“Oh! Um… Sure”: Children and adults use other’s linguistic surprisal to reason about expectations and learn stereotypes

Ben Morris
benmorris@uchicago.edu
Department of Psychology
University of Chicago

Alex Shaw
ashaw1@uchicago.edu
Department of Psychology
University of Chicago

Abstract

While people may be reluctant to explicitly state social stereotypes, their underlying beliefs may nonetheless leak out in subtler conversational cues, such as surprisal reactions that convey information about expectations. Across 3 experiments with adults and children (ages 4-9), we compare permissive responses (“Sure, you can have that one”) that vary the presence of surprisal cues (interjections “oh!” and disfluencies “um”). In Experiment 1 (n = 120), children by 6-to-7 use surprisal reactions to infer that a boy more likely made a counter-stereotypical choice. In Experiment 2, we demonstrate that these cues are sufficient for children (n = 120) and adults (n = 80) to learn a novel expectation about a group of aliens. In Experiment 3, adults (n = 150) use the distribution of surprisal information to infer whether a novel behavior is gender-stereotyped. Across these experiments, we see emerging evidence that conversational feedback may provide a crucial and unappreciated avenue for the transmission of social beliefs.

Keywords: emotion, conversation, feedback, stereotypes, cognitive development

Introduction

In conversation, much is communicated without being directly said. Imagine a young boy expressing a gender counter-stereotypical preference (e.g., wanting to buy a Barbie doll) and his caregiver provides a permissive, gender egalitarian response. However, imagine that response comes slowly, with markers of surprise and production difficulty (e.g., “Oh! Um… Sure”). What message does that young boy really receive? In this paper, we explore how children and adults reason about surprisal in these situations and how these cues provide data to infer speaker expectations to learn about normative behavior (and even stereotypes).

Surprise is a basic emotion that occurs in the face of unexpectedness, and thus witnessing *others’* surprise can license inferences about others’ expectations, a kind of vicarious surprise. Adults show sophisticated abilities to reason about others’ emotional expressions (including surprise), rationally and flexibly inferring underlying mental states accordingly (e.g., Wu et al., 2018). In this paper, we investigate how others’ surprise might provide rich information about the structure of social expectations. For adults, reasoning about others’ reactions in this way would provide crucial insights into a speaker’s expectations, extant stereotyped beliefs, and even for learning norms in a new social environment (e.g., how casually to dress in a new workplace). For children, the consequences may be even more profound.

Conversations with caregivers and other adults provide a fundamental venue for children to learn about the social world, and consequently for the transmission of stereotypes. Even ostensibly well-meaning messages can often have unintended consequences, with subtle linguistic cues highlighting stereotype information (e.g., Chestnut et al., 2021; Moty & Rhodes, 2021; Rhodes et al., 2012). For example, explicitly egalitarian statements like “Girls are just as good as boys at math” can still perpetuate gendered ability stereotypes by setting boys as the reference point (Chestnut et al., 2021).

Beyond isolated messages, others’ feedback and responsiveness also hold rich social information. While research has demonstrated that others’ non-verbal affect may foster stereotype transmission (Skinner et al., 2020), we argue that others’ expressions of surprise may hold particularly stereotype-relevant information by communicating their expectations. We know that children ages 6-to-8 can use others’ marked facial expressions of surprise to derive social inferences, e.g. about another agent’s competence (Asaba et al., 2020). For example, if two children successfully score a basket, but only one’s success leaves the teacher visibly shocked (actually dropping her jaw), we can infer who the better player is. Others’ emotional expressions– even non-valenced reactions like surprisal– can thus convey substantive information about the social world (Asaba et al., 2020; Wu et al., 2021).

But of course subtle social information is not just written on our faces; it also leaks out through the *linguistic* channel– specifically, surprisal interjections (e.g., “oh”) and disfluencies (e.g., filled pauses like “um”). These cues are ubiquitous features of casual, everyday language use and seem to emerge early in children’s own productions (Casillas, 2014; Fox Tree, 1995). Surprisal interjections definitionally index speaker expectations, and two key observations suggest disfluencies may also license inferences about a speaker’s expectations. First, decades of cognitive science experiments demonstrate that violations of expectations delay response times in both children and adults (Meyer et al., 1997; Schützwohl & Reisenzein, 1999). As a result, conversational responses may be slowed following unexpected information or behavior. Second, adults interpret others’ disfluencies in contentious conversations (e.g., about gun control) as reflecting underlying discomfort with the topic and potential dishonesty (Fox Tree, 2002). Together, these findings suggest that these cues reliably co-occur with speaker surprisal and

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

thus may lead adults and children to form inferences about a speaker's underlying expectations.

To explore how these cues to speaker expectations could inform stereotype transmission, we focus on the domain of gender stereotypes as a case study (Experiments 1 and 3). While the general inferential process could support learning many kinds of expectations (as we explore in Experiment 2), the development of gender stereotypes provides an important and ecologically-valid test case. Gender stereotypes emerge early in development; as young as 3, children show robust gender stereotypes about toy preferences, and report that their parents would be less approving of counter-stereotypical toy choices (Eisenberg et al., 1982; Freeman, 2007). By age 6, children show gender biases in their beliefs about ability and this affects their own decisions about which opportunities to pursue (Bian et al., 2017). To be able to combat such stereotypes, we must better understand the transmission processes underlying stereotype transmission.

General Approach

In three experiments, we take a social learning approach to ask how children and adults can use linguistic cues of surprisal to reason and learn about what kinds of behaviors are expected, even when these cues leak information that is counter to the speaker's explicit messaging. In each experiment, an adult figure affirms a character's choice (e.g. "Sure, you can have that one") and shows no facial expressions of surprise (maintaining a consistent, positive facial expression). However between conditions, we vary the presence or absence of conversational markers that tip the adult's hand—indicating whether they *did* or *did not* expect the child to make such a choice.

In Experiment 1, we ask whether children use surprisal feedback to infer if a target boy's toy choice is in line with gender stereotypes. In Experiment 2, we explore this same inference in novel categories to probe whether these cues could serve as a plausible mechanism for both adults and children to learn about the descriptive and normative expectations of the social world. In Experiment 3 with adults only, we connect these two experiments explicitly to ask how surprisal cues can lead adults to learn a novel gender stereotype.

Stimuli Creation

For each experiment, we followed the same general procedure to create test utterances that varied across conditions. We started by having native speakers record surprisal utterances that contained interjections and disfluencies (e.g., "Oh really? Um... Sure, honey. Uh... We can buy you that one"), reading them as naturally as possible. We then digitally removed the surprisal markers to create corresponding fluent utterances that were well matched (e.g., "Sure, honey. We can buy you that one."). Thus, the only features that varied across test utterances were the presence or absence of interjections and disfluencies. Utterances may have included additional paralinguistic markers outside of the interjections or disfluencies themselves (e.g., rising intonation in other phrases), but this information was matched across our conditions.

Experiment 1

In a pre-registered experiment, children were shown videos in which a target boy is choosing between two gender stereotyped toy options (e.g., a doll or a truck), and his choice was ambiguous from the participant's perspective. Children then saw an adult figure respond approvingly, but either with cues to surprise (surprise condition) or fluently (fluent baseline condition). Children were then asked to infer which toy the boy had selected. This experiment asks how children use feedback to reason about whether a choice was expected (i.e. stereotypical) or unexpected (i.e. counter-stereotypical). The key prediction was that children would be more likely to infer the boy had selected a girl-stereotyped toy in the surprise condition, as compared to the fluent baseline condition.

Method

Participants We pre-registered a sample size of 120 children ages 4-to-9, with 20 children in each condition in each of three pre-registered age bins (4-5, 6-7, 8-9). Families were recruited online, primarily through a US University database of families who have expressed interest in doing research. Children completed this experiment over Zoom, interacting with a live experimenter who navigated a slide-style, animated Qualtrics survey. Based on a pre-registered exclusion criterion, children who failed to answer all of the questions were excluded and replaced (an additional 6 children).

Procedure Participants were shown two short animated stories that featured different protagonists and toys. Each story was about a young boy and an adult man looking at two familiar toys (one gender-stereotyped for boys, and one gender-stereotyped for girls). The experimenter introduced each story, and then the rest played as a pre-recorded video. The Toy Store trial involved a boy and his uncle buying a toy from the toy store (doll vs. truck). The Carnival trial involved a boy winning a game at a fair and choosing a prize (pink bear vs. blue bear). Across participants, trial order and toy position were counterbalanced.

Note that both stories were always about a young boy and a male adult. While the underlying inferences here could well hold with gendered stereotypes about young girls (as we explore more in Experiment 3), we focused on boys because their gender counter-stereotypical behaviors and preferences are typically policed more by adults than girls (e.g., Kane, 2006), and thus we expected that the inference from speaker surprisal would be most likely.

Each video showed a brief conversation. In both conditions, the target boy initially requested a toy (e.g., "Can we get a toy for my birthday?") and the adult acknowledged and accepted the request fluently (e.g., "Yeah, let's get one of those toys for your birthday"). This initial back-and-forth was included to establish that the child is allowed to choose a toy, and to demonstrate that the adult sometimes responds fluently. Next, the target boy requested one of the toys (e.g., "I want that one please"). Critically, the target boy's selection was ambiguous from the participant's perspective, as there was no visual cue to indicate which toy the child selected.

Test. In both conditions, the test utterances were positive and affirming of the character's choice. In the fluent baseline condition, the adult responded fluently (e.g., "Sure, honey. We can buy you that one"). In the surprise condition, the adult responded with the same permissive message but with markers of surprise and production difficulty (e.g., "Oh really? Um... Sure, honey. Uh... We can buy you that one"). Participants were then asked which toy the target boy asked for (our primary dependent measure).

Results

As pre-registered, we test for sensitivity to feedback with separate regressions predicting toy choice from condition for each age group. We see a significant effect of condition on 6- to 7-year-old children's responses ($\beta = 0.26$, $p = .011$) and 8- to 9-year-old children's responses ($\beta = 0.26$, $p = .007$). These condition effects showed that older children were selecting the "girl" stereotyped toy more frequently in the surprise condition (see Figure 1). There was no effect of condition on 4- to 5-year-old children's responses ($\beta = -0.01$, $p = .949$). Note also that, unsurprisingly, children in the fluent baseline showed significant gender stereotypes, predicting boys would select a "boy" stereotyped toy in all three age groups (*ps* < .001).

Discussion

We find that by age 6-to-7 children are more likely to infer that a boy chose a counter-stereotypical toy (e.g., a doll) if an adult responds with surprisal markers, compared to baseline. While children at all ages showed clear gender stereotypes at baseline, older children were able to partly override this stereotype based on an adult's surprisal. These data provide an initial demonstration that children are connecting conversational cues of surprisal with expectations about gender stereotypes. Thus, even though the parent gave a permissive and egalitarian response, when their linguistic markers revealed that they seemed surprised, 6-to-7 year-old children were relatively more likely to assume counter gender-normative behavior.

Experiment 2

In Experiment 2, we use a novel alien environment to ask whether these surprisal cues can provide a possible learning mechanism for developing new expectations about normative behavior. While Experiment 1 demonstrates that children connect surprisal cues with extant beliefs about other's expectations, this may or may not implicate these cues in the learning of new expectations (e.g., forming a new stereotype may be more complicated than linking a reaction to an established stereotype). Thus, Experiment 2 directly tests whether conversational surprisal cues can enable *learning* a novel expectation.

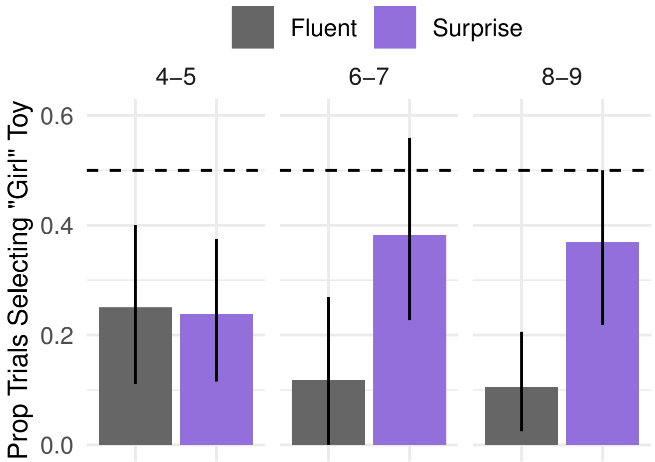

Figure 1: Children's toy selections across conditions for each of our three pre-determined age bins for Experiment 1. Error bars show bootstrapped 95% confidence intervals.

Rather than relying on pre-existing gender stereotypes to inform participants' priors about what is expected, Experiment 2 used novel behaviors and categories (aliens called "Hibbles" wearing hats). By manipulating surprisal cues, we aimed to differentially establish the exact same novel behavior as either unmarked and equally expected (fluent baseline) or marked and potentially unexpected (surprise condition). To test this, our primary measure asked participants to directly evaluate the markedness of the target behavior (judging it as normal or weird), rather than inferring which behavior evoked surprise (as in Experiment 1).

Method

Participants We collected data from a pre-registered sample of 120 children ages 4-to-9, with 20 children in each condition in each of three pre-registered age bins (4-5, 6-7, 8-9). As with Experiment 1, children completed this experiment over Zoom, interacting with a live experimenter who navigated a slide-style, animated Qualtrics survey. Based on pre-registered exclusion criteria, an additional 5 children were excluded and replaced due to technical difficulties, failing to answer all the questions, or parent interference.

A separate sample of 80 adults were recruited via MTurk and paid $0.75 for their participation. Adult participants completed the same task, but navigated the task on their own via Qualtrics. Participants who failed a CAPTCHA or a simple auditory attention check were prescreened and unable to complete the study.

Procedure Participants were shown an animated story that the experimenter narrated. Participants were introduced to a novel alien group ("Hibbles") and told about a school with a Hibble teacher and three Hibble students getting ready for a party. The rest of the story played out in a pre-recorded video wherein each Hibble child put on a hat one-at-a-time and the Hibble teacher responded affirmatively to each one.

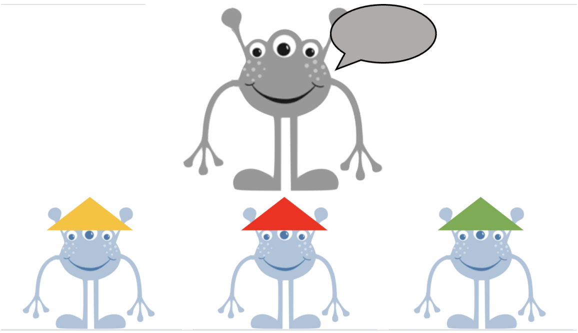

Figure 2: A still from Experiment 2 showing the Hibbles and their hats (colors counterbalanced).

Each Hibble put on a different colored hat (red, green, and yellow, with colors counterbalanced across participants, see Figure 2). The three response utterances followed the same structure with some variation (i.e. varying the initial response token across the three utterances "nice", "yeah", "cool").

In both conditions, the pattern of choices was identical and the teacher responded fluently to the first two Hibbles' hats. Across conditions, we manipulated the teacher's response to the third Hibble's choice (hereafter referred to as the target). In the fluent baseline condition, the teacher responded fluently, comparable to the past selections (e.g., "Cool. You look great!"). In the surprise condition, the teacher responded with stilted surprise, while still affirming the choice as before (e.g., "Oh! Um... Cool. You look uh... great!").

As our primary measure, participants were then asked to evaluate the normality of the target's choice ("Do you think it's normal or weird for a Hibble to wear a [green] hat?", with a two-point contingent follow-up question, e.g., "a little [weird] or really [weird]?"). As follow-up measures, participants were also asked to predict what color hat a novel Hibble would wear (prediction measure), and told about a Hibble who had been teased and asked to infer which color hat that the Hibble *had been* wearing (teasing measure).

Results

Following our pre-registered analysis plan, for each measure, we first report overall regression models, testing for the effects of condition, age (measured continuously), and their interaction, and then follow-up analyses testing the effect of condition in each predetermined age bin.

For children's weirdness judgments (our primary measure, see Figure 3), we see a significant effect of condition ($\beta = 0.48$, $p = .013$) such that children judged the target behavior as weirder in the surprise condition, and marginal interaction effect between condition and age ($\beta = 0.23$, $p = .050$). Examining children's weirdness judgements separately for each age bin, we see a significant effect of condition with the 8- to 9-year-olds ($\beta = 0.85$, $p = .008$), a marginal effect with the 6- to 7-year-olds ($\beta = 0.6$, $p = .052$), and no effect with the 4- to 5-year-olds ($\beta = -0.06$, $p = .894$). That is, older children, but not younger, judged wearing the target hat color as weirder when it had elicited a surprise, compared with the fluent condition.

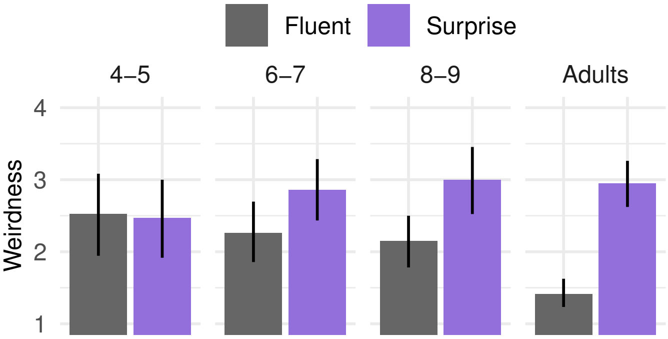

Figure 3: Children's weirdness judgements across conditions for each of our three pre-determined age bins for Experiment 2, with the adult sample for comparison. Error bars show bootstrapped 95% confidence intervals.

We next turn to our two follow-up measures. Examining children's predictions, we found no significant effects of condition ($\beta = 0.1$, $p = .199$), age ($\beta = -0.01$, $p = .853$), or their interaction ($\beta = 0.03$, $p = .566$). Examining children's responses for the teasing measure, we found a significant effect of condition ($\beta = 0.4$, $p = < .001$) such that children were more likely to expect that a teased character had been wearing the target hat color in the surprise condition. When asked about a novel Hibble who was teased, children in every age group were more likely to infer that Hibble had been wearing the target hat color in the surprise condition, compared with the fluent condition (all ps$<$ 0.05). We found no significant effect of age ($\beta = 0.04$, $p = .244$) or their interaction ($\beta = 0.02$, $p = .639$).

Adult Results For adults, we see significant effects of condition for all our measures. Adults judged the target hat color as significantly weirder in the surprise condition, relative to the fluent baseline ($\beta = 1.54$, $p < 0.001$). Adults were less likely to predict that a new Hibble would wear the target hat color in the surprise condition than the fluent baseline ($\beta = -0.19$, $p = 0.04$). Adults were also more likely to expect that a Hibble who was teased had been wearing the target hat color in the surprise condition, relative to the fluent baseline ($\beta = 0.44$, $p < 0.001$).

Discussion

Experiment 2 demonstrates that conversational cues to surprisal may serve as a viable learning mechanism for transmitting novel speaker expectations, and potentially stereotypes. Adults readily use other's surprisal reactions to learn a novel expectation, generate predictions, and infer social consequences. The developmental data clearly show that older children are sensitive to the feedback type in their weirdness evaluations (our primary measure), while 4-5 year old children do not show any sensitivity to feedback (as in Experiment 1). For children's predictions about a novel Hibble, we saw no effect of feedback type which could suggest children are not incorporating surprise into their own predictions, although null effects are difficult to interpret and there might have been difficulties detecting this effect with a choice measure (i.e. a reduction in selections against a 33% chance baseline). Interestingly for the teasing measure, children at all ages in the surprise condition inferred that a character was teased for wearing the target hat, more so than the fluent condition. Overall, these results suggest that surprise cues license additional inferences not just about extant expectations (as in Experiment 1), but also for learning entirely new and consequential expectations.

Experiment 3

Experiment 3 returns to the domain of gender stereotypes to ask how adults might use surprisal cues to learn a novel, gendered expectation. We introduced participants to a novel kids game called "Blickets" and showed some students who were playing Blickets (always an equivalent number of boys and girls). Unlike the prior experiments, Experiment 3 also contrasts two surprisal conditions to further probe the flexibility of adults' inferences. In one surprisal condition, the surprisal reactions covary with gender (gendered-surprise condition), while in the other they happen for both boys and girls (control-surprise condition). We contrast these conditions with a third fluent baseline condition.

We predicted that adults would incorporate information about both the presence and distribution of surprisal feedback when drawing inferences. We again used a perceived weirdness measure to capture unexpectedness, and predicted both surprisal conditions would lead to perceived unexpectedness relative to baseline. We also included two measures probing the extent to which adults saw the game as gendered, and predicted that only the gendered-surprise condition would stand out on those measures, and not the control-surprise condition (where surprise may be attributed to something more idiosyncratic).

Method

Participants A pre-registered sample of 150 adults (50 per condition) were recruited via Prolific and paid $0.80 for their participation. Participants who failed a CAPTCHA or a simple auditory attention check were prescreened and unable to complete the study.

Procedure Participants were randomly assigned to one of three conditions: fluent baseline, gendered-surprise, or a control-surprise. Participants read a short animated story about a classroom where some of the kids like to play a game called "Blickets". Four children (two boys and two girls) come to the teacher one at a time to ask for a toy to play Blickets. After each child asks for a toy, participants heard pre-recorded audio of the teacher affirming the child. The four response utterances followed the same structure with some variation (i.e. varying the initial response token across the four utterances "nice", "yeah", "cool", "sure").

Across conditions, we varied the surprisal of the teacher's responses. In the fluent baseline condition, the teacher provided unmarked responses to all four (e.g., "Yeah, you can play Blickets."). In the gendered-surprise condition, the teacher provided fluent responses for two students of one gender, but used conversational markers of surprisal for two students of the other gender (e.g., "Oh! Um... Yeah, uh... you can play Blickets."). In the control-surprise condition, the teacher also provided surprisal responses for two students, but now for one boy and one girl. The last child was the "target" (and always received a surprisal response in the two surprise conditions). Across participants, we counterbalanced the order of the children with two orders varying the final target's gender: boy-target order (girl, boy, girl, boy) and girl-target order (boy, girl, boy, girl). Please refer to Figure 4 for a simplified schematic of each condition.

Participants were then asked 3 dependent measures in a fixed order (using 7-point bipolar scales, with 0 indicating neutrality). For the weirdness measure, participants were asked to judge if the teacher thought it was normal or weird that the target character wanted to play "Blickets" (1 - really weird to 7 - really normal). For the teasing measure, participants saw two novel characters (a boy and a girl) who also played "Blickets" and were asked to predict which had been teased (1 - probably Bryan to 7 - probably Olivia). Lastly for the stereotype measure, participants were asked who usually plays "Blickets" (1 - mostly boys to 7 - mostly girls). Note that for analysis purposes, we reverse coded the teasing and stereotype scales for the girl-target-order, so that we could compare responses across orders.

Results

First for weirdness judgments, adults inferred the teacher thought the target's behavior was weirder in both the gendered-surprise (β = -3.66, $p < 0.001$) and control-surprise conditions (β = -2.81, $p < 0.001$), relative to the fluent baseline. Comparing our two surprisal conditions, adults inferred the teacher thought the target's behavior was significantly weirder in the gendered-surprise condition (β = -0.85, $p < 0.01$), compared with the control-surprise condition.

For teasing predictions (see Figure 5), adults were more likely to infer the target's gender was teased in the gendered surprise condition relative to the control-surprise condition (β = -1.78,$p < 0.001$) and fluent baseline (β = -1.67, $p < 0.001$). Similarly, adults were also more likely to infer that the game was gendered in the gendered-surprise condition relative to the control surprise condition (β = 1.81, $p < 0.001$) and fluent baseline (β = 2.09, $p < 0.001$). They did not differentiate the control surprise and fluent baseline conditions on either measure (all *ps* > 0.22).

Discussion

Experiment 3 demonstrates that adults readily integrate surprisal information and statistical covariance. After hearing a surprisal reaction (in both surprisal conditions), adults rated the target's behavior as weirder in the teacher's eyes, com-

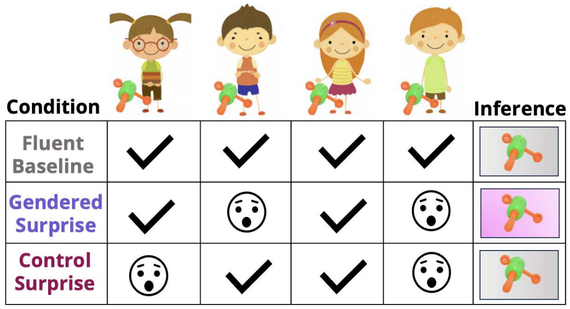

Figure 4: A schematic showing the logic for each of the three conditions for Experiment 3 (check marks indicate a fluent reaction, surprise icons indicate a surprisal reaction). Note, this schematic shows only the boy-target order for simplicity.

pared with the fluent baseline. However it was only when those surprisal reactions covaried with gender (gendered-surprise) that adults inferred that the novel game was gendered and also used gender to infer who was teased. Adults did not infer that the game was gendered or use gender to infer who was teased in the control-surprise condition. Interestingly, adults also rated that the teacher thought the target behavior was weirder in the gendered-surprise condition than the control-surprise condition (despite equivalent amounts of surprise), which may be further evidence that they are inferring a possible norm in the gendered-surprise condition.

General Discussion

Across 3 experiments, we see consistent evidence that even well-intentioned feedback about a child's behavior can nonetheless reveal one's underlying expectations. Across conditions, the feedback was closely matched but the addition of markers of surprise and production difficulty (interjection "oh" and disfluencies "um") was sufficient to generate differentiated inferences in both children and adults. Experiment 1 demonstrates that children by age 6 to 7 use conversational markers of others' surprise to reason about whether a boy made a stereotypical or counter-stereotypical choice. Experiment 2 demonstrates that adults and older children use these same cues to learn a novel expectation and predict social consequences. Experiment 3 combines these approaches to show that adults use others' conversational surprise to learn a novel gendered expectation. This work contributes to the recent "Emotion as Information" framework that argues emotional expressions are useful not just for reasoning about emotions, but for learning unobservable states in the social and physical world (Wu et al., 2021).

While these surprisal inferences clearly reflect reasoning about the speaker's expectations, we remain agnostic as to whether they are seen as capturing descriptive or prescriptive information. Either way, these cues could serve as one mechanism for transmission of social stereotypes. We have focused on gender stereotypes as a pernicious and naturalistic case study of stereotyped expectations, however our proposal applies to learning a variety of expectations and stereotypes.

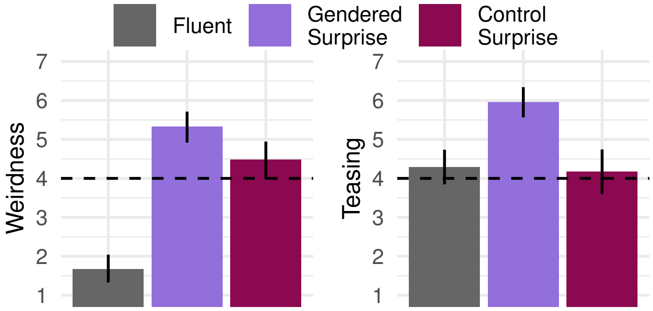

Figure 5: Adults' judgements for the weirdness (left) and teasing (right) measures for each of the three conditions in Experiment 3 (note we did not collect developmental data for this experiment). For teasing, higher values indicate selecting the character who was same gender as the target. Error bars show bootstrapped 95% confidence intervals.

Our work adds new insights to the literature on belief transmission that demonstrates the surprising efficacy of subtle linguistic framing (e.g., Chestnut et al., 2021; Cimpian et al., 2007; Rhodes et al., 2012). Specifically, the current work shows that it is not just what we say, but how we say it that matters. Our results suggest children can use subtle features of casual language to make deep the about speakers' mental states. One exciting question for future research is the extent to which children are doing so by beginning to model the production process that generated speech to draw inferences about the presence of these markers. Alternatively, participants could be reasoning about these cues more heuristically, or even relying on other inferences about a speaker's underlying discomfort or dishonesty (Fox Tree, 2002).

Across Experiments 1 and 2, the data suggest 4- to 5-year-olds are not reliably using others' surprisal to draw inferences. While even infants connect surprisal reactions with expectations about the physical world (Wu et al., 2024), it is possible that younger children in our experiments struggle to connect their representation of the adult's expectations with an additional representation of others' behaviors and mental states. We note that the developmental pattern we observe is consistent with related work on reasoning about an agent's competence on the basis of others' facial expressions of surprise (Asaba et al., 2020). However, it is also possible that younger children can draw the key inference, but their performance is burdened by task demands.

Conversations carry a wealth of social information, especially conveying a speaker's underlying beliefs (e.g., Rhodes et al., 2012). Even well-meaning or explicitly egalitarian messages can sometimes still carry pernicious social messages (Chestnut et al., 2021). Children burgeoning abilities to extract underlying belief information from language helps them learn about the social world very quickly, which might be unfortunate in cases where adults are inadvertently conveying stereotype information.

References

Asaba, M., Wu, Y., Carrillo, B., & Gweon, H. (2020). You're surprised at her success? Inferring competence from emotional responses to performance outcomes. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, 2650–2656.

Bian, L., Leslie, S.-J., & Cimpian, A. (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science*, *355*(6323), 389–391.

Casillas, M. (2014). Taking the floor on time. *Language in Interaction: Studies in Honor of Eve V. Clark*, *12*.

Chestnut, E. K., Zhang, M. Y., & Markman, E. M. (2021). "Just as good": Learning gender stereotypes from attempts to counteract them. *Developmental Psychology*, *57*(1), 114.

Cimpian, A., Arce, H.-M. C., Markman, E. M., & Dweck, C. S. (2007). Subtle linguistic cues affect children's motivation. *Psychological Science*, *18*(4), 314–316.

Eisenberg, N., Murray, E., & Hite, T. (1982). Children's reasoning regarding sex-typed toy choices. *Child Development*, 81–86.

Fox Tree, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, *34*(6), 709–738.

Fox Tree, J. E. (2002). Interpreting pauses and ums at turn exchanges. *Discourse Processes*, *34*(1), 37–55.

Freeman, N. K. (2007). Preschoolers' perceptions of gender appropriate toys and their parents' beliefs about genderized behaviors: Miscommunication, mixed messages, or hidden truths? *Early Childhood Education Journal*, *34*, 357–366.

Kane, E. W. (2006). "No way my boys are going to be like that!" Parents' responses to children's gender nonconformity. *Gender & Society*, *20*(2), 149–176.

Meyer, W.-U., Reisenzein, R., & Schützwohl, A. (1997). Toward a process analysis of emotions: The case of surprise. *Motivation and Emotion*, *21*, 251–274.

Moty, K., & Rhodes, M. (2021). The unintended consequences of the things we say: What generic statements communicate to children about unmentioned categories. *Psychological Science*, *32*(2), 189–203.

Rhodes, M., Leslie, S.-J., & Tworek, C. M. (2012). Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences*, *109*(34), 13526–13531.

Schützwohl, A., & Reisenzein, R. (1999). Children's and adults' reactions to a schema-discrepant event: A developmental analysis of surprise. *International Journal of Behavioral Development*, *23*(1), 37–62.

Skinner, A. L., Olson, K. R., & Meltzoff, A. N. (2020). Acquiring group bias: Observing other people's nonverbal signals can create social group biases. *Journal of Personality and Social Psychology*, *119*(4), 824.

Wu, Y., Baker, C. L., Tenenbaum, J. B., & Schulz, L. E. (2018). Rational inference of beliefs and desires from emotional expressions. *Cognitive Science*, *42*(3), 850–884.

Wu, Y., Merrick, M., & Gweon, H. (2024). Expecting the unexpected: Infants use others' surprise to revise their own expectations. *Open Mind*, *8*, 67–83.

Wu, Y., Schulz, L. E., Frank, M. C., & Gweon, H. (2021). Emotion as information in early social learning. *Current Directions in Psychological Science*, *30*(6), 468–475.