Proceedings of SALT 35: 881–900, 2025

First-person metalinguistic disagreements update the standard of precision, both up *and* down *

Yifan Wu
Cornell University

Helena Aparicio
Cornell University

Abstract The *standard of precision* (SoP) is the discourse parameter governing how much imprecision is tolerated in a given context. Speakers can implicitly negotiate the SoP through metalinguistic disagreements (MDs)—moves that challenge the assertability of an (im)precise utterance in order to signal that the SoP should be updated. Prior work has claimed that such negotiations are asymmetric: MDs can effectively raise the SoP but cannot lower it (e.g., Klecha 2018; Lewis 1979). However, recent experimental work has failed to find evidence for this purported asymmetry (Wu & Aparicio 2025a), motivating further empirical investigation into the validity of this claim. We present results from two studies in which participants engaged in metalinguistic (dis)agreements over (im)precise utterances. Our findings reveal a bidirectional update pattern: MDs can shift the SoP both upward, as previously claimed, and downward, contrary to earlier accounts (Klecha 2018; Lewis 1979). However, these two update types differ qualitatively: participants making imprecise utterances tend to abandon their original stance in favor of the disagreeing view, whereas participants making precise utterances tend to accommodate the disagreement without abandoning their own position. We suggest that this contrast may underlie earlier intuitions about the unidirectionality of SoP updates.

Keywords: Imprecision, disagreements, maximum standard absolute adjectives.

1 Introduction

Strict adherence to literal truth is not a prerequisite for successful communication. In fact, speakers often deviate from truth in favor of utterances that convey propositions which, while knowingly false, are sufficiently close to the truth to be perceived as informative, and sometimes even more felicitous than their strictly true counterparts. This phenomenon, known as imprecision or loose talk (Aparicio, Xiang & Kennedy 2015; Aparicio Terrasa 2017; Kao, Wu, Bergen & Goodman 2014; Kennedy 2007; Klecha 2018; Krifka 2002, 2007; Lasersohn 1999; Lauer 2012, 2013; Leffel, Xiang

* We are grateful to the members of the Cornell LiMe Lab, Cornell C. Psyd and Cornell Cognitive Science, to the anonymous reviewers and audiences at SALT 35, and to Julian Grove for insightful discussions and valuable feedback.

& Kennedy 2016; Lewis 1979; Ronderos, Noveck & Falkum 2024; Solt 2015; Syrett, Kennedy & Lidz 2010; Wu & Aparicio 2025a,b: a.o.), is illustrated in Figure 1, where Alex chooses to describe the bottle as *empty* despite knowing that it contains some amount of water.

Alex: This bottle is empty.

Figure 1 Imprecise description of a bottle containing some water.

Imprecision can affect a range of lexical categories, including maximum-standard absolute adjectives (e.g., '*empty*'), round numerals (e.g., '*one thousand*'), and prepositions (e.g., '*in*'). Across these domains, it has been argued that the acceptability of an imprecise utterance is modulated by features of the utterance context, such as conversational goals (Aparicio Terrasa 2017; Burnett 2014; Mathis & Papafragou 2022; Ronderos et al. 2024; van der Henst, Carles & Sperber 2002) or speaker personae (Beltrama & Schwarz 2021, 2022, 2024; Beltrama, Solt & Burnett 2023).

Following Lewis (1979), we assume that the degree of imprecision tolerated in a given context is regulated by the *standard of precision* (SoP), a latent discourse parameter that forms part of the conversational score. Lewis further observes that the value of the SoP can be implicitly negotiated by interlocutors through disagreements such as (1a), where Andy takes issue with Alex's imprecise utterance in (1a-i) and challenges it via the metalinguistic denial in (1a-ii). While such challenges can succeed in raising the SoP, Lewis argues that parallel attempts to lower it are ineffective: challenging a (more) precise utterance with a metalinguistic denial such as (1b-ii) does not succeed at lowering the SoP. The claim is that updates to the SoP are asymmetric: raising the standard is considerably easier than lowering it, at least through the kind of implicit negotiation illustrated in (1a).[1]

(1) Metalinguistic Disagreements:

a. **Upward update**

i. *Alex:* This bottle is empty. [Description of Figure 1]

ii. *Andy:* I disagree. This bottle is not empty.

b. **Downward update**

i. *Alex:* This bottle is not empty. [Description of Figure 1]

1 See also Klecha (2018) for a similar argument.

ii. *Andy:* I disagree. This bottle is empty.

However, recent experimental work has failed to find evidence for the asymmetry in SoP updates claimed by Lewis. In particular, Wu & Aparicio (2025a) do not detect upward adjustments in metalinguistic disagreements such as (1a); instead, they find that imprecise assertions continue to be perceived as felicitous even after a metalinguistic denial, suggesting that such denials do not trigger an upward shift in the SoP. One possibility is that the lack of upward updates was an artifact of the experimental task. In that study, participants acted as bystanders to the disagreement dialogues rather than as interlocutors, which may have inflated the rates of *faultless disagreement* responses (Barker 2013; Kaiser & Rudin 2020, 2021; Kennedy 2013; Kölbel 2004), i.e., judgments indicating that both speakers could be right. Be that as it may, Wu & Aparicio only tested precisification dialogues. Properly evaluating the asymmetry claim requires examining both cases that attempt to strengthen the SoP and parallel moves aimed at weakening it.

In what follows, we present two studies that directly test whether SoP updates are asymmetrical by examining both strengthening and weakening updates, as well as the conditions under which such asymmetries may arise. Participants engaged in *first-person* dialogues with an interlocutor who exhibited either a preference or a dispreference for imprecision. In Experiment 1, participants interacted with an imprecise interlocutor, whereas in Experiment 2 they were exposed to a precise one. We were interested in whether participants would adjust the SoP to align with that of their interlocutor. We consider three hypotheses. Hypothesis 1 (H1) assumes that updates to the SoP are asymmetric, in line with previous claims in the literature (Klecha 2018; Lewis 1979). On this account, exposure to a precise interlocutor should lead to upward updates, whereas interaction with an imprecise interlocutor should fail to induce a corresponding downward shift. H1 therefore predicts raising effects in Experiment 2 and no differences in Experiment 1. Hypothesis 2 (H2) posits that metalinguistic disagreements can update the SoP bidirectionally, exerting both upward and downward pressure on standards. This hypothesis predicts weakening effects in Experiment 1 and raising effects in Experiment 2. Finally, Hypothesis 3 (H3), building on the bystander results of Wu & Aparicio (2025a), states that first-person metalinguistic disagreements, like bystander ones, do not trigger updates to the SoP. Accordingly, H3 predicts no effects in either experiment.

Our results show that first-person metalinguistic disagreements can update the SoP both upward, as previously claimed, *and* downward, contrary to earlier claims, supporting H2. However, we find that these updates are qualitatively different: imprecise participants display a more categorical update pattern; they are more likely to judge their original imprecise utterance as wrong in favor of the disagreeing view. Precise participants, on the other hand, accept the disagreeing view without

abandoning their own. We speculate that these qualitative differences might underlie previous intuitions in the literature about the unidirectionality of precisification (Klecha 2018; Lewis 1979).

The remainder of this paper proceeds as follows. In Section 2, we present Experiment 1; Section 3 details Experiment 2. Finally, Section 4 provides a general discussion of our findings and Section 5 concludes the paper.

2 Experiment 1

Experiment 1 tests whether the SoP can be lowered or weakened through first-person metalinguistic disagreements. Specifically, we investigate whether interacting with an imprecise interlocutor prompts participants to adjust toward a lower SoP.

2.1 Materials, design and procedure

Our experiments test maximum standard absolute adjectives. We adopted a set of 24 five-point scales used by Wu & Aparicio (2025a), representing varying degrees along maximum-standard absolute adjectival dimensions (see Figure 2). The visual stimuli had previously been normed by Wu & Aparicio to ensure that the lower scale points (S1-S4) are compatible with imprecise interpretations.

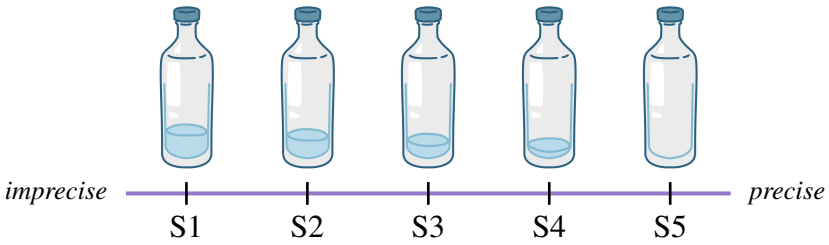
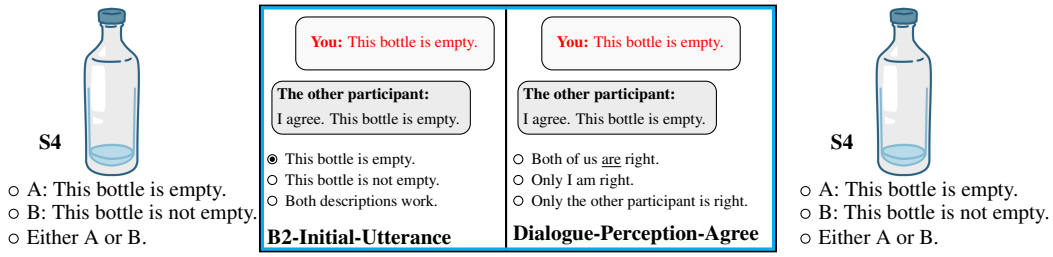


Figure 2 Five scale points corresponding to the scale '*empty bottle*'.

Our experiment tested S4, the near-maximal scale point that displayed the highest tolerance for imprecision in Wu & Aparicio (2025a)'s norming study. We also included S5 as a control, which was compatible only with a precise interpretation of the predicate (e.g., *empty*), and S1, the scale point that displayed the least tolerance for imprecision, as a filler. Experiment 1 followed a three-block adaptation paradigm (Heim, Peiseler & Bekemeier 2020). In the pre-exposure phase, **Block 1** (**B1**), participants provided interpretational preferences for each scale point by selecting one of three options: '*A: This* [OBJECT] *is* [ADJECTIVE]' (e.g., '*This bottle is empty*'), '*B: This* [OBJECT] *is not* [ADJECTIVE]' (e.g., '*This bottle is not empty*'),

or '*Either A or B*' (see Figure 3(a)). Block 1 therefore provided a baseline for participants' judgments when evaluating the stimuli in isolation.



(a) Block 1 (b) Block 2 (c) Block 3

Figure 3 Experiment 1 item example.

In **Block 2** (**B2**), participants engaged in a conversation about the same object they had previously seen in B1. Participants were told they would be remotely paired with a human participant (i.e., *the other participant*) who could see the same object on the screen as them. As shown in the B2-Initial-Utterance panel of Figure 3(b), their task was to describe the object to their interlocutor by choosing one of three options: '*This* [OBJECT] *is* [ADJECTIVE]' (e.g., '*This bottle is empty*'), '*This* [OBJECT] *is not* [ADJECTIVE]' (e.g., '*This bottle is not empty*'), and '*Both descriptions work*'. This initial selection was later used as a baseline, since it allows us to measure participants' interpretational preferences at the onset of the discourse. In reality, participants interacted with a chatbot (*bot*) designed to simulate conversational exchanges in real time. The bot was programmed to display specific preferences. Experiment 1 featured an *imprecise bot* (blue in Figure 4), which accepted the predication at S4 and S5 (e.g., '*This bottle is empty*') and always rejected the predication at S1 (e.g., '*This bottle is empty*') for realistic simulation. The bot agreed or disagreed with the participant's initial utterance based on its own preferences. The example in Figure 3(b) features an imprecise participant who initiates the dialogue by accepting the predication at S4, and an imprecise bot (Experiment 1) that agrees with the participant's description of the object. After the dialogue, participants were asked to judge which interlocutor was right by selecting one of three answers: '*Both of us are/can be right*', '*Only I am right*', or '*Only the other participant is right*'. The first option was phrased as '*Both of us are right*' for agreement dialogues, see the Dialogue-Perception-Agree panel of Figure 3(b), whereas for disagreement dialogues, it was phrased as '*Both of us can be right*' to indicate that the disagreement was perceived as *faultless* (Barker 2013; Kaiser & Rudin 2020, 2021; Kennedy 2013; Kölbel 2004), where neither interlocutor is

at fault. The aim of this task was to determine whether participants engaged in first-person dialogues perceive disagreements in a manner comparable to bystander participants in Wu & Aparicio (2025a).

Finally, **Block3** (**B3**), the post-exposure block, was an exact replication of B1 (see Figure 3(c)). B3 therefore allowed us to examine whether the type of dialogue participants were exposed to in B2 modulated selection rates in B3 relative to B1.

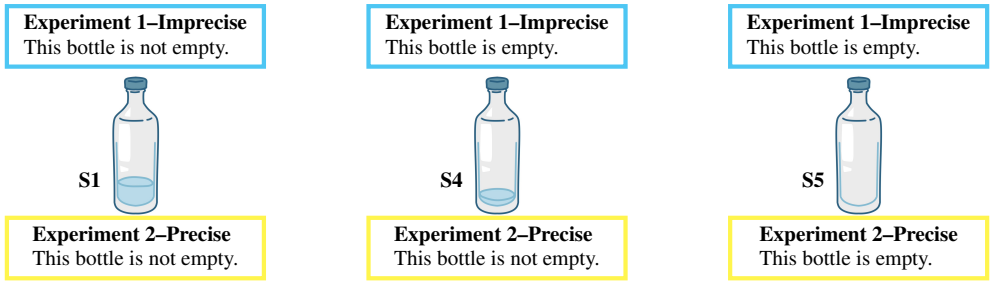


Figure 4 Bot preferences.

Additionally, we incorporated 24 color fillers from Wu & Aparicio (2025b). These fillers served as unambiguous baseline controls to ensure participant engagement and task comprehension, and were subsequently used as attention checks. In half of these trials, the depicted object was a perfect match for both the noun and the property described in the predication, e.g., the left panel of Figure 5 required a categorical acceptance of the sentence '*This flag is blue*', whereas the other half featured objects that were clearly incompatible with the adjectival property being discussed, e.g., the right panel of Figure 5 required a categorical rejection of the sentence '*This fish is green*'. To maintain the integrity of the simulated interaction, the bot was programmed to be always factual on these fillers, responding according to the objective truth. Participants' selections in B1 and the B2-initial utterance allowed us to flag and exclude individuals who provided non-factual responses to these unambiguous controls.

Figure 5 Filler item examples: Left: "*blue flag*"; Right: "*green fish*".

The experiment was administered remotely through the *PCIbex Farm* platform (Zehr & Schwarz 2018). Prior to the main experiment, participants provided informed consent, completed a demographic questionnaire, and engaged in three

practice trials designed to acclimate them to the experimental setup and response protocol.

2.2 Participants

Thirty native speakers of American English, who were at least 18 years old and self-reported having normal vision with no colorblindness, were recruited through the web platform *Prolific*. Participation was compensated at a rate of $15 per hour. Participants were excluded if their accuracy on attention check trials fell below 90%. One participant failed to meet the accuracy threshold, resulting in a final sample of 29 participants.

2.3 Predictions

Before detailing specific predictions, we first outline our interpretation of each response type. Selections accepting the predication (i.e., *Accept*-responses, e.g., '*This bottle is empty*' in B1, B2, and B3) indicate that the imprecise interpretation was accepted and that a low SoP was therefore adopted. Conversely, selections rejecting the predication (i.e., *Reject*-responses, e.g., '*This bottle is not empty*' in B1, B2, and B3) indicate that the imprecise interpretation of S4 was rejected and that a precise SoP was therefore adopted. Finally, *Either*-responses (e.g., '*Either A or B*' in B1 and B3 and '*Both descriptions work*' in B2) indicate that either predication is accepted and that both a high and a low SoP are deemed acceptable.

We begin with the critical scale point, S4. When participants first enter B2, they have not yet encountered any utterance from their interlocutor. We therefore do not expect changes in selection rates between B1 and the initial utterance choice in B2, nor do we expect participants to deviate from their isolation baseline SoP across all tested scale points at the beginning of each dialogue. Regarding B3, our predictions depend on the dialogue type encountered in B2. For agreement dialogues, we anticipate no updates to the SoP. Since the bot's response aligns with the participant's initial utterance, there is no need for a revision of the current threshold. Accordingly, selection rates in B3 are expected to remain similar to those observed in B2. Finally, for the metalinguistic judgment task concerning which interlocutor is correct, we expect participants to predominantly choose '*Both of us are right*', reflecting the interlocutors' agreement.

The disagreement dialogues constitute the critical test for the present study, as they provide insight into the effect of first-person metalinguistic disagreements on the SoP. H1 assumes no weakening effects. It therefore predicts that the proportion of rejections of imprecise predications at S4—indicating that participants held a higher SoP—should remain stable in B3 relative to the dialogue baseline in B2.

Accordingly, we expected no significant change in *Accept*-, *Reject*-, or *Either*-responses. H2 predicts that the proportion of rejections of imprecise utterances at S4 will decrease following the disagreement, as the bot's metalinguistic challenge should lead participants to lower their SoP. Consequently, we expect a decrease in *Reject*-responses in B3 compared to B2, with a corresponding shift toward *Accept*- or *Either*-responses as participants align with the lower standard. H3 predicts no updates following the disagreement; accordingly, selection rates in B3 should remain comparable to those in B2. Finally, given the exploratory nature of the dialogue perception task, we do not advance specific predictions for scale point S4.

We now turn to our predictions for S5. Scale point S5 functioned as a truth-conditional control involving no meaning uncertainty. We therefore anticipate that participants will provide near-ceiling *Accept*-responses in B1 and B2. Since the bot also accepts the predication at S5, we expect participants to largely be exposed to agreement dialogues, resulting in no significant shifts in selection rates between B2 and B3. In line with this, we also expect participants to select the option "*Both of us are right*" in the dialogue perception task. Regarding the filler S1, we anticipate higher proportions of *Reject*-responses than *Accept*-responses. Furthermore, imprecise interpretations should be significantly less acceptable than at S4, given its distance from the maximal endpoint.

2.4 Results and Discussion

Participants' responses across the three experimental blocks are illustrated in Figure 6. To statistically assess whether participants updated the SoP in the critical imprecise scale point (S4), we fit a series of mixed-effects logistic regression models. Each possible response type (*Accept*, *Reject*, and *Either*) was analyzed by binarizing responses into three distinct dependent variables. For the ACCEPT model, *Accept*-responses were coded as 1, while the other two responses were assigned 0. This coding scheme was mirrored for the other two possible responses (i.e., REJECT and EITHER models). Each model predicted the probability of the target response from the fixed effect of BLOCK (B1, B2, B3). By-item and by-participant random intercepts and random slopes by BLOCK were also included.

We start with the results pertaining to S4. B1 served as the isolation baseline for the B1-B2 comparison. As expected, utterance choices in B2 did not differ from B1 in any response type (ACCEPT: $\hat{\beta} = 0.20$, $SE = 0.32$, $z = 0.62$, $p > 0.1$; REJECT: $\hat{\beta} = -0.04$, $SE = 0.33$, $z = -0.13$, $p > 0.1$; EITHER: $\hat{\beta} = 0.29$, $SE = 0.37$, $z = 0.77$, $p > 0.1$). For the B2–B3 comparison, B2 served as the dialogue baseline for examining the discursive effect of metalinguistic disagreements on the SoP. Model outputs reveal that participants' *Reject*-responses (Figure 7(d)) significantly decreased ($\hat{\beta} = -0.79$, $SE = 0.28$, $z = -2.77$, $p < 0.01$) and *Either*-responses

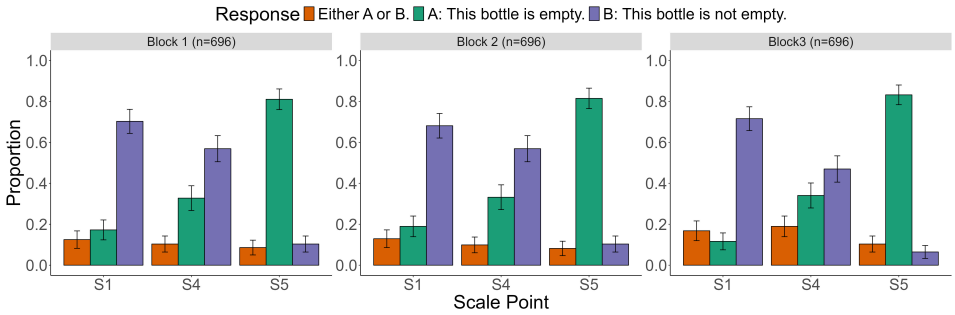
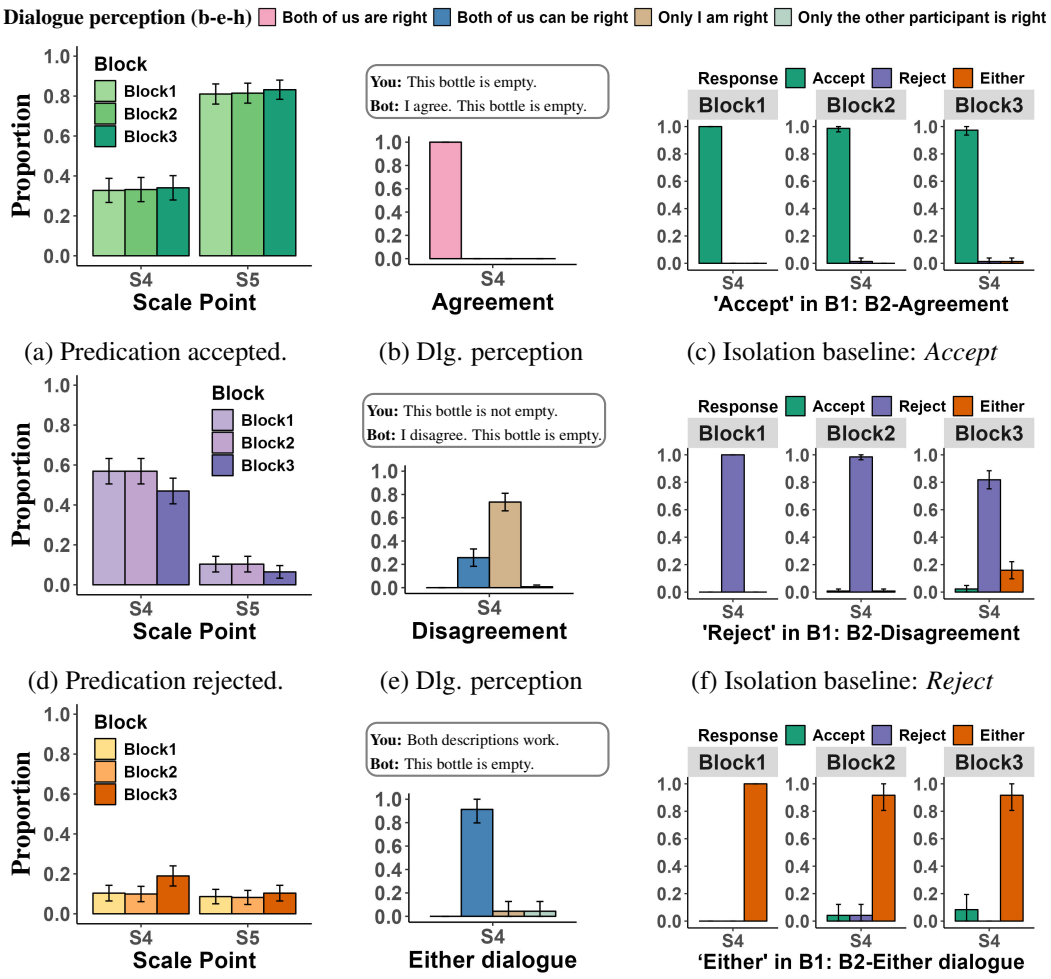


Figure 6 Experiment 1 participants' interpretational responses results.

(Figure 7(g)) significantly increased ($\hat{\beta} = 1.19$, $SE = 0.32$, $z = 3.70$, $p < 0.001$) in B3 compared to B2. No effects were detected for ACCEPT-responses ($\hat{\beta} = -0.03$, $SE = 0.31$, $z = -0.11$, $p > 0.1$; Figure 7(a)). Importantly, these effects were driven by the bot's disagreeing replies to the *Reject*-responses in B2. This can be observed in Figure 7(f), showing that participants who selected a *Reject*-response in B1 and B2, disfavored this response in B3—after the bot's disagreeing reply—in favor of *Either*-responses. The other two dialogue-types did not exert a significant influence on response patterns (see Figure 7(c) and Figure 7(i)). We also compared B3 directly to B1 as a robustness check; results showed a significant decrease in *Reject*-responses ($\hat{\beta} = -0.82$, $SE = 0.35$, $z = -2.33$, $p < 0.05$), a significant increase in *Either*-responses ($\hat{\beta} = 1.41$, $SE = 0.35$, $z = 4.06$, $p < 0.001$), and no change in *Accept*-responses ($\hat{\beta} = 0.16$, $SE = 0.31$, $z = 0.50$, $p > 0.1$).

No differences were found for the control S5 (all p's> 0.05). Filler trials (S1) displayed the lowest tolerance for imprecision, with overall lower *Accept*-responses compared to S4. No robust updates were observed for S1: *Accept*-responses increased between B1 and B2 ($\hat{\beta} = 0.78$, $SE = 0.36$, $z = 2.16$, $p < 0.05$), and decreased between B2 and B3 ($\hat{\beta} = -0.82$, $SE = 0.35$, $z = -2.34$, $p < 0.05$), but no significant difference between B1 and B3 was detected ($\hat{\beta} = -0.07$, $SE = 0.41$, $z = -0.18$, $p > 0.1$).

Regarding the dialogue perception results, agreement trials behaved as predicted: when the bot aligned with the participant's interpretation, participants predominantly selected the response '*Both of us are right*', as shown in Figure 7(b). Interestingly, in disagreement dialogues—where the bot challenged the participant's higher SoP—responses were split between '*Both of us can be right*' and '*Only I am right*' (see Figure 7(e)). These results align closely with the interpretive patterns observed in the block comparisons (Figure 7(f)): participants who perceived the disagreement as faultless ('*Both of us can be right*') accounted for the vast majority (87.5%) of participants who updated their SoP, shifting away from a *Reject*-response in B3;



(a) Predication accepted. (b) Dlg. perception (c) Isolation baseline: *Accept*

(d) Predication rejected. (e) Dlg. perception (f) Isolation baseline: *Reject*

(g) *Either*-responses. (h) Dlg. perception (i) Isolation baseline: *Either*

Figure 7 Experiment 1: (a, d, g): Binarized predication-judgment responses; (b, c, h): dialogue perception responses; (c, f, i) predication-judgment responses subsetting by response-type in Block 1.

however, the subset of participants who maintained their original strict standard was almost entirely composed of participants who viewed the disagreement as faulty, with 88.0% of non-updating trials associated with the '*Only I am right*' response. Finally, as expected, *Either Dialogues* were overwhelmingly judged as '*Both of us can be right*' (Figure 7(h)).

The finding that selection rates remained stable between B1 and the initial utterance in B2 confirms that participants entered the dialogue maintaining their baseline

interpretation. We now turn to the findings pertaining to trials involving disagreement dialogues at S4. The significant decrease in *Reject*-responses and corresponding increase in *Either*-responses from B2 to B3 indicate that first-person metalinguistic challenges can successfully trigger a downward update of the SoP. These results are therefore incompatible with H1, which posits that metalinguistic denials can raise but not lower the SoP, as well as with H3, which holds that first-person disagreements alone do not trigger updates. The current results are better accommodated by H2, which states that metalinguistic disagreements can bidirectionally update the SoP. Notably, the update results in a shift toward *Either*-responses rather than a categorical adoption of *Accept*-responses, alongside a perception of the disagreement as faultless (*Both of us can be right*). This suggests that the update does not involve a full displacement of the original standard, but rather an acknowledgment of the viability of the opposing view.

3 Experiment 2

Experiment 2 examines whether the SoP can be strengthened through first-person metalinguistic disagreements. In particular, the current study tests whether interacting with a precise interlocutor leads to an upward update of the SoP.

3.1 Materials, design and procedure

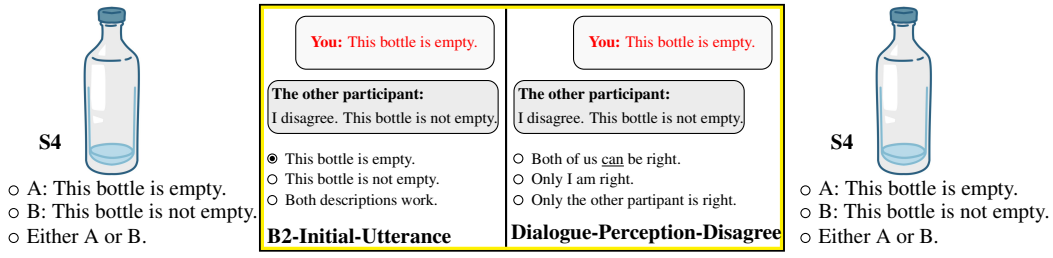
Experiment 2 mirrored Experiment 1 except for Block 2 (B2), see Figure 8. While the participants' task in B2 remained identical to that in Experiment 1, the bot was programmed to display a preference for *precision*, accepting the predication only at S5 and rejecting it at both S1 and S4 (yellow in Figure 4). Consequently, in critical S4 trials, the bot agreed with participants who chose a *Reject*-response, but disagreed with those who chose either an *Accept*- or *Either*-response. The visual stimuli, filler trials, and procedure were the same as in Experiment 1.

3.2 Participants

Experiment 2 followed the same recruitment criteria as Experiment 1. A total of 30 participants took part in the study. No participants were excluded for failing to meet the accuracy threshold on attention-check trials.

3.3 Predictions

We start with the predictions for the critical scale point S4. Consistent with Experiment 1, we expect no significant changes in the initial utterance choices in B2



(a) Block 1 (b) Block 2 (c) Block 3

Figure 8 Experiment 2 item example.

compared to the isolation baseline in B1. For agreement dialogues—which in this experiment occur when participants reject the imprecise predication—we hold the same expectations as in Experiment 1: we predict no updates to the SoP in B3 compared to the dialogue baseline B2, and we predict participants will predominantly select the option '*Both of us are right*' in the dialogue preception task.

Regarding disagreement dialogues, both H1 and H2 predict strengthening effects after a disagreement dialogue. We therefore expect a drop in proportions of *Accept*-responses in B3 compared to B2 towards either *Reject*- or *Either*-responses. Finally, H3, predicts no changes across blocks. Predictions for the dialogue perception task, as well as for S1 and S5, remain the same as in Experiment 1.

3.4 Results and Discussion

The results from the three experimental blocks of Experiment 2 are visualized in Figure 9. We employed the same binarized coding schemes and mixed-effects logistic regression models as in Experiment 1. Control S5 and filler S1 trials yielded parallel patterns to those observed in Experiment 1, consistent with our predictions. In the Control S5 trials, where *Accept*-responses were at near ceiling, no changes were observed across blocks (all p's > 0.05). Proportions of *Either*-responses were overall very low at S5 across the three blocks, though they increased significantly in B2 compared to B1 ($\hat{\beta} = 1.15$, $SE = 0.57$, $z = 2.02$, $p < 0.05$) and decreased significantly in B3 compared to B2 ($\hat{\beta} = -1.43$, $SE = 0.63$, $z = -2.26$, $p < 0.05$); rates remained stable between B1 and B3. *Reject*-responses, on the other hand, decreased significantly between B2 and B3 ($\hat{\beta} = -2.75$, $SE = 0.76$, $z = -3.61$, < 0.001) and between B1 and B3 ($\hat{\beta} = -2.03$, $SE = 0.70$, $z = -2.88$, $p < 0.01$). While the theoretical significance of these fluctuations remains unclear, all S5 patterns are consistent with the predicted dispreference for imprecision at this scale

point. Similarly, filler S1 trials exhibited stable *Reject*-response patterns across all blocks (all p's > 0.05). A late-stage increase in *Either*-responses was observed in B3 compared to B2 ($\hat{\beta} = 1.26$, $SE = 0.48$, $z = 2.63$, $p < 0.01$). Comparisons between B1 and B3 further indicated a decrease in *Accept*-responses ($\hat{\beta} = -1.11$, $SE = 0.39$, $z = -2.84$, $p < 0.01$) alongside a corresponding increase in *Either*-responses ($\hat{\beta} = 1.42$, $SE = 0.48$, $z = 2.97$, $p < 0.01$). Updates in the interpretation of S1 trials were overall stable, with the observed fluctuations suggesting a reduced tolerance for the imprecise interpretation following the dialogue phase.

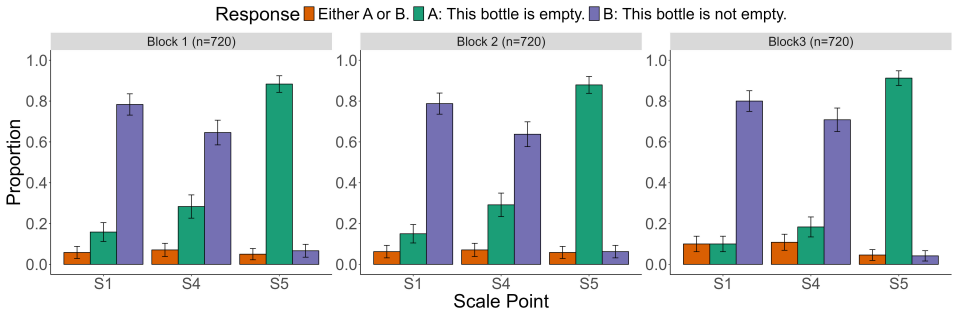
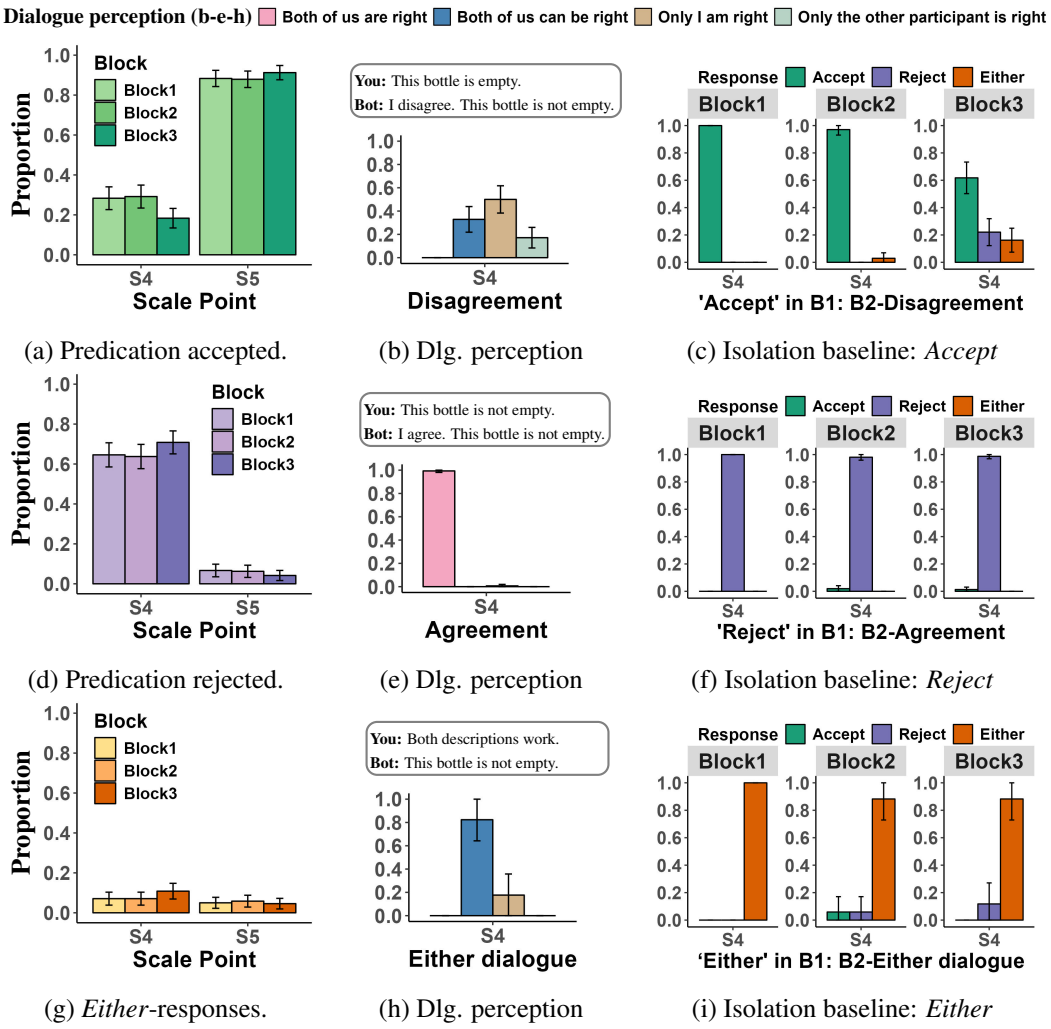


Figure 9 Experiment 2 participants' interpretational responses results.

Turning to the critical scale point S4, as in Experiment 1, the B1-B2 comparison showed no differences in the ACCEPT model ($\hat{\beta} = 0.03$, $SE = 0.28$, $z = 0.10$, $p > 0.1$) and the REJECT model ($\hat{\beta} = -0.04$, $SE = 0.27$, $z = -0.13$, $p > 0.1$), while a decrease was detected for *Either*-responses ($\hat{\beta} = -4.65$, $SE = 1.08$, $z = -4.30$, $p < 0.01$). As illustrated in Figure 10(g), this likely represents a statistical artifact driven by the near-zero frequency of *Either*-responses in B2 (a floor effect), rather than a substantive interpretational shift, as it did not transfer to either the ACCEPT or REJECT measures. In the B2-B3 comparison, we detected a significant drop in *Accept*-responses ($\hat{\beta} = -1.16$, $SE = 0.30$, $z = -3.86$, $p < 0.001$, Figure 10(a)), along with significant increases in the REJECT ($\hat{\beta} = 0.65$, $SE = 0.28$, $z = 2.35$, $p < 0.05$, Figure 10(d)) and EITHER models ($\hat{\beta} = 3.54$, $SE = 0.77$, $z = 4.62$, $p < 0.001$, Figure 10(g)). Similar to Experiment 1, this change was specifically driven by the bot's disagreeing replies to *Accept*-responses in B2. As illustrated in Figure 10(c), participants who chose an *Accept*-response in B2 significantly dispreferred this response in B3 in favor of *Reject*- and *Either*-responses, after encountering the bot's challenge to their lower SoP. The other two dialogue-types did not incur significant changes in B3 (Figure 10(f) and Figure 10(i)), mirroring Experiment 1. The B1-B3 robustness check confirmed these findings, revealing a significant drop in *Accept*-responses ($\hat{\beta} = -1.15$, $SE = 0.30$, $z = -3.81$, $p < 0.001$) and a significant increases in *Reject*- ($\hat{\beta} = 0.62$, $SE = 0.28$, $z = 2.21$, $p < 0.05$) and *Either*-responses

($\hat{\beta} = 1.64$, $SE = 0.46$, $z = 3.54$, $p < 0.001$).



(a) Predication accepted. (b) Dlg. perception (c) Isolation baseline: *Accept*

(d) Predication rejected. (e) Dlg. perception (f) Isolation baseline: *Reject*

(g) *Either*-responses. (h) Dlg. perception (i) Isolation baseline: *Either*

Figure 10 Experiment 2: (a, d, g): Binarized predication-judgment responses; (b, c, h): dialogue perception responses; (c, f, i) predication-judgment responses subsetting by response-type in Block 1.

Perception of agreement dialogues aligned with the results of Experiment 1, such that participants predominantly selected the response '*Both of us are right*', as shown in Figure 10(e). However, in disagreement dialogues (Figure 10(b))—where the bot challenged participants' lower SoPs—all three response options were selected at non-negligible rates. As in Experiment 1, shifts in the SoP were primarily driven by

participants who regarded the bot's position as plausible. Participants who retained their original lower SoP largely regarded themselves as uniquely correct (82.5% of non-updaters selected '*Only I am right*'). By contrast, participants who revised their SoPs were substantially more likely to endorse either the bot's stricter standard ('*Only the other participant is right*'; 40% of updaters) or the faultless interpretation (53.3% of updaters). Regarding the *Either*-dialogues, we again observed a high frequency of responses indicating that the disagreement was perceived as faultless (*Both of us can be right*'), alongside a smaller proportion of *Only I am right*' responses (see Figure 10(h)).

The significant decrease in *Accept*-responses, together with the corresponding increases in both *Reject*- and *Either*-responses in B3 relative to B2, indicates that metalinguistic disagreements can effectively raise the SoP, in line with prior claims (Klecha 2018; Lewis 1979). These findings therefore rule out H3, which predicted that the disagreements would not trigger interpretive updates, while remaining compatible with the predictions of both H1 (i.e., that metalinguistic denials only strengthen the SoP) and H2 (i.e., that they permit bidirectional modulation).

To better assess which hypothesis provides the best account of our findings, we considered the results of Experiment 2 alongside those of Experiment 1 and used Cohen's d to quantify the effect sizes associated with the discursive updates induced by metalinguistic denials—namely, the decrease in *Reject*-responses observed in Experiment 1 and the decrease in *Accept*-responses observed in Experiment 2. Both comparisons yielded effect sizes of approximately $d \approx 0.2$ (Experiment 1: $d = 0.20$, 95% CI: $[-0.38, -0.02]$; Experiment 2: $d = 0.26$, 95% CI: $[-0.44, -0.08]$). The comparable magnitude of these effects suggests that metalinguistic disagreements are equally effective in updating the standard in either direction, thereby challenging previous unidirectional accounts (Klecha 2018; Lewis 1979) and lending stronger support to H2 than to H1.

To further compare the perception data for disagreement dialogues across experiments we constructed a binary variables for each of the three response options: '*Only I am right*' (henceforth *Only-I*), '*Only the other participant is right*' (henceforth *Only-other*), and '*Both of us can be right*' (henceforth *Both-can*). These variables were appended and coded based on 1) whether the observation belonged to Experiment 1 or 2, which we refer to as EXPERIMENT; and 2) which interlocutor participants perceived to be right (i.e., *Only-I*, *Only-other*, *Both-can*), referred to as RIGHT. Response proportions by item, scale-point and experiment were obtained within each RIGHT level. We fitted a mixed-effects regression model to predict this new dependent variable from the fixed effects of EXPERIMENT, RIGHT, and their interaction. The model further included by-item random intercepts and fully-specified random slopes. Results revealed a significant interaction effect between EXPERIMENT and RIGHT for the contrast between *Only-I* and *Both-can* ($\hat{\beta} = -1.42$,

$SE = 0.06,\ t = -23.36,\ p < 0.001$). This interaction was primarily driven by the significantly higher proportion of participants selecting *Only-I*-responses in Experiment 1 compared to Experiment 2. In contrast, no corresponding interaction emerged for the comparison between *Only-other* and *Both-can* ($\hat{\beta} = -0.02$, $SE = 0.06$, $t = -0.28$, $p > 0.1$), indicating that participants across both experiments acknowledged the interlocutor's challenging SoP at comparable rates. This statistical asymmetry in the dialogue perception data mirrors the patterns observed in the block comparison results. In Experiment 1, precise participants who updated their SoP did so primarily by shifting to *Either*-responses rather than categorically adopting the bot's lower standard. In Experiment 2, however, imprecise participants more readily abandoned their baseline opinion, with the drop in *Accept*-responses translating more directly into higher *Reject*-responses. We speculate that these qualitative differences might underlie previous intuitions in the literature about the unidirectionality of SoP updates (Klecha 2018; Lewis 1979).

4 General Discussion

The present study investigated whether metalinguistic disagreements can bidirectionally update the standard of precision (SoP), addressing prior claims that such disagreements can successfully raise the SoP but fail to lower it (Klecha 2018; Lewis 1979). To evaluate this asymmetry claim, we conducted two experiments examining both strengthening and weakening updates in first-person dialogue contexts involving precise and imprecise interlocutors. More specifically, we asked whether participants would adjust their own interpretive standards in response to interlocutors who either challenged or endorsed imprecise utterances. Our goals were twofold: first, to determine whether metalinguistic disagreements can modulate the SoP in both directions, and second, to characterize the qualitative nature of these updates and the extent to which they align with participants' perceptions of disagreement.

We considered three hypotheses. Our first hypothesis (H1) proposed that metalinguistic disagreements can only induce upward updates of the SoP while failing to weaken it (Klecha 2018; Lewis 1979). The findings from Experiment 1, where *Reject*-responses significantly decreased following disagreement dialogues, indicate that participants successfully lowered the SoP, thereby directly challenging this unidirectional account. Our findings likewise argue against the third hypothesis (H3), which predicted that first-person metalinguistic disagreements, like bystander disagreements (Wu & Aparicio 2025a), would fail to trigger interpretive updates altogether. Overall, our results are most consistent with the second hypothesis (H2), according to which metalinguistic disagreements can update the SoP bidirectionally by both weakening and strengthening interpretive standards. In Experiment 1, disagreement dialogues produced a significant decrease in *Reject*-responses, re-

flecting a successful weakening of the SoP. Conversely, Experiment 2 revealed a corresponding decrease in *Accept*-responses, indicating a successful strengthening of the SoP. Importantly, the effect size of these updates was comparable across the two experiments, as shown by their similar Cohen's d value. Taken together, these results suggest that interlocutors accommodate shifts in the SoP in both directions in response to their interlocutors' linguistic behavior.

Some evidence of directional effects nevertheless emerged. While imprecise participants in Experiment 2 readily abandoned their lower SoP in favor of a stricter SoP (i.e., switching from *Accept*- to *Reject*-responses), precise participants were less willing to relinquish their own SoP and instead accommodated the bot's challenge by shifting toward the *Either*-response. This pattern was also reflected in the dialogue perception task, where precise participants in Experiment 1 selected '*Only I am right*' at significantly higher rates than imprecise participants in Experiment 2. The question remains how to capture bidirectional updates while preserving the observed qualitative asymmetry. In ongoing work (Wu, Grove & Aparicio In prep), we tentatively argue that the observed experimental results can be derived from the underlying scale structure of maximum-standard adjectives (Kennedy 2007; Kennedy & McNally 2005) and interlocutors' probabilistic reasoning about likely SoPs. We provide a proof of concept by implementing this account within the Probabilistic Dynamic Semantics (PDS) framework (Grove & White 2025a,b, 2026).

5 Conclusion

In two experiments, we investigate how interlocutors coordinate on contextual standards of precision (SoP) via metalinguistic disagreement. Our results indicate that disagreements can both raise *and* lower the SoP, challenging previous claims that such implicit negotiations can only effectively raise the standard (Klecha 2018; Lewis 1979). Importantly, these updates were qualitatively asymmetric: precise participants incorporated the challenger's perspective while retaining their own baseline commitments, whereas imprecise participants were more likely to shift away from their initial stance toward the alternative SoP. Participants' perception of the disagreements exhibited parallel asymmetries, with precise participants more often endorsing judgments consistent with exclusive self-assessment of correctness than imprecise participants. Finally, our results suggest that first-person metalinguistic disagreements more effectively elicit participants' intuitions about the discourse dynamics of imprecision than the bystander disagreements used in previous work (Wu & Aparicio 2025a).

References

Aparicio, Helena, Ming Xiang & Christopher Kennedy. 2015. Processing gradable adjectives in context: A visual world study. *Semantics and Linguistic Theory (SALT)* 25. 413–432. doi:10.3765/salt.v25i0.3128.

Aparicio Terrasa, Helena. 2017. *Processing context-sensitive expressions: The case of gradable adjectives and numerals*: The University of Chicago PhD dissertation.

Barker, Chris. 2013. Negotiating taste. *Inquiry* 56(2-3). 240–257. doi:10.1080/0020174X.2013.784482.

Beltrama, Andrea & Florian Schwarz. 2021. Imprecision, personae, and pragmatic reasoning. *Semantics and Linguistic Theory (SALT)* 31. 122–144. doi:10.3765/salt.v31i0.5107.

Beltrama, Andrea & Florian Schwarz. 2022. Social identity, precision and charity: when less precise speakers are held to stricter standard. *Semantics and Linguistic Theory (SALT)* 32. 575–598. doi:10.3765/salt.v1i0.5406.

Beltrama, Andrea & Florian Schwarz. 2024. Social identity affects imprecision resolution across different tasks. *Semantics and Pragmatics* 17. 10–EA. doi:10.3765/sp.17.10.

Beltrama, Andrea, Stephanie Solt & Heather Burnett. 2023. Context, precision, and social perception: A sociopragmatic study. *Language in Society* 52(5). 805–835. doi:10.1017/S0047404522000240.

Burnett, Heather. 2014. A delineation solution to the puzzles of absolute adjectives. *Linguistics and Philosophy* 37. 1–39. doi:10.1007/s10988-014-9145-9.

Grove, Julian & Aaron Steve White. 2025a. Modeling the prompt in inference judgment tasks. *Proceedings of Experiments in Linguistic Meaning* 3. 176–187. doi:10.3765/elm.3.5857.

Grove, Julian & Aaron Steve White. 2025b. Probabilistic dynamic semantics https://lingbuzz.net/lingbuzz/008478/.

Grove, Julian & Aaron Steve White. 2026. Factivity, presupposition projection, and the role of discrete knowledge in gradient inference judgments. *Natural Language Semantics* 34. 1–45. doi:10.1007/s11050-025-09244-9.

Heim, Stefan, Natalja Peiseler & Natalia Bekemeier. 2020. Few or many? an adaptation level theory account for flexibility in quantifier processing. *Frontiers in Psychology* 11. 382.

Kaiser, Elsi & Deniz Rudin. 2020. When faultless disagreement is not so faultless: What widely-held opinions can tell us about subjective adjectives. *Proceedings of the Linguistic Society of America* 5(1). 698–707. doi:10.3765/plsa.v5i1.4757.

Kaiser, Elsi & Deniz Rudin. 2021. Arguing with experts: Subjective disagreements on matters of taste. *Proceedings of the Annual Meeting of the Cognitive Science*

Society 43(43). 924–930. https://escholarship.org/uc/item/8921n58s.

Kao, Justine T., Jean Y. Wu, Leon Bergen & Noah D. Goodman. 2014. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences* 111(33). 12002–12007. doi:10.1073/pnas.1407479111.

Kennedy, Christopher. 2007. Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy* 30. 1–45. doi:10.1007/s10988-006-9008-0.

Kennedy, Christopher. 2013. Two sources of subjectivity: Qualitative assessment and dimensional uncertainty. *Inquiry* 56(2–3). 258–277. doi:10.1080/0020174X.2013.784483.

Kennedy, Christopher & Louise McNally. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language* 81(2). 345–381. doi:10.1353/lan.2005.0071.

Klecha, Peter. 2018. On unidirectionality in precisification. *Linguistics and Philosophy* 41. 87–124. doi:10.1007/s10988-017-9216-9.

Kölbel, Max. 2004. Faultless disagreement. *Proceedings of the Aristotelian Society* 104(1). 53–73. doi:10.1111/j.0066-7373.2004.00081.x.

Krifka, Manfred. 2002. Be brief and vague! and how bidirectional optimality theory allows for verbosity and precision. In David Restle & Dietmar Zaefferer (eds.), *Sounds and Systems: Studies in Structure and Change. A Festschrift for Theo Vennemann*, 439–458. Berlin, New York: De Gruyter Mouton. doi:10.1515/9783110894653.439.

Krifka, Manfred. 2007. *Approximate interpretation of number words*. Berlin: Humboldt-Universität zu Berlin, Philosophische Fakultät II. doi:10.18452/9508.

Lasersohn, Peter. 1999. Pragmatic halos. *Language* 75(3). 522–551. doi:10.2307/417059.

Lauer, Sven. 2012. On the pragmatics of pragmatic slack. *Proceedings of Sinn und Bedeutung* 16(2). 389–402.

Lauer, Sven. 2013. *Towards a dynamic pragmatics*: Stanford University Doctoral dissertation.

Leffel, Timothy, Ming Xiang & Christopher Kennedy. 2016. Imprecision is pragmatic: Evidence from referential processing. *Semantics and Linguistic Theory (SALT)* 26. 836–854. doi:10.3765/salt.v26i0.3937.

Lewis, David. 1979. Scorekeeping in a language game. *Journal of Philosophical Logic* 8. 339–359. doi:10.1007/BF00258436.

Mathis, Ariel & Anna Papafragou. 2022. Agents' goals affect construal of event endpoints. *Journal of Memory and Language* 127. 104373. doi:10.1016/j.jml.2022.104373.

Ronderos, Camilo R., Ira Noveck & Ingrid Lossium Falkum. 2024. Straight enough: Deriving imprecise interpretations of maximum standard absolute adjectives.

Glossa Psycholinguistics 3(1). 1–36. doi:10.5070/G60111411.

Solt, Stephanie. 2015. Vagueness and imprecision: Empirical foundations. *Annu. Rev. Linguist.* 1(1). 107–127. doi:10.1146/annurev-linguist-030514-125150.

Syrett, Kristen, Christopher Kennedy & Jeffrey Lidz. 2010. Meaning and context in children's understanding of gradable adjectives. *Journal of Semantics* 27(1). 1–35. doi:10.1093/jos/ffp011.

van der Henst, Jean-Baptiste, Laure Carles & Dan Sperber. 2002. Truthfulness and relevance in telling the time. *Mind & Language* 17(5). 457–466. doi:10.1111/1468-0017.00207.

Wu, Yifan & Helena Aparicio. 2025a. Disagreements do not automatically raise the standard of precision. *Proceedings of Experiments in Linguistic Meaning* 3. 435–446. doi:10.3765/elm.3.5835.

Wu, Yifan & Helena Aparicio. 2025b. Meaning adaptation in the discourse dynamics of imprecision. *Proceedings of the Annual Meeting of the Cognitive Science Society* 47(0). 102–108. https://escholarship.org/uc/item/6r9914v5.

Wu, Yifan, Julian Grove & Helena Aparicio. In prep. Predicting the discourse dynamics of imprecision from adjectival scale structure.

Zehr, Jeremy & Florian Schwarz. 2018. Penncontroller for internet based experiments (ibex) doi:10.17605/OSF.IO/MD832.

Yifan Wu
203 Morrill Hall
Cornell University
Ithaca, NY 14850
yw2578@cornell.edu

Helena Aparicio
203 Morrill Hall
Cornell University
Ithaca, NY 14850
haparicio@cornell.edu