# Disagreements do not automatically raise the standard of precision

Yifan Wu & Helena Aparicio*

**Abstract.** Speakers often choose to utter imprecise sentences that, albeit felicitous, are, strictly speaking, false (e.g., using '*This bottle is empty*' to describe a bottle with a bit of water in it). The acceptability of an imprecise utterance hinges on the standard of precision (SoP), a discourse parameter that governs how much imprecision is tolerated in a context. Previous theoretical accounts (e.g., Lewis 1979, Klecha 2018) have argued that metalinguistic denials that target the assertability of an imprecise utterance (e.g., '*No, this bottle is not empty!*') more or less force accommodation to a higher SoP. The present study investigates the nature of this accommodation process. In particular, we ask whether metalinguistic disagreements result in an automatic update of the SoP. In two acceptability judgment experiments, we show that imprecise utterances are not deemed unacceptable when embedded in a disagreement dialogue. Our findings instead suggest that metalinguistic denials act as a request to raise the SoP and that any potential updates ought to be signaled overtly in subsequent conversational moves.

**Keywords.** Imprecision; standard of precision; metalinguistic disagreement; maximum standard absolute adjectives; discourse processing

**1. Introduction.** During conversation, speakers often choose to stretch the boundaries of lexical representations by speaking loosely. For instance, in Figure 1, Alex describes the bottle as *empty*, despite being aware that it contains a small amount of water. Such an instance showcases the phenomenon of imprecision, or loose talk (Lewis 1979, Lasersohn 1999, Krifka 2002, 2007, Kennedy 2007, Syrett et al. 2010, Lauer 2012, Aparicio et al. 2015, Leffel et al. 2016, Aparicio Terrasa 2017, Klecha 2018, Ronderos et al. 2024; a.o.), wherein a speaker chooses to utter a sentence that they judge to be pragmatically felicitous, despite it being strictly speaking false.

*Alex:* This bottle is empty.

Figure 1: Imprecise description of a bottle containing some water.

Imprecision is pervasive in everyday communication, manifesting across a wide range of lexical items, such as maximum standard absolute adjectives (e.g., *empty*), verb phrases denoting events with incremental themes (e.g., *peel the apple*), or round numerals (e.g., *one hundred*), among others. Predicates that can be subject to imprecision usually denote upper closed scalar meanings. Here, we exemplify this core property of imprecision through the case of maximum standard absolute adjectives, which the current work uses as a testbed. Maximum standard absolute adjectives have been argued to associate with adjectival scales (i.e., the scale encoding the dimension denoted by the adjective) that have an upper bound corresponding to the maximum degree on the relevant scale (e.g., the

[*] We are grateful to the Cornell LiMe Lab, Cornell Interdisciplinary Semantics Group, Cornell C. Psyd, and the audiences at HSP2024 and ELM3 for discussions. Authors: Yifan Wu, Cornell University (yw2578@cornell.edu) & Helena Aparicio, Cornell University (haparicio@cornell.edu).

maximum degree of emptiness; Rotstein & Winter (2004), Kennedy & McNally (2005), Kennedy (2007)), see Figure 2.[1]

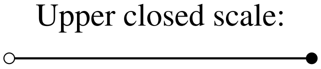


Figure 2: Scale topology for maximum standard absolute adjectives.

A precise interpretation of a maximum standard absolute adjective obtains when the adjective is predicated of an individual that instantiates the property to its maximum degree (e.g., a bottle that is *completely* empty). Imprecise interpretations, on the other hand, obtain when the predicate is applied to an individual that bears the adjectival property to a non-maximal degree, as exemplified in Figure 1. The felicity of an imprecise utterance has been shown to be contingent on features of the discourse context (van Der Henst et al. 2002, Burnett 2014, Aparicio Terrasa 2017, Ronderos et al. 2024). For instance, a movie theater with only three people sitting in it can be felicitously described as *empty* in the context of a highly anticipated movie premiere that was expected to be a blockbuster. In this case, the imprecise use of the predicate is licensed: three people is a small enough number for it to be safely ignored. However, in a higher-stakes context such as an emergency fire evacuation, the same imprecise interpretation of the predicate stops being available, since ignoring the three people sitting in the theater could have fatal consequences.[2] Additionally, recent studies have shown that comprehenders take into account information about the speaker's goals (Mathis & Papafragou 2022), as well as social information about the speaker's identity (Beltrama & Schwarz 2021, 2022, 2024) to determine whether to adopt an imprecise interpretation of the utterance.

Here, we take the acceptability of an imprecise utterance to be determined by the *Standard of Precision* (SoP, Lewis 1979), a latent discourse parameter—or in Lewis' terms, an element of the conversational score—that governs the degree of imprecision tolerated in a given discourse. Because the SoP is not observable, conversational agents must rely on contextual cues (some of which have already been mentioned above) to infer likely parametrizations of the SoP. While in most cases speakers and listeners successfully align on the value of the SoP, this implicit coordination process is not infallible; occasionally, interlocutors assume conflicting parametrizations that can eventually cause conversational disruptions. Of particular interest to us are instances where the speaker asserts an utterance whose felicity is contingent on a sufficiently low SoP, whereas the listener has all along been assuming a stricter one. In such situations, the listener may choose to go along with the speaker's conversational move by updating their beliefs about the SoP to a sufficiently low value so as to render the speaker's utterance felicitous. Alternatively, the listener might be unwilling to accommodate. In such cases, the only discourse move available to the listener is to object to the assertability of the utterance through a metalinguistic denial or disagreement (Horn 1989, Barker 2002, 2013), as shown in (1). The disagreement in (1) is therefore not about the factual state of the world, both Alex and Andy acknowledge that the bottle contains some water.

[1] While maximum adjectives minimally have upper bounds, adjectives like *empty* actually associate with fully closed scales, i.e., scales that are closed on both the upper and the lower end. Antonymous pairs of gradable adjectives map their arguments onto the same scale but impose inverse orderings on their shared domains. For the antonym pair *full/empty*, both maximum standard absolute adjectives, the maximum of *full* is *empty*'s minimum and vice versa.

[2] This example is based on an example provided by Burnett (2014).

Rather, Andy is taking issue with the appropriateness of the predicate *empty* as a suitable descriptor of the referent. Put differently, Andy is indirectly challenging the low SoP assumed by Alex and signaling that a higher SoP should be adopted.[3]

(1) a. *Alex:* This bottle is empty. [Uttered as a description of the bottle in Figure 1]
b. *Andy:* No, this bottle is not empty; there's a bit of water in it!

Previous authors have observed that metalinguistic denials such as (1-b) are *hard to resist* (Klecha 2018: 92), more or less forcing the listener, in our running example *Alex*, to accommodate to a higher SoP (Lewis 1979, Lauer 2012, Klecha 2018). Here we investigate whether this need to accommodate is a by-product of the discursive import of metalinguistic denials. More specifically, we ask whether metalinguistic denials such as (1-b) lead to an automatic update of the SoP. We consider two hypotheses. Hypothesis 1 states that challenging the SoP through a metalinguistic denial automatically updates this discourse parameter, thereby superseding previous parametrizations. Hypothesis 1 makes the prediction that disagreements should decrease the acceptability of a previous imprecise utterance, since after the challenge only the higher SoP should be operative. Contra Hypothesis 1, Hypothesis 2 posits that metalinguistic denials act as a request to raise the SoP, but do not directly update it. This hypothesis therefore predicts that the looser SoP can remain operative after the challenge and that the acceptability of the original imprecise utterance should not decrease. We report results from two acceptability judgment studies, where we find evidence for Hypothesis 2. Our results indicate that imprecise utterances are not deemed unacceptable in disagreement dialogues, suggesting that the lower SoP remains operative even after being challenged. This implies that any potential updates to the SoP ought to occur in subsequent conversational moves.

The remainder of this paper proceeds as follows. Section 2 presents the results pertaining to Experiment 1. In Section 3, we present Experiment 2 and the comparison analysis between the two experiments. Finally, Section 4 provides a general discussion of our findings and Section 5 concludes the paper.

## 2. Experiment 1.

The goal of Experiment 1 was to obtain interpretational preferences for imprecise utterances in isolation. Results pertaining to Experiment 1 were later used as a baseline for comparison with results from Experiment 2, in which the same visual stimuli were paired with disagreement dialogues (see Section 3).

### 2.1. Materials & Design.

We constructed 24 five-point scales instantiating different maximum standard absolute properties to varying degrees (see Figure 3). Individual scale-points were combined with a written statement of the form '*This* [OBJECT] *is* [ADJECTIVE]' (e.g., *This bottle is empty*), where the noun was always an appropriate descriptor of the depicted object and the adjective matched the property represented in the scale the picture was part of (see left panel of Figure 5). The five scale-points pertaining to each of the 24 scales were distributed across five lists following a Latin-square design. This ensured that each participant judged one single scale-point

[3] Metalinguistic disagreements over imprecise utterances have been argued to be unidirectional (e.g., Lewis 1979, Klecha 2018), meaning that such implicit challenges can be used to raise the SoP, but not to lower it. In Klecha's terms, the SoP can be raised *incidentally to the content of an expression*. However, in order to lower the SoP, Klecha argues that speakers must engage in an explicit metalinguistic negotiation (Klecha 2018: 93).

per item. Trial order was randomized within each list and for each participant.

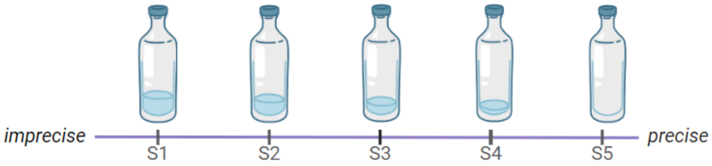


Figure 3: Five scale-points corresponding to the scale ‘*empty bottle*’.

2.1.1. NORMING STUDY. The 24 scales tested in Experiment 1 were normed in order to ensure that the lower scale points (S1-S4) tolerated some non-negligible amount of imprecision. In each trial, the full five-point scale was paired with a statement of the form ‘*This* [OBJECT] *is* [ADJECTIVE]’. As in Experiment 1, the noun always matched the depicted object and the adjective encoded the property represented in the visual scale. The study consisted of a forced-choice picture-matching task in which participants were instructed to assess whether the statement was an appropriate description of each individual scale-point. Unlike in Experiment 1, the full scale was available to participants at the moment of providing their judgements. Participants gave their answers by choosing one of three possible responses (i.e., ‘*yes*’, ‘*no*’ or ‘*unsure*’) for each of the five points in the scale (see left panel of Figure 4).

Thirty adult native speakers of American English recruited through the crowd-sourcing platform *Prolific* participated in the study. Participants were compensated at a rate of $15 per hour. Results are shown in the right panel of Figure 4. As can be observed in the plot, precise interpretations were preferred over imprecise ones. This is shown by the fact that the precise scale-point (S5) received the highest amount of ‘*yes*’ responses (green bars). More important for us, participants demonstrated some degree of tolerance for imprecise interpretations in all the lower scale-points, as indexed by the ‘*yes*’ responses in S1-4. The acceptability of imprecise interpretations, however, was gradient: the further the scale point strayed from the endpoint-oriented interpretation, the less available imprecise interpretations became.

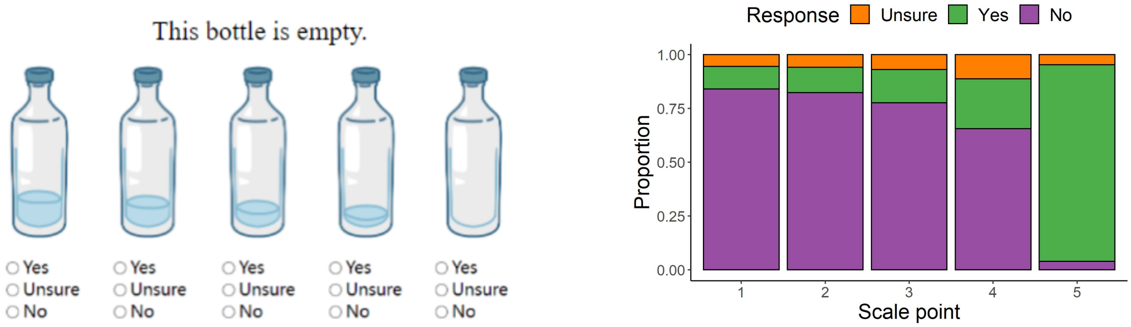


Figure 4: Left: Norming study item example; Right: Norming study results.

2.2. PROCEDURE. The experiment was administered remotely through the *PCIbex Farm* platform (Zehr & Schwarz 2018). At the beginning of the 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. In the main part of the experiment, participants saw a visual stimulus accompanied by a sentence (e.g., *This bottle is empty*, see left panel of Figure 5). The experimental task consisted of a forced-choice picture-matching task, where participants were asked to judge whether the written statement was an appropriate description of the picture. Participants provided their answer by choosing one of three possible options: '*Yes*', '*No*', and '*Unsure*'.

2.3. PARTICIPANTS. Thirty participants were recruited through the web platform *Prolific* and were compensated at a rate of $15/hour. All participants were native speakers of American English and were at least 18 years old.

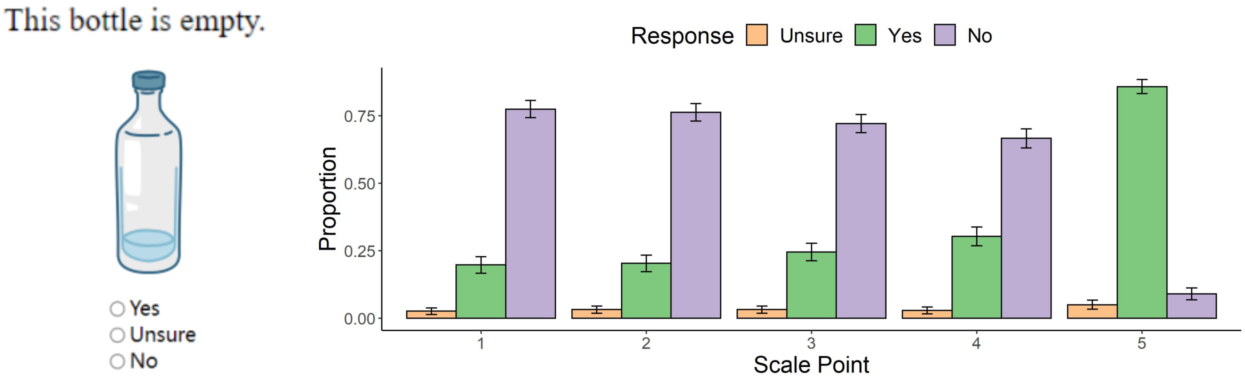


Figure 5: Left: Experiment 1 item example; Right: Experiment 1 results.

2.4. RESULTS & DISCUSSION. Results are shown in Figure 5. Experiment 1 qualitatively replicates the results obtained in the norming study. First, participants continue to display a strong preference for precision over imprecision. To confirm this statistically, we fit a logistic mixed-effects regression model to the binarized response variable (i.e., *Yes*-responses were coded as 1, and *No*- and *Unsure*-responses were coded as 0), using SCALE POINT as a fixed effect with S5 as the reference level. The model also included random intercepts by-item and by-participant, as well as by-condition random slopes. Model outputs, shown in Table 1a, reveal significant effects for all comparisons (all $p$'s $< 0.001$), confirming that the precise scale-point (S5) received a significantly higher proportion of *Yes*-responses when compared to those scale-points that only supported an imprecise interpretation.

Second, all lower scale-points (S1-S4) exhibit some tolerance for imprecision (see the green bars corresponding to *Yes*-responses in the right panel of Figure 5). As in the norming study, tolerance for imprecision follows a gradient: the closer the scale-point is to the maximum scalar degree, the more available imprecise interpretations become. Among the scale-points that were incompatible with a precise interpretation (i.e., S1-S4), S1 showed the lowest proportion of *Yes*-responses ($20\%$), whereas S4 showed the highest ($30\%$). To further examine this gradient effect, we recoded our categorical predictor SCALE POINT using a forward-difference contrast scheme (a coding system that compares the mean of the dependent variable for one level of a categorical variable to the mean of the next level). None of the comparisons reached significance (see Table 1b). Finally, we note that participants displayed very little uncertainty regarding the appropriateness of the predicates, as reflected by the low proportion of *Unsure*-responses (all scale-points displayed proportions lower than $5\%$).

Table 1: Model outputs for Experiment 1.

(a) Model 1: Preference for precision analysis.

| | $\beta$ | SE | $z$ | $p$ |
|---|---|---|---|---|
| (Intercept) | 2.81 | 0.44 | 6.44 | 1.18e-10*** |
| S5 vs. S1 | -5.85 | 0.86 | -6.84 | 7.81e-12*** |
| S5 vs. S2 | -6.00 | 0.90 | -6.67 | 2.56e-11*** |
| S5 vs. S3 | -5.53 | 0.88 | -6.28 | 3.47e-10*** |
| S5 vs. S4 | -5.11 | 0.92 | -5.56 | 2.70e-08*** |

(b) Model 2: Gradience analysis.

| | $\beta$ | SE | $z$ | $p$ |
|---|---|---|---|---|
| (Intercept) | -2.99 | 0.64 | -4.69 | 2.68e-06*** |
| S1-S2 | -0.19 | 0.36 | -0.53 | 0.5934 |
| S2-S3 | 0.27 | 0.35 | 0.78 | 0.4368 |
| S3-S4 | 0.71 | 0.38 | 1.86 | 0.0629. |

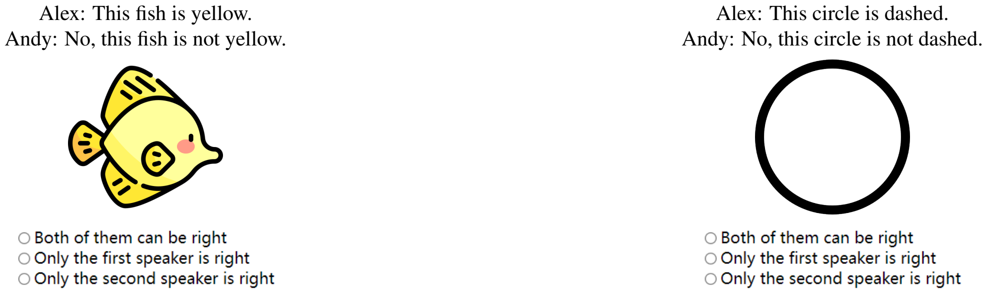
significance levels: <.001*** <.01** <.05* <.1

Taken together, Experiment 1 results show that judgments elicited by presenting the scale-points in isolation are qualitatively comparable to the norming study results, where the same items were judged as part of the five-point scale: overall, precision was preferred over imprecision, but all the lower scale-points allowed for some degree of imprecision, especially scale-points that were closer to the endpoint. More importantly, results from Experiment 1 provide us with a baseline for comparison with results pertaining to Experiment 2, to be presented in the next section.

**3. Experiment 2.** The goal of Experiment 2 was to determine whether the acceptability of an imprecise utterance declines when judged as part of a disagreement dialogue.

3.1. MATERIALS & DESIGN. Experiment 2 followed the same design and tested the same stimuli as Experiment 1, with one crucial modification: the visual stimuli were paired with a disagreement dialogue. All the disagreements consisted of a speaker assertion of the form '*This* [OBJECT] *is* [ADJECTIVE]' (e.g., '*This bottle is empty*') followed by a metalinguistic denial of the form '*No, this* [OBJECT] *is not* [ADJECTIVE]' (e.g., '*No, this bottle is not empty*', see left panel of Figure 7).

Additionally, twenty-four fillers were included. In the filler trials, participants were presented with a visual stimulus coupled with a disagreement dialogue. However, unlike experimental trials, the predicates under discussion were not maximum standard adjectives but rather color adjectives (e.g., *yellow*) or minimum standard gradable adjectives (e.g., *dashed*). In half of the filler trials, the image clearly matched both the adjectival property and the noun included in the first speaker's utterance, whereas in the other half of the filler trials the visual stimulus could not be described with the adjectival property under discussion, only with the noun (see Figure 6). Experimental materials corresponding to the critical trials were distributed in five lists following a Latin square design. Each list was complemented with the 24 filler trials. Each participant was randomly assigned to a list and the order of the 48 trials within each list were randomized for each participant.


Alex: This fish is yellow.
Andy: No, this fish is not yellow.

○ Both of them can be right
○ Only the first speaker is right
○ Only the second speaker is right

Alex: This circle is dashed.
Andy: No, this circle is not dashed.

○ Both of them can be right
○ Only the first speaker is right
○ Only the second speaker is right

Figure 6: Left: Color filler example; Right: Minimum standard absolute adjective filler example.

3.2. PROCEDURE. Experiment 2 followed the same procedure used in Experiment 1 with one important distinction. In Experiment 2, participants' task was to judge whether only one of the speakers was right, or whether both of them could be right by selecting one of three options: '*Only the first speaker is right*', '*Only the second speaker is right*', or '*Both of them can be right*' (henceforth, '*First*', '*Second*' and '*Both*'; see left panel of Figure 7). In scale-points S1-4, which are only compatible with an imprecise interpretation of the first speaker's utterance, a *First*-response indicates that the imprecise utterance was judged to be acceptable, even after having been targeted by a metalinguistic denial. *First*-responses therefore entail that the lower SoP remains operative, even after the second speaker takes issue with the assertability of the imprecise utterance. In contrast, *Second*-responses indicate that the imprecise utterance was deemed unacceptable. This response therefore indexes alignment with the higher SoP adopted by the second speaker. Finally, a *Both*-response indicates that the disagreement was judged to be *faultless* (Kölbel 2004, Barker 2013, Kennedy 2013, Kaiser & Rudin 2020, 2021, Pecsok & Aparicio 2024). Faultless disagreements are a type of disagreement where neither party in the discourse is judged to be at fault. We take this type of answer to be compatible with a lower SoP.

3.3. PARTICIPANTS. Participants consisted of 60 native speakers of American English who were at least 18 years old. All participants were recruited through the web platform *Prolific*. Participation was compensated at a rate of $15 per hour. Two participants were removed from data analysis due to failure to reach a 90% accuracy threshold in the filler trials.

3.4. PREDICTIONS. As discussed in Section 1, our first hypothesis (H1) states that metalinguistic denials automatically raise the SoP. H1 therefore predicts that the acceptability of the imprecise utterance should decrease in Experiment 2 compared to Experiment 1, where the same utterance is judged in isolation. More specifically, in Experiment 2, S1-4 should display a substantial increase in *Second*-responses compared to the proportion of *No*-responses observed in Experiment 1. Crucially the increase in *Second*-responses should be accompanied by a decrease in *First*-responses, such that the proportion of *First*-responses in Experiment 2 should be lower compared to the proportion of *Yes*-responses in Experiment 1. No differences are predicted for S5.

Our second hypothesis (H2) posits that metalinguistic challenges only act as a request to update the SoP. H2 therefore predicts that the lower SoP should remain operative after a metalinguistic denial, and that the acceptability of the imprecise utterance should not be negatively impacted by the disagreement. Under this hypothesis, we expect comparable acceptability rates for imprecise utterances across Experiments 1 and 2. This pattern should materialize as comparable rates of *Yes*/*First*-responses and *No*/*Second*-responses respectively across Experiments 1 and 2. Alternatively, if participants select *Both*-responses at a high rate in Experiment 2, we would expect this choice to be at the expense of *Second*-responses. This should result in lower selection rates of *Second*-responses in Experiment 2, compared to *No*-responses in Experiment 1.

3.5. RESULTS & DISCUSSION. Results for Experiment 2 are presented in the right panel of Figure 7. We note that the preference for precision remains evident in Experiment 2, as shown by the high proportion of *First*-responses (green bars) in S5 compared to all other scale-points. To statistically confirm this preference, we fit a mixed effects logistic regression model to the binarized response data (i.e., *First*-responses were coded as 1 and *Both*- and *Second*-responses as 0). This dependent variable was predicted from SCALE POINT, with S5 as the reference level. The model also included by-item and by-participant random intercepts, as well as random slopes by scale-point for both participants and items. Model outputs revealed significant differences in all comparisons (see Table 2a).

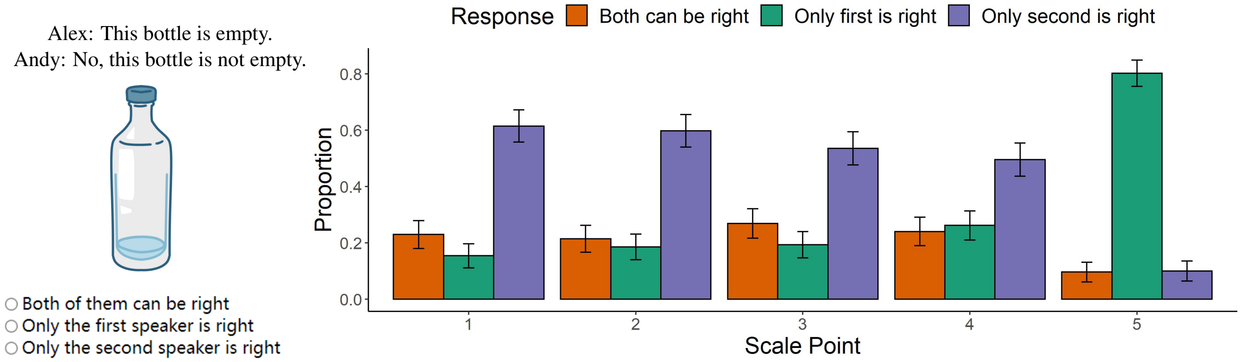


Figure 7: Left: Experiment 2 item example; Right: Experiment 2 results.

Results also reveal that participants selected *Both* at a higher rate in S1-S4 compared to S5. To statistically assess this qualitative pattern, we binarized our response data, such that *Both*-responses were coded as 1, while the remaining two options (*First* and *Second*) were coded as 0. A logistic mixed effects regression model was fit to this new binomial variable using SCALE POINT as a fixed effect. S5 was coded as the reference level. Random intercepts by items and participants were also included. No random slopes were added to the model due to convergence issues. As shown in Table 2b, all comparisons reached significance.

Table 2: Model outputs for Experiment 2.

(a) Model 1: Preference for precision analysis.

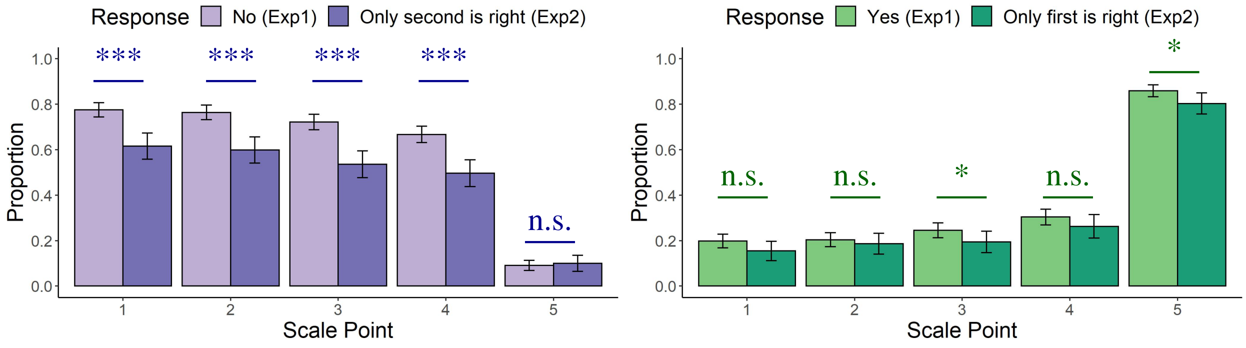
|  | $\beta$ | SE | $z$ | $p$ |
|---|---|---|---|---|
| (Intercept) | 1.48 | 0.19 | 7.83 | 5.00e-15*** |
| S5 vs. S1 | -5.56 | 1.03 | -5.41 | 6.24e-08*** |
| S5 vs. S2 | -5.99 | 1.32 | -4.55 | 5.28e-06*** |
| S5 vs. S3 | -5.22 | 1.00 | -5.22 | 1.77e-07*** |
| S5 vs. S4 | -3.50 | 0.61 | -5.79 | 7.18e-09*** |

(b) Model 2: *Both*-responses analysis.

|  | $\beta$ | SE | $z$ | $p$ |
|---|---|---|---|---|
| (Intercept) | -3.10 | 0.38 | -8.21 | 2.26e-16*** |
| S5 vs. S1 | 1.35 | 0.28 | 4.89 | 1.01e-06*** |
| S5 vs. S2 | 1.19 | 0.28 | 4.28 | 1.90e-05*** |
| S5 vs. S3 | 1.58 | 0.28 | 5.74 | 9.24e-09*** |
| S5 vs. S4 | 1.36 | 0.28 | 4.93 | 8.27e-07*** |

In order to determine whether the predictions of our two hypotheses (see Section 3.4) are borne out, we conducted comparison analyses between Experiment 1 and Experiment 2. Specifically, we compared *Yes*-responses in Experiment 1 to *First*-responses in Experiment 2, as these responses indicate that participants took the lower standard to be operative. Additionally, we compared *No*-responses in Experiment 1 to *Second*-responses in Experiment 2, as these responses indicate that participants rejected the imprecise utterances in S1-S4, and therefore did not take the lower standard to be operative. We did not perform a comparison between the *Unsure*- and *Both*-responses in Experiments 1 and 2 respectively, as these responses are not directly comparable.

Four new binary variables were constructed. For Experiment 2, *First*-responses were coded as 1 and *Second*- and *Both*-responses were coded as 0. This variable reflected the acceptability of the imprecise utterance in S1-S4. A second variable in which *Second*-responses were coded as 1 and all other options as 0 was also created. This second random variable reflected the unacceptability of the imprecise utterance in S1-S4. The same procedure was followed for Experiment 1 with *Yes*- versus *No*-responses, with *Unsure*-responses always being coded as 0. The four binomial variables were appended and coded based on 1) whether the observation belonged to Experiment 1 or 2. We refer to this factor as EXPERIMENT; and 2) whether the imprecise utterance was *accepted* (i.e., *Yes* in Experiment 1, and *First* in Experiment 2), or *rejected* (i.e., *No* in Experiment 1, and *Second* in Experiment 2). We refer to this factor as ACCEPTABILITY. Response proportions by item, scale-point and experiment were obtained within each ACCEPTABILITY level. This summarized variable was used as the dependent measure in all subsequent analyses. A series of linear mixed effects models were fit to the data pertaining to each scale-point, with EXPERIMENT, ACCEPTABILITY and their interaction as fixed effects, and random intercepts and slopes by item scale. At all scale-points, we found a significant interaction effect between EXPERIMENT and ACCEPTABILITY (S1: $\beta = -0.12$, $t = -3.04$, $p < 0.01$; S2: $\beta = -0.15$, $t = -5.07$, $p < 0.001$; S3: $\beta = -0.13$, $t = -4.25$, $p < 0.001$; S4: $\beta = -0.13$, $t = -3.64$, $p < 0.001$; S5: $\beta = 0.07$, $t = 2.87$, $p < 0.01$).[4]



(a) Imprecise utterance rejected.

(b) Imprecise utterance accepted.

Figure 8: Experiments 1 & 2 comparison.

In order to further probe the interactions, we conducted simple effects analyses for each scale-point with EXPERIMENT as the independent factor and by-item random intercept and slopes. This analysis is visualized in Figure 8. Results show that proportions of responses rejecting the lower standard (left panel in Figure 8) were significantly lower in Experiment 2 compared to Experiment 1 in all the lower scale-points (S1: $\beta = -0.16$, $t = -4.02$, $p < 0.001$; S2: $\beta = -0.17$, $t = -5.99$, $p < 0.001$; S3: $\beta = -0.19$, $t = -5.14$, $p < 0.001$; S4: $\beta = -0.17$, $t = -4.05$, $p < 0.001$). No significant difference was observed at the precise scale point (S5: $\beta = 0.01$, $t = 0.65$, $p > 0.1$). We now move on to the analysis of responses indicating participants' acceptance of the lower standard (right panel of Figure 8b). No significant differences were found between *Yes*- and *First*-responses in most of the lower scale points (S1: $\beta = -0.04$, $t = -1.35$, $p > 0.1$; S2: $\beta = -0.02$, $t = -0.67$, $p > 0.1$; S4: $\beta = -0.04$, $t = -1.74$, $p > 0.05$), although S3 and S5 did reach significance (S3: $\beta = -0.05$, $t = -2.74$, $p < 0.05$; S5: $\beta = -0.06$, $t = -2.69$, $p < 0.05$).

Comparison analyses of Experiments 1 and 2 show that proportions of *First*-responses in Ex-

[4] Due to space constraints, we do not include model estimates and significance levels for the main effects. However, we note that such comparisons do not have any bearing on the hypotheses under consideration.

periment 2 and *Yes*-responses in Experiment 1 were comparable across scale-points S1-S4, with the exception of S3, where the proportion of *First*-responses was significantly lower. Conversely, the proportions of *Second*-responses in Experiment 2—a choice that is only compatible with a higher SoP—were systematically lower than *No*-responses in Experiment 1 across all the imprecise scale-points (S1-S4). This decrease in selection rates of *Second*-responses was therefore incurred by the high selection rates observed for *Both*-responses, a point that we return to in the General Discussion. Overall, our findings suggest that the acceptability of the imprecise utterance remains largely stable across the two experiments, a finding that can be better accommodated by H2, which states that metalinguistic denials act only as a request to raise the SoP. Importantly, we also find that rejections of the imprecise utterance *decrease* across the two experiments. This second finding argues against H1, which predicts higher rejection rates of the imprecise utterance after a metalinguistic denial. Taken together, our results therefore allow us to conclude that challenging the assertability of an imprecise utterance does not incur an automatic update of the SoP.

**4. General discussion.** In this section, we discuss our findings in the larger context of the literature on imprecision. We first note that both Experiments 1 & 2, as well as our norming study, showed a clear preference for precise interpretations over imprecise ones. These results replicate previous findings showing that precise interpretations are not only judged to be more acceptable, but are also processed faster (Syrett et al. 2010, Aparicio et al. 2015, Aparicio Terrasa 2017, Leffel et al. 2017, Ronderos et al. 2024).

We now turn to the research question addressed in the current paper. Our starting point was the observation that the move to raise the SoP by means of a metalinguistic denial is hard to resist (Lewis 1979, Klecha 2018). Our first hypothesis (H1) proposed that this difficulty is a direct consequence of the discursive import of metalinguistic denials. In particular, H1 states that denials effectively raise the SoP. As has been discussed, our results argue against this view: the acceptability of an imprecise utterance overall was not negatively affected when embedded in a disagreement dialogue. Our results can be better accounted for by the second hypothesis under consideration (H2), which states that metalinguistic denials act only as a request to raise the SoP. It is important to note that our findings do not argue against previous claims that metalinguistic denials more or less cause the listener to accommodate to a higher SoP. Our results however suggest that, to the extent that this accommodation takes place, it ought to be overtly signaled in a subsequent conversational move, such as a concession or a retraction. In this respect, accommodating to a higher SoP differs from better understood types of accommodation, such as presupposition accommodation (Cf. Klecha (2018) for a similar observation).

If H2 is on the right track, and metalinguistic denials do indeed act as a request to raise the SoP, it remains an open question what the specific effect of this move is on the SoP. One possibility is that the lower SoP remains operative until the listener either overtly agrees to adopt a raised threshold, or further challenges it. A second possibility is that the denial has the effect of suspending the SoP, such that it is temporarily undefined until the subsequent conversational move. This view is partially supported by one aspect of our findings, namely the high rates of *Both*-responses obtained in S1-4 in Experiment 2. The fact that in this experiment participants judged the disagreement to be faultless at high rates suggests that they took both the lower and the higher standard, to be operative. One possible way of reconciling these two incompatible parametrizations would be to assume that as long as the two speakers have incompatible beliefs about the SoP, this parameter ought to be undefined in the common ground.

**5. Conclusion.** This paper investigates the discourse dynamics of imprecision, focusing on whether metalinguistic disagreements affect the acceptability of imprecise utterances. Building on previous claims that such disagreements trigger accommodation to a higher standard of precision (SoP) (Lewis 1979, Klecha 2018), we tested two hypotheses. Hypothesis 1 (H1), consistent with prior work, proposes that metalinguistic denials automatically raise the SoP, rendering the original imprecise utterance unacceptable. Hypothesis 2 (H2) suggests that metalinguistic denials serve as a request to raise the SoP rather than enforcing an update, allowing the lower SoP to remain operative. Our results show that imprecise utterances remain as acceptable after a metalinguistic denial as they are in isolation, in line with H2. This indicates that lower SoPs persist even after metalinguistic challenges, with disagreements prompting but not mandating higher SoP adoption. In ongoing work, we investigate whether and how the discourse commitments (Lauer 2012) incurred by subsequent conversational moves (e.g., concessions vs. retractions) update the SoP.

## 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. https://doi.org/10.3765/salt.v25i0.3128.

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

Barker, Chris. 2002. The dynamics of vagueness. *Linguistics and Philosophy* 25(1). 1–36.

Barker, Chris. 2013. Negotiating taste. *Inquiry* 56(2-3). 240–257.

Beltrama, Andrea & Florian Schwarz. 2021. Imprecision, personae, and pragmatic reasoning. *Semantics and Linguistic Theory (SALT)* 31. 122–144. https://doi.org/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. https://doi.org/10.3765/salt.v1i0.5406.

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

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

Horn, Laurence R. 1989. *A natural history of negation*. Chicago: University of Chicago Press.

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. https://doi.org/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.

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

Kennedy, Christopher. 2013. Two sources of subjectivity: Qualitative assessment and dimensional uncertainty. *Inquiry* 56(2–3). 258—277. https://doi.org/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.

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

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. https://doi.org/10.1515/9783110894653.439.

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

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

Lasersohn, Peter. 1999. Pragmatic halos. *Language* 75(3). 522–551.

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

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

Leffel, Timothy, Ming Xiang & Christopher Kennedy. 2017. Interpreting gradable adjectives in context: Domain distribution vs. scalar representation. Unpublished Manuscript.

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

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

Pecsok, Emily & Helena Aparicio. 2024. How can they both be right?: Faultless disagreement and semantic adaptation. *Proceedings of the Annual Meeting of the Cognitive Science Society* 46. 3939–3945.

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. https://doi.org/10.5070/G60111411.

Rotstein, Carmen & Yoad Winter. 2004. Total adjectives vs. partial adjectives: Scale structure and higher-order modifiers. *Natural Language Semantics* 12. 259–288.

Syrett, Kristen, Christopher Kennedy & Jeffrey Lidz. 2010. Meaning and context in children's understanding of gradable adjectives. *Journal of Semantics* 27(1). 1–35. https://doi.org/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. https://doi.org/10.1111/1468-0017.00207.

Zehr, Jeremy & Florian Schwarz. 2018. *PennController for Internet Based Experiments (IBEX)*. https://doi.org/10.17605/OSF.IO/MD832.