibility of test statistics under null hypothesis distributions. Rather, they are statements about posterior confidence in directional hypotheses about $x$ and can be read as posterior odds $\frac{\Pr(\beta_i > x|\boldsymbol{D})}{1-\Pr(\beta_i > x|\boldsymbol{D})}$. Thus, even "weak" evidence that $\beta_i > x|\boldsymbol{D}$ nonetheless permits $\beta_i > x|\boldsymbol{D}$ to be nearly 9 times more probable than $\beta_i < x|\boldsymbol{D}$, while "moderate" evidence permits $\beta_i > x|\boldsymbol{D}$ to be nearly 19 times more probable than $\beta_i < x|\boldsymbol{D}$**.** Sensitivity analysis of fixed effects estimates in the complex analyses' mixed models were adequate (Schad et al., 2021). Sensitivity analysis investigates the influence of prior on posterior belief for fixed data and model.

## Results

The complex analyses precede the simple analyses. The Results Section ends with a series of comprehension and manipulation checks. Briefly, these analyses present evidence that suggests that 2- and 4-year-olds understood that *we* includes speakers (though, evidence was stronger among 4-year-olds) and that the spatial cues manipulation worked as intended among 2-year-olds.

**Complex analyses**. There were two sets of complex analyses. The first set of complex analyses investigated trial-level behavior (Figure 2A). Stack weights $\boldsymbol{w_{M_i}}$ for the posterior null $M_{null}$, reduced $M_{redu}$, and full $M_{full}$ models favored the full model, $w_{M_i} \in \{w_{M_{null}} = .07, w_{M_{redu}} = .20, w_{M_{full}} = .73\}$. As predicted, the full model better predicted interpretations than competing models. This pattern of model comparison results was robust to weaker and more informative priors on the fixed effects of the candidate models (see code).

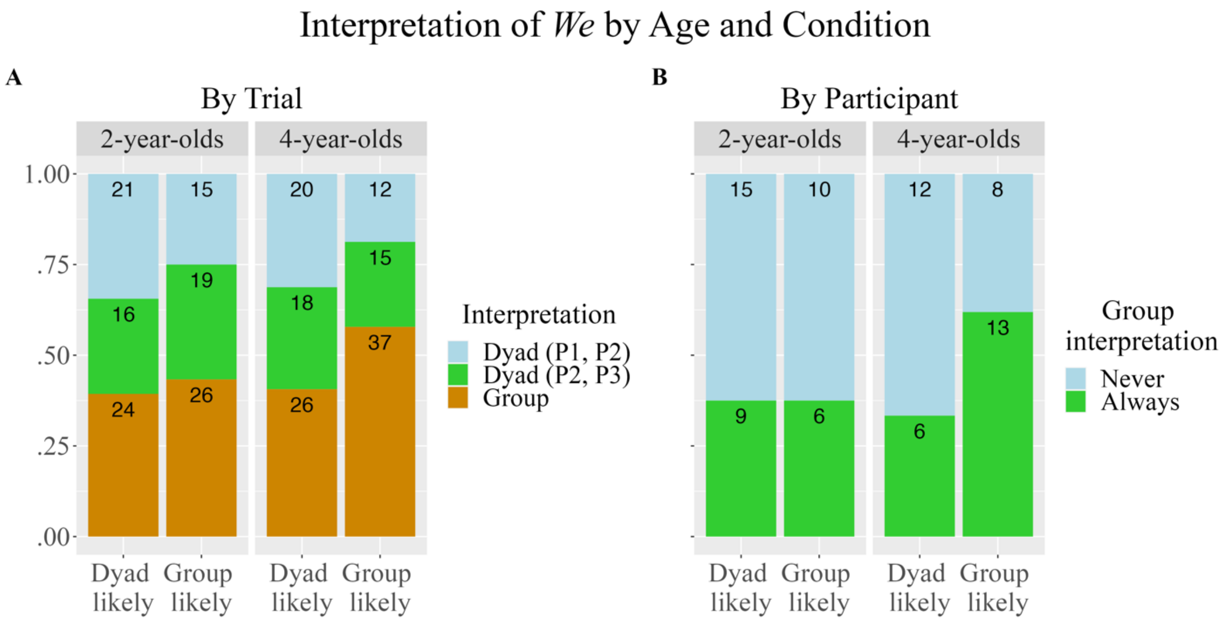

**Figure 2**. Panels faceted by age; condition on the *x*- and proportion on the *y*-axis (*N* inset).

In the posterior full model, there was weak evidence for experimenter, order, age, and condition effects on interpretations, all $\Pr(\beta > 0|\boldsymbol{D})$ between .602 and .783. There was moderate evidence that females made group interpretations less often than males, $\text{HDI} = [-0.88, 0.01]$, $\Pr(\beta > 0|\boldsymbol{D}) = .055$ (for data and analysis, see Supplementary Figure 2). There was weak evidence for the predicted age *x* condition interaction, $\text{HDI} = [-0.25, 1.64]$, $\Pr(\beta > 0|\boldsymbol{D}) =$ .898. Sampling the full model with 4-year-olds as reference level showed strong evidence that 4-year-olds made group interpretations relatively more often in the group likely than dyad likely condition, $\text{HDI} = [0.09, 1.35]$, $\Pr(\beta > 0|\boldsymbol{D}) = .972$. On balance, these results suggest that, in referring situations like that of the present study, 4-year-olds distinguish between dyadic and group interpretations, whereas 2-year-olds do not. This first set of results accords with study predictions.

The second set of complex analyses investigated participant-level behavior (Figure 2B). Stack weights $\boldsymbol{w_{M_i}}$ for the posterior null, reduced, and full models favored the full model, $w_{M_i} \in$

$\{w_{M_{null}} = .00, w_{M_{redu}} = .00, w_{M_{full}} = 1.00\}$. As predicted, the full model better predicted interpretations than competing models. This pattern of model comparison results was robust to weaker and more informative priors on the fixed effects of the candidate models (see code).

In the posterior full model, there was weak evidence for experimenter, order, age, and condition effects on interpretations, all $\Pr(\beta > 0|\boldsymbol{D})$ between .441 and .792. There was moderate evidence that females made group interpretations always, as opposed to never, less often than males, $\text{HDI} = [-1.76, 0.27]$, $\Pr(\beta > 0|\boldsymbol{D}) = .075$ (see Supplementary Figure 2). There was weak evidence for the predicted age *x* condition interaction, $\text{HDI} = [-0.65, 2.05]$, $\Pr(\beta > 0|\boldsymbol{D}) = .848$. Sampling the full model with 4-year-olds as reference level showed strong evidence that 4-year-olds made group interpretations always, as opposed to never, relatively more often in the group likely than dyad likely condition, $\text{HDI} = [-0.02, 1.46]$, $\Pr(\beta > 0|\boldsymbol{D}) = .971$. On balance, these results suggest that, in referring situations like that of the present study, 4-year-olds distinguish between dyadic and group interpretations, whereas 2-year-olds do not. This second set of results accords with the first set of results and with study predictions.

Specifically, model comparison more strongly favored the full model in the second set of complex analyses than in the first set. However, even though the by-participant interaction was nominally starker than the by-trial interaction (Figure 2), the smaller by-participants sample size meant that the posterior full model in the second set of analyses was more uncertain about the size of the interaction than was the posterior full model in the first set ($\Pr(\beta > 0|\boldsymbol{D}) = .848$ vs. .898).

**Simple analyses**. Two sets of simple analyses were conducted. The first set investigated whether participants made group interpretations above chance (Figure 2A). These models included a participant random intercept and binomial likelihood. Chance was defined in two ways. One definition of chance assumed participants compared each of the three interpretations separately.

Under this definition, chance equaled $\mathrm{logit}(.33) = -0.69$. Another definition of chance assumed participants compared only a dyadic and a group interpretation. Under this definition, chance equaled $\mathrm{logit}(.50) = 0.00$. In the group likely condition, assuming that participants compared three interpretations, there was weak evidence that 2-year-olds chose group interpretations above chance, HDI = $[-1.07, 0.38]$, $\Pr(\beta > -0.69|\boldsymbol{D})$ = .862; whereas, there was strong evidence that 4-year-olds did so, HDI = $[-0.42, 1.37]$, $\Pr(\beta > -0.69|\boldsymbol{D})$ = .993. However, assuming that participants compared only a dyadic and a group interpretation, there was weak evidence that 2-year-olds made dyadic interpretations above chance, $\Pr(\beta > 0.00|\boldsymbol{D})$ = .183; whereas, there was weak evidence that 4-year-olds made group interpretations above chance, $\Pr(\beta > 0.00|\boldsymbol{D})$ = .846. Model comparison favored a model with an age term over one without, $\boldsymbol{w_{M_i}} \in \{w_{M_{nuil}} =$ $.13, w_{M_{age}} = .87\}$, with weak evidence that 4-year-olds made group interpretations more often than did 2-year-olds, HDI = $[-0.33, 1.41]$, $\Pr(\beta > 0.00|\boldsymbol{D})$ = .887. On balance, these results suggest that, in referring situations like that of the group likely condition, 4-year-olds form group interpretations of *we* more often than do 2-year-olds. This accords with study predictions.

In the dyad likely condition, assuming that participants compared three interpretations, there was weak evidence that 2-year-olds chose dyadic interpretations above chance, HDI = $[-2.19, 0.62]$, $\Pr(\beta > -0.69|\boldsymbol{D})$ = .470; whereas, there was weak evidence that 4-year-olds chose group interpretations above chance, HDI = $[-1.16, 0.19]$, $\Pr(\beta > -0.69|\boldsymbol{D})$ = .794. However, assuming that participants compared only a dyadic and a group interpretation, there was moderate evidence that 2-year-olds chose dyadic interpretations above chance, $\Pr(\beta > 0.00|\boldsymbol{D})$ = .131; and moderate evidence that 4-year-olds chose dyadic interpretations above chance, $\Pr(\beta > 0.00|\boldsymbol{D})$ = .089. Model comparison favored a model without an age term over one with, $\boldsymbol{w_{M_i}} \in \{w_{M_{null}} = 1.00, w_{M_{age}} = .00\}$. On balance, these results suggest that, in referring situations like that of the dyad likely condition, 2- and 4-year-olds interpret *we* similarly and generally favor dyadic interpretations. This accords with study predictions.

Altogether, the first set of simple analyses suggests that, in appropriately ambiguous contexts, 4-year-olds more readily form group interpretations of uses of *we* than do 2-year-olds.

The second set of simple analyses investigated whether participants picked group interpretations always, as opposed to never, above chance (Figure 2B). Thus, this analysis considered only the extreme response profiles of participants who completed both trials in a condition. Because we collapsed across trials within participants, observations were independent. Thus, these models included no random effects and Bernoulli likelihood. Chance was defined in two ways. One definition of chance assumed participants compared each of the three interpretations. Under this definition, chance equaled $\text{logit}(.33^2 = .11) = -2.08$. Another definition of chance assumed participants only compared only a dyadic and a group interpretation. Under this definition, chance equaled $\text{logit}(.50^2 = .25) = -1.10$. In the group likely condition, assuming that participants compared three interpretations, there was strong evidence that participants of both age groups chose group interpretations always, as opposed to never, above chance, both $\Pr(\beta > -2.08|\boldsymbol{D}) \geq 0.999$. However, assuming that participants compared only a dyadic and a group interpretation, there was moderate evidence that 2-year-olds chose group interpretations always, as opposed to never, above chance, $\text{HDI} = [-1.45, 0.49]$, $\Pr(\beta > -1.10|\boldsymbol{D}) = .903$; whereas there was strong evidence that 4-year-olds chose group interpretations always, as opposed to never, above chance, $\text{HDI} = [-0.37, 1.31]$, $\Pr(\beta > -1.10|\boldsymbol{D}) = 1.00$. Model comparison favored a model with an age term over one without, $\boldsymbol{w_{M_i}} \in \left\{w_{M_{null}} = .14, w_{M_{age}} = .86\right\}$, with moderate evidence that 4-year-olds more often chose group interpretations always, as opposed to never, than did 2-year-olds, $\text{HDI} = [-0.42, 1.63]$,

$\Pr(\beta > 0.00|\boldsymbol{D}) = .869$. On balance, these results suggest that, in referring situations like that of the group likely condition, 4-year-olds form group interpretations of *we* always, as opposed to never, more often than do 2-year-olds. This accords with study predictions.

In the dyad likely condition, assuming that participants compared three interpretations, there was strong evidence that participants of both age groups chose group interpretations always, as opposed to never, above chance, both $\Pr(\beta > -2.08|\boldsymbol{D}) \geq 0.998$. However, assuming that participants compared only a dyadic and a group interpretation, there was moderate evidence that 2-year-olds chose group interpretations always, as opposed to never, above chance,: HDI = $[-1.28, 0.31]$, $\Pr(\beta > -1.10|\boldsymbol{D}) = .940$; and weak evidence that 4-year-olds did so, HDI = $[-1.58, 0.27]$, $\Pr(\beta > -1.10|\boldsymbol{D}) = .845$. Model comparison favored a model without an age term over one with, $\boldsymbol{w_{M_i}} \in \left\{w_{M_{null}} = 1.00, w_{M_{age}} = .00\right\}$. On balance, these results suggest that, in referring situations like that of the dyad likely condition, 2- and 4-year-olds interpret *we* similarly (though group interpretations were chosen often). This accords with study predictions.

Taken together, the second set of simple analyses suggests that, in appropriately ambiguous contexts, 4-year-olds more readily form group interpretations of uses of *we* than do 2-year-olds. This pattern accords with that of the first set of simple analyses and with study predictions.

**Comprehension and Manipulation Checks**. If participants understood that referents of *we* include speakers, then they should have disfavored interpretations that excluded speaker puppets. For example, participants should have chosen the P1-P2 interpretation less often than chance when P3 spoke. Indeed, assuming that 2-year-olds compared three interpretations separately, there was weak evidence that they made the P1-P2 interpretation less often than chance when P3 spoke, HDI = $[-1.63, -0.50]$, $\Pr(\beta > -0.69|\boldsymbol{D}) = .103$; and strong evidence if one assumes that they compared only a dyadic and a group interpretation, $\Pr(\beta > 0.00|\boldsymbol{D}) = .000$. Among 4-year-olds, assuming that they compared three interpretations, there was strong evidence that they made the P1-P2 interpretation less often than chance when P3 spoke, $\Pr(\beta > -0.69|\boldsymbol{D}) = .000$; and strong evidence if one assumes that that they compared only a dyadic and a group interpretation, $\text{HDI} = [-2.02, -0.84], \Pr(\beta > 0.00|\boldsymbol{D}) = .006$. Along the same lines, participants should have chosen the P2-P3 interpretation less often than chance when P1 spoke. Indeed, assuming that 2-year-olds compared only a dyadic interpretation and a group interpretation, there was moderate evidence they made the P2-P3 interpretation less often than chance when P1 spoke, $\text{HDI} = [-1.24, 0.14], \Pr(\beta > 0.00|\boldsymbol{D}) = .063$. However, surprisingly, assuming that 2-year-olds compared three interpretations, there was weak evidence that they chose the P2-P3 interpretation more often than chance, $\Pr(\beta < -0.69|\boldsymbol{D}) = .679$. In contrast, assuming that 4-year-olds compared only a dyadic and a group interpretation, there was strong evidence they made the P2-P3 interpretation less often than chance when P1 spoke, $\text{HDI} = [-2.19, -0.59], \Pr(\beta > 0.00|\boldsymbol{D}) = .000$; and strong evidence if one assumes that they compared three interpretations, $\Pr(\beta > -0.69|\boldsymbol{D}) = .046$. In sum, these results suggest that, in referring situations like that of the present study, 2- and 4-year-olds understand that the referent of *we* includes speakers, although 4-year-olds may understand this better than 2-year-olds.

The present manipulation relied on participants' ability to leverage spatial cues to disambiguate plural person reference. Specifically, the spatial proximity of a speaker puppet relative to other puppets was the key manipulation. If participants chose dyadic interpretations when P2 spoke, then they should have chosen the P1-P2 interpretation more often than the P2-P3 interpretation (i.e., because P1 was nearer to P2 than was P3 to P2). Indeed, there was strong evidence that, if 2-year-olds chose dyadic interpretations when P2 spoke, then they favored the P1-P2 interpretation, $\text{HDI} = [0.16, 2.14], \Pr(\beta > 0.00|\boldsymbol{D}) = .989$. Interestingly, there was weak evidence that, if 4-year-olds chose a dyadic interpretation when P2 spoke, they favored the P2-P3 interpretation, $\mathrm{HDI} = [-1.39, 0.39], \Pr(\beta > 0.00|\boldsymbol{D}) = .142$. The key point that these results suggest is that the manipulation worked as intended among 2-year-olds.

Why did 4-year-olds behave unexpectedly in the manipulation check? One possibility is that 4-year-olds perceived or utilized the spatial cues differently than did 2-year-olds, and so behaved differently (i.e., as documented in the manipulation check). Evidence for this could come from differential rates of P1-P2-P3 group interpretations, relative to P1-P2 dyadic interpretations, when P2 spoke (i.e., excluding trials in which the P2-P3 dyadic interpretation was chosen). This is because, if groupmindedness were irrelevant for participants' interpretations in the P2 dyadic condition (as we had expected), then both 2- and 4-year-olds should favor the P1-P2 interpretation over the P1-P2-P3 interpretation. However, if groupmindedness were (unexpectedly) related to participants' interpretations even in the P2 dyadic condition, then 4-year-olds' rate of choosing P1-P2 over P1-P2-P3 should diverge from that of 2-year-olds'. Indeed, in line with the latter possibility, there was moderate evidence that 4-year-olds favored group interpretations over P1-P2 dyadic interpretations, $\mathrm{HDI} = [-1.47, 0.29], \Pr(\beta > 0.00|\boldsymbol{D}) = .096$; while there was weak evidence that 2-year-olds favored P1-P2 dyadic interpretations over group interpretations, $\mathrm{HDI} = [-0.47, 1.13], \Pr(\beta > 0.00|\boldsymbol{D}) = .788$. Moreover, a model that included age found moderate evidence that 4-year-olds chose group interpretations more often than P1-P2 dyadic interpretations, compared to 2-year-olds, $\mathrm{HDI} = [-2.30, 0.24], \Pr(\beta > 0.00|\boldsymbol{D}) = .059$. In sum, 4-year-olds may have perceived or utilized cues in situations in which P2 spoke differently than did 2-year-olds. This was unexpected but aligns with the idea that groupmindedness emerges at 3 years, as 4-year-olds favored group interpretations and 2-year-olds dyadic interpretations.

## Discussion

In the current study, participants were tasked with interpreting the intended referent of a speaker's use of *we*. It was predicted that 2-year-olds would stick with dyadic interpretations, whereas 4-year-olds' interpretations would depend on nonlinguistic contextual cues. Accordingly, we found that 4-year-olds used nonlinguistic contextual cues to make dyadic or group interpretations, while 2-year-olds made mostly dyadic interpretations. This pattern of results is best explained by age-related developments in children's social-conceptual structure.

Referential interpretation relies on listeners' conceptual structure. If listeners cannot conceive of a referent, then they cannot appropriately form interpretations of the referent. This is why 2-year-olds in this study did not appropriately form group interpretations of *we*, even when nonlinguistic cues pulled for group interpretations. That is, 2-year-olds have not yet undergone the groupminded shift. Consequently, 2-year-olds lack the requisite conceptual structure for group interpretations and so do not appropriately make them. Meanwhile, 4-year-olds can conceive of groups. Thus, only 4-year-olds appropriately formed group interpretations.

The emergence of groupmindedness "expands" the set of conceivable plural person referents to include groups, in addition to dyads. "Certain thoughts cannot be communicated to children even if they are familiar with the necessary words [because] the adequately generalized concept that alone ensures full understanding may still be lacking," (Vygotsky, 1962, p. 8).

Alternative interpretations based on autonomous linguistic developments cannot explain the documented age-related changes in interpretation of *we*. Participants heard identical language in both conditions and this language was well within the linguistic skills of 2.5-year-olds. Importantly, there is evidence from looking time studies suggesting that infants have some conception of social groups from a third-person perspective based on contiguity and similarity among members (e.g., Powell & Spelke, 2013). However, this is arguably a different concept than the concept of first-person "we" focused on groups in which the speaker participates.

Did 2-year-olds respond randomly in the test trials? Three considerations suggest that 2-year-olds responded non-randomly. First, the warmup reduced the likelihood that participants failed to understand the task structure. Second, comprehension checks suggested that 2-year-olds understood the semantic constraint on *we* that speakers are part of intended referents of *we*. While evidence for this was stronger among 4- than 2-year-olds, this is to be expected because children rarely hear and use first-person plural forms (Vasil et al., 2023). Third, a manipulation check suggested that 2-year-olds responded appropriately when P2 spoke (i.e., by choosing P1-P2 over P2-P3 interpretations). These considerations suggest that 2-year-olds responded non-randomly.

It is unclear what to make of the finding that 4-year-olds interpreted P2 differently than did 2-year-olds (i.e., by favoring P2-P3 over P1-P2 dyadic interpretations). Evidence was presented that 4-year-olds perceived or utilized the spatial cues differently than did 2-year-olds when P2 spoke, and so behaved differently. Regardless of the explanation for 4-year-olds, the key message from the manipulation checks is that 2-year-olds were appropriately sensitive to the manipulation.

In this study, children's sense of groupmindedness was assumed from age, based on previous studies (Tomasello, 2019). Instead, following Bates (1979), future research might relate nonlinguistic manifestations of groupmindedness (e.g., norm enforcement) to children's interpretation of *we*. Additionally, one nonlinguistic cue was investigated in this study. However, children's interpretation of *we* may be sensitive to other nonlinguistic cues (e.g., perceptual similarity). Linguistic cues likely play a role, too, such as prior discourse about what "we" are doing (Vasil, 2023). Future research might investigate effects of other nonlinguistic and linguistic cues on the development of interpretation of *we*. Furthermore, it should be noted that we investigated exclusive *we*, in which participants were excluded from intended referents. However, the development of inclusive *we* – in which the "we" includes listeners – should be investigated, too. Indeed, although first-person plural pronouns are rare in children's input (Vasil et al., 2023), inclusive *we* input probably exceeds that of exclusive *we*. Finally, crosslinguistic investigation is key. Does formal marking of, e.g., clusivity and number influence children's interpretation of *we*?

In conclusion, evidence was reported that suggests that age-related developments in conceptualization of social relations influence early language use. Two-year-olds inflexibly favor dyadic interpretations of *we*. Four-year-olds flexibly interpret *we* as referring to dyads or groups.

**Acknowledgements**. Data and code are freely available to reproduce all results and figures in the manuscript ([*anonymized*] https://osf.io/9hygn/?view_only=c096f9dc90af46a18aba2d76475cf5b3.). Materials necessary to attempt to replicate the findings reported in this article are available upon request of the corresponding author. Hypotheses and certain of the analyses were preregistered ([*anonymized*] https://aspredicted.org/blind.php?x=RVL_8ZQ).

## References

Bates, E. (1979). *The Emergence of Symbols: Cognition and Communication in Infancy*. Academic Press.

Bohn, M., Le, K. N., Peloquin, B., Köymen, B., & Frank, M. C. (2020). Children’s interpretation of ambiguous pronouns based on prior discourse. *Developmental Science*, e13049. https://doi.org/10.1111/desc.13049

Bruner, J. S. (1983). *Child’s Talk: Learning to Use Language*. Norton.

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1), Article 1. https://doi.org/10.18637/jss.v080.i01

Fillmore, C. J. (1975). An Alternative to Checklist Theories of Meaning. *Annual Meeting of the Berkeley Linguistics Society*, *1*, Article 0. https://doi.org/10.3765/bls.v1i0.2315

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian Data Analysis* (3rd ed.). CRC Press/Taylor & Francis.

Girouard, P. C., Ricard, M., & Décarie, T. G. (1997). The acquisition of personal pronouns in French-speaking and English-speaking children. *Journal of Child Language*, *24*(2), 311–326. https://doi.org/10.1017/S030500099700305X

Gräfenhain, M., Behne, T., Carpenter, M., & Tomasello, M. (2009). Young children’s understanding of joint commitments. *Developmental Psychology*, *45*(5), 1430–1443.

Ibbotson, P., Hartman, R. M., & Björkenstam, K. N. (2018). Frequency filter: An open access tool for analysing language development. *Language, Cognition and Neuroscience*, *33*(10), 1325–1339. https://doi.org/10.1080/23273798.2018.1480788

Langacker, R. W. (1987). *Foundations of Cognitive Grammar: Theoretical prerequisites: Vol. I*. Stanford University Press.

Liebal, K., Carpenter, M., & Tomasello, M. (2013). Young children's understanding of cultural common ground. *The British Journal of Developmental Psychology*, *31*(Pt 1), 88–96.

Liszkowski, U. (2018). Emergence of shared reference and shared minds in infancy. *Current Opinion in Psychology*, *23*, 26–29. https://doi.org/10.1016/j.copsyc.2017.11.003

Oberauer, K. (2022). The Importance of Random Slopes in Mixed Models for Bayesian Hypothesis Testing. *Psychological Science*, *33*(4), 648–665. https://doi.org/10.1177/09567976211046884

Orvell, A., Kross, E., & Gelman, S. A. (2018). That's how "you" do it: Generic you expresses norms during early childhood. *Journal of Experimental Child Psychology*, *165*, 183–195. https://doi.org/10.1016/j.jecp.2017.04.015

Powell, L. J., & Spelke, E. S. (2013). Preverbal infants expect members of social groups to act alike. *Proceedings of the National Academy of Sciences*, *110*(41), E3965–E3972. https://doi.org/10.1073/pnas.1304326110

R Core Team. (2018). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing.

Ricard, M., Girouard, P. C., & Décarie, T. G. (1999). Personal pronouns and perspective taking in toddlers. *Journal of Child Language*, *26*(3), 681–697. https://doi.org/10.1017/S0305000999003943

Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled Bayesian workflow in cognitive science. *Psychological Methods*, *26*(1), 103–126. https://doi.org/10.1037/met0000275

Schmidt, M. F. H., Rakoczy, H., & Tomasello, M. (2012). Young children enforce social norms selectively depending on the violator's group affiliation. *Cognition*, *124*(3), 325–333.

Tomasello, M. (2019). *Becoming Human: A Theory of Ontogeny*. Harvard University Press.

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

Vasil, J., Moore, C., & Tomasello, M. (2023). Thought and language: Association of groupmindedness with young English-speaking children's production of pronouns. *First Language*, *43*(5), 516–538. https://doi.org/10.1177/01427237231169398

Vasil, J., & Tomasello, M. (2022). Effects of "we"-framing on young children's commitment, sharing, and helping. *Journal of Experimental Child Psychology*, *214*, 105278. https://doi.org/10.1016/j.jecp.2021.105278

Vygotsky, L. S. (1962). *Thought and Language* (4th ed.). The MIT Press.

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using Stacking to Average Bayesian Predictive Distributions (with Discussion). *Bayesian Analysis*, *13*(3). https://doi.org/10.1214/17-BA1091