Could both be right? Children's and adults' sensitivity to subjectivity in language

Ruthe Foushee (foushee@berkeley.edu) **Mahesh Srinivasan (srinivasan@berkeley.edu)**
Department of Psychology, University of California, Berkeley, Berkeley, CA 94704

Abstract

While some word meanings, like "spotted," depend on intersubjectively accessible properties of the world, others like "pretty" invoke speakers' subjective beliefs. We explored children and adults' sensitivity to the subjectivity of a range of adjectives, including words like "spotted" and "pretty," but also words like "tall," which are evaluated relative to a standard. Participants saw two speakers who had independently experienced sets of exemplars of a novel object kind disagree about whether a critical exemplar was, e.g., "tall," "pretty," and "spotted." In Experiments 1 and 3, speakers had seen distinct sets of exemplars, while in Experiments 2 and 4, the sets were identical. Adults always judged disagreements over words like "pretty" as *faultless*—indicating that both speakers "could be right"—and permitted less faultless disagreement for ones like "tall" when the speakers had experienced identical sets of exemplars. Strikingly, children did not respond in an adult-like manner until age 8 or 9, but their explanations for speakers' conflicting assertions suggested some sensitivity to the kinds of knowledge relevant for evaluating different adjectives.

Keywords: metalinguistic development; theory of mind

Introduction

What goes into our understanding of what other people say? While comprehension of some words, like "spotted" or "striped," does not appear to rely on our knowledge of others' beliefs and experiences, comprehension of other words, like "pretty" or "tasty," does. Thus, when someone says that they saw a "spotted bird," we can understand what they mean by leveraging our understanding of what "spotted" and "bird" refer to. We understand that the bird must have *some spots* on it to be called "spotted," and that were we also to see the bird, we would agree with the speaker's description. The meaning of "spotted" is in this sense *intersubjective*, based on properties in the external world all speakers can access. In contrast, when someone says that they saw a "pretty bird," it is not immediately apparent what property of the bird she is describing, nor that we would agree that the same bird is "pretty." This is because the meanings of words like "pretty" are not intersubjective, but instead depend on the speaker's belief.

The present studies are motivated by the idea that many words—far beyond clearly subjective predicates of personal taste like "pretty," "tasty," and "funny"—are interpreted relative to their speakers. In particular, we consider the cases of *relative adjectives* like "big" and "cold," which have to be interpreted relative to the nouns they modify: a "big bird" is smaller than a "big lion." We test whether these predicates are *also* interpreted relative to the prior experiences and beliefs of speakers. Do we adjust the imagined size of a "big apartment" or temperature of a "cold day" depending on the sample of apartments and weather we believe our interlocutor has experienced? Speakers may have different thresholds for calling an apartment "big" or a day "cold" depending on their prior experiences, such that they may not always agree about whether a specific apartment or day is "big" or is "cold."

As we review below, the fact that speakers may have different things in mind when using words like "big" and "pretty" may pose a challenge for successful communication. This is especially true in light of evidence that children (and even adults) have an overarching tendency toward *naive realism*; i.e., to behave as though their own perception reflects reality and their judgments are objective (Ross & Ward, 1996).

Background

Previous work demonstrates that children have sophisticated knowledge of relative adjectives, but leaves open whether they incorporate information about their interlocutors into their interpretations. Four-year-olds understand that the meanings of words like "big" and "tall" depend on distributions of referents within a given class. For example, they appropriately identify "tall pimwits" as ones at the higher end of the distribution of *only* pimwit heights, even if that means ignoring other, taller objects (Barner & Snedeker, 2008). Five-year-olds also understand that the frame of reference for what counts as "high" or "low" varies with the class of object in question (e.g., "high" for a bird is much higher than "high" for a bunny: Smith, Cooney, & McCord, 1986).

One reason to think that children may have difficulty interpreting the meanings of words that depend on speakers' beliefs is that they appear to begin life heavily influenced by naive realism, and behave as if their own construal of a stimulus will be shared by others. In one study, for example, children were shown an image that was then covered so only an unidentifiable quadrant of it was left visible. Four-year-olds predicted that others would still be able to identify the largely occluded image, seemingly discounting their previous subjective access to it in full (Taylor, Cartwright, & Bowden, 1991).

In general, the literature suggests that young children might be able to understand *why* people say the things they say, but still have difficulty thinking that others can have different meanings for words than they do, perhaps due to their fundamental assumptions about language itself. For example, although toddlers recognize that others might want a different snack from one that they themselves find delicious (Repacholi & Gopnik, 1997), they judge statements of unconventional snack opinions like "ice cream is yucky" as unacceptable well into the preschool years (Holubar, 2015). Thus, preschool-aged children understand that others may have different preferences, but struggle with understanding that an unqualified statement about a preference that they think is wrong can still be "right." Children's eventual success on false belief tasks (e.g., Wellman & Liu, 2004) demonstrates their understanding that an individual's experience leads to their beliefs. But

it may be more challenging for children to understand that experience might result in speakers of the same language having different meanings for the same word (e.g., such that speakers have different temperatures in mind when they talk about a "cold day"). Consistent with this, studies on children's beliefs about conventionality in language have argued that children expect object labels to be shared by other speakers of their language, even given evidence to the contrary (e.g., when other speakers were absent when a novel object's label was taught: Diesendruck, 2005).

While adults "succeed" on some of these tasks that stump children, they are also not immune to the influence of naive realism, suggesting some form of continuity over development. For example, adults often overestimate the prevalence of their own attitudes in the general population, and are reluctant to attribute those attitudes to their own subjective experience, rather than to objective features of the world (see Ross & Ward, 1996 for a review). When it comes to language, adults have the metalinguistic knowledge to be able to explicitly judge words and phrases like "pretty shirt" as *subjective* when they are presented in the absence of a referent that they could evaluate (Scontras, Degen, & Goodman, 2016). However, to our knowledge there have been no empirical investigations of whether adults permit different word meanings for speakers when the adults themselves are confident of whether the word applies: e.g., when they are confident that a particular shirt is "pretty" or "big." Additionally, no studies have explored whether adults' tolerance of disagreement about different word uses is influenced by their knowledge of a speaker's relevant prior experience.

The Present Studies

Here, we explore adults' and children's sensitivity to linguistic subjectivity across four experiments. In particular, we ask whether, in addition to considering the real-world distribution of a specific noun's referents along a given dimension, listeners also interpret adjectives like "tall" relative to what they know of the distribution that the *speaker* has experienced. To test these ideas, we manipulate whether two speakers experience different distributions of exemplars of a novel object kind (Exps. 1 and 3), or identical distributions of exemplars (Exps. 2 and 4), and assess whether this affects adults' (Exps. 1 and 2) and children's (Exps. 3 and 4) judgments of whether the two speakers can disagree about how to describe a novel *target exemplar* that they can both see.

Across our studies, the disagreeing assertions that participants judge involve adjective-noun phrases that describe the same target object: e.g., "That's [not] a tall pimwit." We introduce novel nouns, but use familiar *gradable adjectives* (GAs) that vary in how intersubjective versus subjective they are. We categorize these adjectives into three classes. Following Syrett, Kennedy, and Lidz (2009), we call words like "spotted" *absolute* GAs. These adjectives require their arguments to possess some minimal degree of a property, and their meaning is largely context-independent. We refer to context-*dependent* adjectives like "tall" as *relative* GAs (Syrett et al., 2009), and refer to adjectives like "pretty" as *subjective* GAs.

To assess individuals' appreciation of the subjectivity of these different kinds of adjectives, we obtain judgments of *faultless disagreement*: disagreements where neither person is wrong (Barker, 2013). Such judgments are closely correlated with direct measures of statements' subjectivity (Scontras et al., 2016), and offer a less metalinguistically demanding measure to use with children. In addition, we elicit qualitative explanations from participants to understand the sources of knowledge that they are drawing on when evaluating speakers' utterances. Critically, given that participants maintain visual access to the complete distribution of exemplars observed by both speakers, they are able to form their own evaluation of whether the adjective-noun phrase applies to the target exemplar that is the subject of the speakers' disagreement. Since they share this evaluation with only one of the two disagreeing speakers (e.g., one will call the pimwit "pretty" and one "not pretty"), we can interpret judgments of faultless disagreement as overcoming naive realism.

Experiment 1

Participants Twenty-five UC Berkeley undergraduates participated in Experiment 1 (18 women, 19.65–27.37, $M = 21.24$ years, $SD = 1.68$ years). All were native speakers of English and received course credit for their participation.

Stimuli and Methods

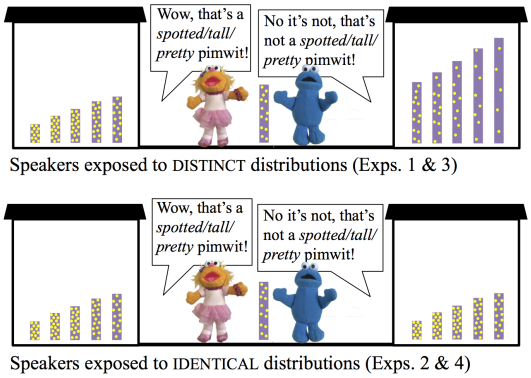


Figure 1: Schematic of experimental setup for Exps. 1–4.

Experimental Setup The stimuli were sets of eleven objects belonging to distinct novel kinds. Critical kinds were *pimwits*, thin purple cylinders ranging from 0.75 to 6.25 inches in height, and from densely to sparsely spotted, and *daxes*, blue and yellow spheres ranging from 0.5 to 3 inches in diameter, and from heavily to lightly striped. Each set was divided into two arrays composed of the five smallest and five largest exemplars, with the exemplar in the middle of the size distribution used as the critical target exemplar (Figure 1).

Participants sat across a table from two wooden house-like structures separated by a narrow stage. The experimenter sat behind the display and animated pairs of puppets representing familiar characters from Sesame Street, who she

explained could not see or hear anything that happened beyond their "classrooms" while they were inside them.

The experiment consisted of two blocks of two training trials each, followed by two blocks of three critical trials each and a post-test. In each block, two speakers were independently introduced to distinct arrays of a novel kind in their classrooms by the experimenter, before emerging to view and disagree about a new exemplar (the target) placed by the experimenter in the middle of the stage.

Training Trials The initial training trials familiarized participants with the paradigm, and provided practice with judging disagreements as faultless and not. In them, characters saw distinct sets of exemplars labeled with the same noun, and then 'disagreed' over a target exemplar that shared properties with both sets. In a faultless training trial, Dawn might see five *feps* that were matte white circles, while Big Bird saw five feps that were sparkly white squares. Dawn would exclaim that the target fep, a sparkly white circle, was "sparkly," while Big Bird would assert that it was "round" (a faultless disagreement). The non-faultless complement would consist of one speaker asserting the target exemplar was "white," while the other said it was "black." Participants received feedback for their answers on only the first block of training trials, and we recorded their judgments prior to feedback.

Critical Trials In critical trials, the characters were introduced to distinct sets of exemplars belonging to the same novel kind (pimwits or daxes). For example, Zoe might see five relatively short and densely spotted pimwits, while Cookie Monster saw five relatively tall and lightly spotted ones. Upon encountering the intermediate target pimwit, Zoe would assert that it was tall, which Cookie Monster would deny. Following the disagreement, participants answered whether each speaker was "wrong" or "could be right," and explained why. Responses where participants answered "could be right" for both speakers were coded as indicating faultless disagreement. For each novel kind, speakers disagreed over an absolute, relative, and subjective GA.

The order of the blocks, which speaker asserted the positive statement, and the block-internal order of the relative versus subjective disagreements were counterbalanced across participants. Disagreements over absolute GAs were always presented last to avoid invalidating one of the speakers. To prevent speakers from being degraded across blocks for wrong assertions, new speakers were introduced each block of trials.

Qualitative Explanations In each critical trial, we collected qualitative explanations of participants' evaluations of speakers' assertions. From explanations collected during piloting, we developed three primary codes to describe participants' responses. Trained coders identified whether each explanation made reference to apparently intersubjective properties of the target exemplar (OBJECT PROPERTY—e.g., "It *is* beautiful;" "There are dots on it"), the distinct arrays of exemplars the speakers had experienced (SPEAKER EXPERIENCE—e.g., "He saw tall pimwits and she saw short ones"), or the speaker's subjective evaluation of the object (SPEAKER OPINION—e.g., "He likes purple and she doesn't like spots."). Explanations could receive multiple codes.

Post-Test We directly assessed participants' own evaluation of the target exemplars in a post-test. Participants saw the entire distribution of exemplars, and answered whether the target exemplars were "spotted," "tall," "pretty," etc.

Results

Faultless Disagreement Disagreements over relative and subjective GAs were almost always judged as faultless (spotted: 24%, striped: 20%, tall: 100%, big: 100%, pretty: 100%, boring: 92%; see Figure 2). There were no significant differences between the proportions of faultless disagreement for the two adjectives in each class, so we collapse them here.

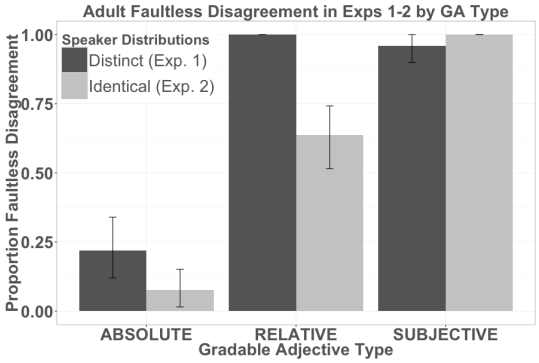


Figure 2: Adult rates of faultless disagreement judgment during critical trials in Exps. 1–2 by gradable adjective type (ABSOLUTE: "spotted," "striped;" RELATIVE: "tall," "big;" SUBJECTIVE: "pretty," "boring"). Participants in Exp. 1 judged speakers exposed to distinct distributions of exemplars, while participants in Exp. 2 judged speakers who had seen *identical* ones. Error bars for this and all plots indicate 95% bootstrapped confidence intervals ($k = 1000$).

Relation to Post-test The relation between participants' post-test evaluations and faultless disagreement judgments qualitatively distinguished the three classes of adjectives. For absolute GAs, participants uniformly answered "yes" when asked whether the target pimwit or dax was "spotted" or "striped," and typically answered that only the speaker who asserted the same could be right. For relative GAs, in contrast, while participants again *all* said that the target exemplar was "tall" or "big," they also *all* responded that both speakers could be right. Despite variability in participants' own evaluations of the critical items' beauty (91% said it was "pretty") or tedium (17% said it was "boring"), they almost always judged disagreements over subjective GAs as faultless.

Qualitative Explanations Participants for the most part cited distinct sources of knowledge to explain their judgments of utterances from different adjective classes (Figure 3). They referred to speakers' opinions (SPEAKER OPINION) *exclusively* when explaining their evaluations of utterances using

subjective GAs (and did so on 80% of all subjective trials).

We fit separate logit models to the data for the two remaining codes (OBJECT PROPERTY and SPEAKER EXPERIENCE) that were used in explanations regarding more than one adjective class, using GA type as our sole predictor. Participants were more likely to refer to object properties to explain absolute GA utterances ($\beta = 1.266$, $p < 0.001$), and less likely to cite them when explaining subjective GA judgments ($\beta = -3.462$, $p < 0.001$). Participants cited speakers' unique experiences most in explaining judgments of relative GA utterances, and were unlikely to do so to explain judgments of absolute GAs ($\beta = -1.153$, $p < 0.001$).

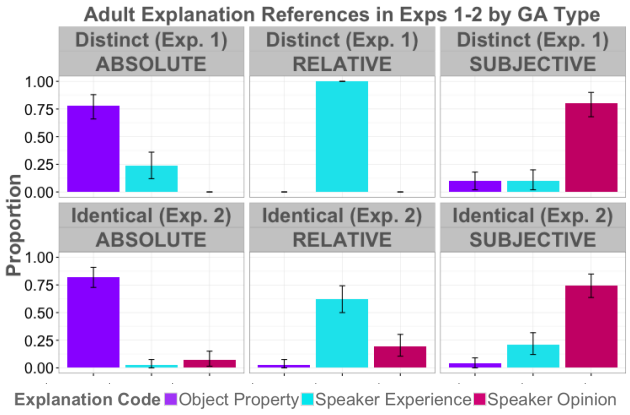


Figure 3: Adult reference to properties of the target exemplar (OBJECT PROPERTY), speakers' distinct experiences of the object kind (SPEAKER EXPERIENCE), and speakers' subjective opinions (SPEAKER OPINION), in explaining their evaluations of assertions. Panels reflect proportions of each code for explanations regarding a given GA type (in columns) in a given experiment (in rows).

Experiment 2

Having demonstrated that adults readily judge disagreements over relative and subjective GAs as faultless when speakers have experienced distinct distributions of exemplars, we explored the limits of listeners' acceptance of subjective meanings by *equating* the disagreeing speakers' experiences.

Participants 33 undergraduate adults (26 women, 18.10–39.83 years, $M = 20.91$, $SD = 3.52$) participated.

Stimuli and Methods

The experimental paradigm was identical to that of Experiment 1 with two changes: 1) speakers saw identical distributions of exemplars in their respective classrooms, and 2) we introduced an additional, plain (i.e., not spotted or striped) target exemplar for each novel kind about which the speakers only disagreed using our subjective GAs.[1]

Results

Faultless Disagreement We fit a logit model to the faultless disagreement judgment data with GA type as a predictor. Participants were highly likely to permit faultless disagreement for subjective ($\beta = 7.377$, $p < 0.001$) and relative ($\beta = 3.061$, $p < 0.001$) GAs. They were unlikely to permit faultless disagreement over absolute ones ($\beta = -2.501$, $p < 0.001$).

Relation to Post-test As in Experiment 1, we see differences among the adjective classes in the relation between participants' own assertion of each GA and their permission of faultless disagreement over it. In the post-test, all participants judged the target pimwit with spots "spotted," "tall," and "pretty." 94% answered "yes" when asked if the plain target pimwit was "pretty." For the dax with stripes, 97% judged it "striped." 64% said it was "big," and only 9% said it was "boring," while 55% said that the plain dax was. Despite substantial variation in their own evaluations of the critical items, participants almost always permitted faultless disagreement for the subjective GAs. For the absolute GAs, which the vast majority of participants accepted as true of the critical items, participants permitted very little faultless disagreement, but judged disagreements over the relative GAs as faultless between half and three-quarters of the time (Figure 2).

Even when listeners do not have an explanation for speakers' differing standards for relative GAs, they may permit faultless disagreement due to the standard's uncertainty. Participants permitted more faultless disagreement over "big" (72% of participants), which a lower proportion (64%) agreed was true of the critical dax, and less over "tall" (56%) which all participants agreed was true of the pimwit.

Qualitative Explanations Fitting logit models to the data for each explanation code and GA type, participants again were likely to refer to object properties in explaining judgments over absolute GAs ($\beta = 1.541$, $p < 0.001$), but not relative ($\beta = -4.616$, $p < 0.001$) or subjective ($\beta = -4.616$, $p < 0.001$) ones. Participants were most likely to refer to speakers' experiences—even though they were identical—in their explanations for relative ($\beta = 3.960$, $p < 0.001$) and subjective ($\beta = 2.060$, $p < 0.01$) GAs, and least likely for absolute ones ($\beta = -3.466$, $p < 0.001$). Finally, participants were unlikely to refer to speakers' opinions in explaining disagreements over absolute GAs ($\beta = -2.501$, $p < 0.001$), but were likely to do so in explaining disagreements over relative GAs ($\beta = 1.096$, $p < 0.05$), and highly likely for subjective ($\beta = 3.521$, $p < 0.001$) ones as well (Figure 3).

Compared to Experiment 1, adults permitted less faultless disagreement for relative and absolute GAs when speakers had experienced identical distributions of exemplars (Figure 2). This was not the case for subjective GAs, which participants continued to permit faultless disagreement over.

Experiment 3

Experiment 3 followed up on the previous experiments with adults by exploring the developmental trajectory of linguistic subjectivity. We tested a large age range to span a broad swath of theory-of-mind and metalinguistic development.

[1] We included plain exemplars of the critical object kinds after most children during piloting for Exp. 3 cited the pimwit's spots to explain why the speaker who had denied it was pretty was wrong.

Participants Seventy-one children across three age groups participated (24 4–5.5 years: 15 girls, $M = 4.83$, $SD = 0.34$; 23 5.5–7 years: 8 girls, $M = 6.05$, $SD = 0.470$; 24 8–9.5 years: 14 girls, $M = 8.90$, $SD = 0.34$). Four children were excluded due to experimenter error or broken stimuli ($n = 2$).

Stimuli and Methods

Experiment 3 used the same method as Experiment 2, except that speakers experienced distinct exemplar distributions as they did in Experiment 1. We also only included one object kind, pimwits, to keep it a more manageable length for children.

Results

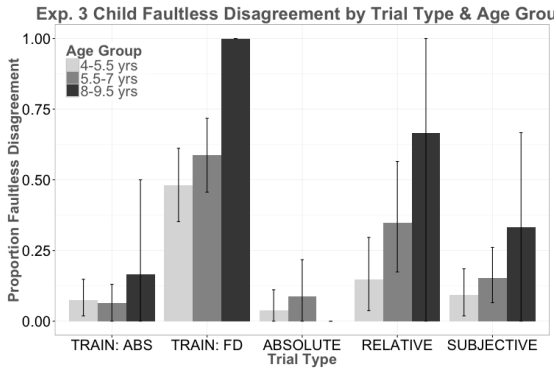


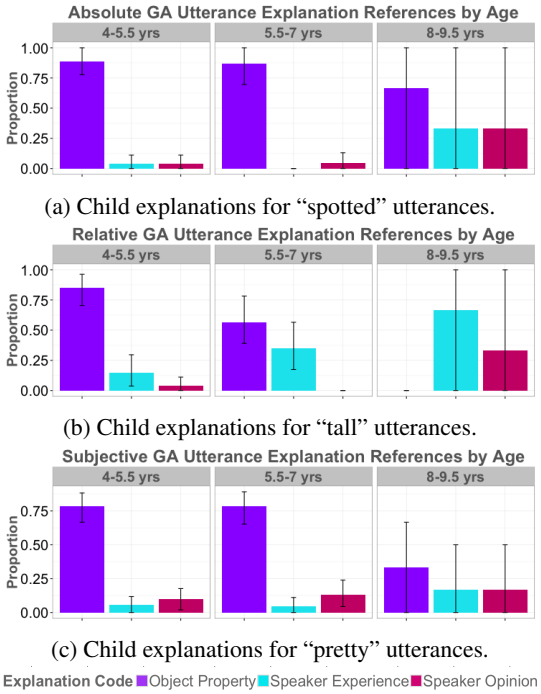
Figure 4: Child rates of faultless disagreement judgment in Exp. 3. TRAIN:ABS and TRAIN:FD trials were training intended to elicit non-faultless and faultless judgments, respectively. Participants made two judgments over the subjective GA "pretty," regarding a spotted as well as a plain pimwit, but there was no significant difference between rates of faultless disagreement between the two.

Faultless Disagreement We fit a logit model to the critical trial faultless disagreement data with GA and age. Children were significantly less likely to permit faultless disagreement for absolute GA "spotted" ($\beta = -7.646$, $p < 0.001$), and significantly more likely for relative and subjective GAs "tall" and "pretty" ("tall": $\beta = 2.047$, $p < 0.001$; "pretty" for spotted pimwit: $\beta = 1.170$, $p < 0.05$; "pretty" for plain pimwit: $\beta = 1.831$, $p < 0.001$). In general, they permitted faultless disagreement more with age ($\beta = 0.764$, $p < 0.001$).

We can think of children's initial judgment rates on the faultless disagreement training trials as baselines (Figure 4). Even in our oldest age group, rates of faultless disagreement on the critical trials are significantly below those of the faultless training trials (for relative trials: $t = -2.164$, $df = 31.373$, $p < 0.05$; for subjective trials: $t = -3.820$, $df = 70.616$, $p < 0.001$). While capable of judging disagreements as faultless, children were reluctant to do so when they themselves agreed with only one of the speakers.

Relation to Post-test 96% of children judged the target pimwit "spotted." More children judged it "pretty" (85%) than "tall" (49%). 55% answered that the plain pimwit was "pretty." For the absolute and relative GAs, we see roughly the same qualitative relation between post-test response and faultless disagreement judgments as with adults: greater post-test consensus meant less faultless disagreement.

Qualitative Explanations There appear to be some children across our age range who understood the source of knowledge most relevant for each GA, though children referred to object properties most frequently for all types until our oldest age group (Figure 5). We fit logit models to the data for each explanation code separately, with GA type, age, and their interaction as predictors. Children were highly likely to refer to properties of the objects in explaining absolute ($\beta = 2.543$, $p < 0.05$) and relative ($\beta = 3.758$, $p < 0.05$) GA utterances, and less likely to do so for relative ($\beta = -0.800$, $p < 0.01$) and subjective ones with age ($\beta = -0.642$, $p < 0.01$). They were least likely to refer to speakers' experiences in explaining absolute utterances ($\beta = -6.608$, $p < 0.01$), though more likely to refer to them at all with age ($\beta = 0.586$, $p < 0.05$). Finally, they became more likely to refer to speakers' opinions to explain subjective utterances as they got older ($\beta = 0.872$, $p < 0.05$).



(a) Child explanations for "spotted" utterances.

(b) Child explanations for "tall" utterances.

(c) Child explanations for "pretty" utterances.

Figure 5: Proportion of children's explanations in Exp. 3 receiving each qualitative code, by GA type and age group (in panels).

Experiment 4

Experiment 4 used the same method as above, with children at the older end of the age range. As in Experiment 2, the distributions that the two speakers saw were identical.

Participants Participants were 24 children 8–9.5 years of age (12 girls; $M = 9.09$, $SD = 0.44$).

Results

Faultless Disagreement Children permitted faultless disagreement on 98% of faultless training trials, and on *none* of non-faultless training trials. They did so most on subjective trials (67% of the time, 95% $CI : 53 - 80$), followed by relative (38%, $CI : 21 - 54$) and absolute (1%, $CI : 0 - 4$).

Children were least likely to permit faultless disagreement over absolute GA "spotted" ($\beta = -3.135$, $p < 0.01$), and highly likely for all other adjectives ("tall": $\beta = 2.625$, $p < 0.05$, "pretty" for spotted pimwit: $\beta = 4.052$, $p < 0.001$, "pretty" for plain pimwit: $\beta = 3.829$, $p < 0.001$).

Relation to Post-Test While all children judged the target pimwit "spotted" and "tall," they resembled adults by still permitting faultless disagreement over "tall" about half the time (and almost never over "spotted"). 79% and 92% participants judged the spotted and plain pimwits "pretty," respectively. Despite general consensus over their beauty, children responded more like adults in nonetheless permitting faultless disagreement over them at relatively high, equivalent rates.

Qualitative Explanations Children's explanations for absolute GA "spotted" were highly likely to receive the OBJECT PROPERTY code ($\beta = 2.398$, $p < 0.01$), while explanations of relative ($\beta = -2.734$, $p < 0.01$), and subjective ($\beta = -3.267$, $p < 0.001$) utterances were unlikely to. Explanations of absolute GA utterances were also unlikely to be coded as referring to SPEAKER EXPERIENCE ($\beta = -3.135$, $p < 0.01$), which was highly likely for relative GA utterances ($\beta = 2.625$, $p < 0.05$). Lastly, explanations of subjective GAs were likely to be coded as citing SPEAKER OPINION ($\beta = 4.005$, $p < 0.001$), in contrast to explanations about absolute GA utterances ($\beta = -3.135$, $p < 0.01$).

General Discussion

We tested theoretical claims about faultless disagreement arising when there is uncertainty about how and whether to assess something as, e.g., "pretty" or "tall" (Barker, 2013). We asked in particular whether individuals consider the reference distribution of their interlocutors in interpreting relative gradable adjectives. Adults reliably permitted faultless disagreement over relative and subjective GAs when two speakers had had distinct personal experiences. Rates of faultless disagreement decreased for relative GAs when speakers had experienced identical distributions, but did not disappear altogether, suggesting that adults were instead permitting faultless disagreement out of an understanding of the standard's uncertainty. Together, these findings provide evidence for the consideration of speaker at the level of semantics, as well as adults' sensitivity to the potential for differing standards of more than just explicitly context-dependent adjectives.

The development of sensitivity to linguistic subjectivity appears to be exceptionally prolonged: for the most part, children 'sided' with the speaker who voiced their own evaluations. Two factors might explain the apparent gap between adults and children in our studies. First, previous work suggests that children better grasp subjectivity when they are able to reason about an individual's *goals* (Holubar, 2015), a dimension that was absent from our experiments. To this end, our ongoing studies explore the effect of goal-oriented contexts (e.g., choosing who you would want to be friends with or learn from), which might be more sensitive to children's nascent understanding of the different implications of being "wrong" about whether something is "spotted" as opposed to "pretty." Second, there may be more continuity between adults and children than it appears. Adults' permission of faultless disagreement and explanation of different GA utterances may reflect social pressures and metalinguistic knowledge, rather than a core belief that their own evaluation is subjective. When it comes to predicates of personal taste, although adults may readily say that the meaning of "good" is subjective, such that both speakers can be right, we have all had the experience of disagreeing about whether a movie or song is "good." Future studies will test for possible continuity between children and adults by examining the contexts in which children may behave more like adults in their metalinguistic judgments of subjectivity, and the contexts in which adults may react similarly to children in their implicit commitment to intersubjectivity.

References

Barker, C. (2013). Negotiating taste. *Inquiry: An Interdisciplinary Journal of Philosophy*, *56*(2–3), 240–257.

Barner, D., & Snedeker, J. (2008). Compositionality and statistics in adjective acquisition: 4-year-olds interpret tall and short based on the size distributions of novel noun referents. *Child Development*, *79*, 594–608.

Diesendruck, G. (2005). The principles of conventionality and contrast in word learning: An empirical examination. *Developmental Psychology*, *41*(3), 451–463.

Holubar, T. F. (2015). *Children's reasoning about unconventional opinions as evidence for naive realism*. Unpublished doctoral dissertation, Stanford University.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology*, *33*(1), 12–21.

Ross, L., & Ward, A. (1996). Naive realism in everyday life: Implications for social conflict and misunderstanding. In T. Brown, E. S. Reed, & E. Turiel (Eds.), *Values and knowledge* (pp. 103–135). Hillsdale, NJ: Erlbaum.

Scontras, G., Degen, J., & Goodman, N. D. (2016). Subjectivity predicts adjective ordering preferences. *Open Mind*.

Smith, L. B., Cooney, N. J., & McCord, C. (1986). What is "high"? The development of reference points for "high" and "low". *Child Development*, *57*, 583–602.

Syrett, K., Kennedy, C., & Lidz, J. (2009). Meaning and context in children's understanding of gradable adjectives. *Journal of Semantics*, *27*(1), 1–35.

Taylor, M., Cartwright, B., & Bowden, T. (1991). Perspective taking and theory of mind: Do children predict interpretive diversity as a function of differences in observers' knowledge? *Child Development*, *62*(6), 1334–1351.