# Straight enough: Deriving imprecise interpretations of maximum standard absolute adjectives

Camilo R. Ronderos[1], Ira Noveck[2], & Ingrid Lossius Falkum[1]

[1] University of Oslo, Department of Philosophy, Classics, History of Art and Ideas

[2] Laboratoire de Linguistique Formelle, UMR 7110 -Université de Paris & CNRS

Author note

The ORCID IDs of the authors are as follows:

Camilo R. Ronderos: https://orcid.org/0000-0002-7779-406X

Ira Noveck: https://orcid.org/0000-0002-3401-9629

Ingrid Lossius Falkum: https://orcid.org/0000-0002-1203-8036

Correspondence concerning this article should be addressed to Camilo R. Ronderos, Blindernveien 31 Georg Morgenstiernes hus, room 533, Oslo. E-mail: camilorr@uio.no

Abstract

While maximum standard absolute adjectives (such as *straight*) typically have a precise meaning (e.g., 'perfectly straight'), they are also regularly used imprecisely (e.g., to mean 'straight enough'). The current study investigates how contextual expectations of precision and a visual referent's conceptual distance from an ideal maximum standard influence the processing effort of precise and imprecise interpretations of these adjectives. In three experiments, we showed native speakers of English images depicting objects that could be referred to precisely or imprecisely via an absolute adjective and asked them to select the image that best matched the written sentence (Experiments 1 and 2) or to read sentences containing maximum standard absolute adjectives (Experiment 3). Experiment 1 presented no discourse context and participants accepted on average only a small degree of imprecision; and when they did, they took longer relative to cases in which the same adjectives were used precisely, which is in line with existing empirical findings. Experiment 2 contrasted two kinds of discourse contexts (raising high or low expectations of precision) before the presentation of the test sentences. When expectations of precision were high, participants tolerated only a small degree of imprecision and when they did, it came at a cost, as in Experiment 1. When expectations of precision were low, much larger degrees of imprecision were tolerated but, critically, participants were still overall faster to reach precise, relative to imprecise, interpretations in supporting contexts, suggesting that accessing the precise meaning is less effortful. Experiment 3 supported these findings by showing how the cost of understanding imprecision is also present in a self-paced reading task. Our results lend support to the view that maximum standards are part of the encoded meaning of these adjectives.

*Keywords:* imprecision, absolute adjectives, semantic underspecification, experimental pragmatics

*Word count:* 10600

Straight enough: Deriving imprecise interpretations of maximum standard absolute adjectives

## 1. Introduction

People routinely use language imprecisely. For example, someone can say that a hand-drawn line is *straight* even if it isn't *perfectly straight*. The relationship between adjectives such as *straight* (which are known as maximum standard absolute adjectives, or MSAAs) and their precise or imprecise interpretation is complex. While one might agree with a four-year-old who describes their freely hand-drawn line as *straight*, it is another matter to accept *straight* as a description from a hired carpenter who is installing a visibly bent beam. However, even if hastily drawn by a child, a line that is too 'squiggly' will hardly be accepted as *straight*. So how does context interact with the meaning of MSAAs to generate precise and imprecise interpretations? And what can this tell us about the encoded meaning of MSAAs?

In general, the way that language is used imprecisely has received a lot of attention through the years. Some have studied how and when numbers are used imprecisely (e.g., saying that it's 3 o'clock when the time is actually 2:56) (Beltrama & Schwarz, 2022; Dehaene & Mehler, 1992; Gibbs & Bryant, 2008; Solt et al., 2017; Van Der Henst et al., 2002). Others have focused on MSAAs specifically, formulating theories on how people reach an imprecise interpretation of these adjectives (e.g., Kennedy, 2007; Lasersohn, 1999). Further, recent work has tested these theories experimentally (e.g., Aparicio et al., 2016; Leffel et al., 2016).

However, how MSAAs interact with contextually raised expectations of precision remains understudied. This is important to elucidate since it can help us better understand how MSAAs are represented in the mental lexicon: Is precision part of the stored representation of maximum standard absolute adjectives such as *straight*? How is their lexical meaning adjusted when context calls for imprecise interpretations? These are the questions we address in the present article.

Concretely, we investigate how two factors - discourse expectations of precision and a referent's conceptual distance from the precise standard - affect the comprehension of MSAAs. Before we present our three Experiments addressing these issues, we first provide a brief overview of the relevant theoretical and empirical literature in the following section.

### 1.1 Interpreting maximum standard absolute adjectives

In the semantics and pragmatics literature, many have claimed that deriving imprecise interpretations of MSAAs involves a pragmatic adjustment (Burnett, 2014; Carston, 2010; Lasersohn, 1999; Leffel et al., 2016; Lewis, 1979; Recanati, 2010; Wilson & Sperber, 2012, i.a.) (see Lassiter & Goodman, 2013 for a different view). One popular theory, for example, describes MSAAs as being associated with closed scales (i.e., with a clear beginning and an end) and as being end-point oriented (Kennedy, 2007; Kennedy & McNally, 2005). MSAAs require that their arguments have the maximal degree of the appropriate property (Syrett et al., 2010): When a line is described as *straight*, it means that it has the maximal degree of 'straightness', i.e., it is at the endpoint of the 'straightness' scale. In this approach, a literally false description is tolerated as being 'close enough' if a high degree of precision is not pragmatically relevant (see Lasersohn, 1999). From a psychological perspective, it could be said that this view would consider precision (i.e., the requirement that the argument possesses the maximal degree of the appropriate property) to be part of the encoded meaning of MSAAs. The meaning can be modified contextually, but this is a potentially costly pragmatic adjustment.

Another view that considers imprecision to involve a pragmatic adjustment is that of Relevance Theory (RT) (Sperber & Wilson, 1986/1995). In RT, imprecision is seen as a variety of lexical broadening, the outcome of a mechanism of ad hoc concept construction. Through this mechanism the encoded meaning of words is contextually loosened in accordance with context-

specific expectations to include a larger set of possible referents (Carston, 2002, 2010). Following this view, there is a continuum of pragmatic instances of broadening ranging from approximation (i.e., imprecision) all the way to cases typically described as figurative language, such as metaphor (Sperber & Wilson, 2008; Wilson & Carston, 2006, 2007). This view, which treats imprecise uses as a variety of ad hoc construction, presupposes a precise semantics of MSAAs which is subject to pragmatic broadening in context. However, the broader relevance-theoretic account of the contextualized interpretation of substantive words (nouns, verbs, adjectives) is also compatible with a view in which these adjectives are semantically underspecified regarding their precision (Carston, 2002, 2013, 2016). In fact, this is the position taken in much of the recent literature on the representation of word meanings: because most substantive (open-class) words are associated with several related senses (i.e., they are polysemous), it is assumed that the different senses are pragmatically derived from a single underspecified representation which encompasses all the semantically related senses of the word that are known to the language user (e.g., Carston, 2013, 2016; Frisson, 2009; Pietroski, 2005; Ruhl, 1989). While there is some experimental evidence to support this claim (Beretta, et al., 2005; Frisson & Pickering, 2001; Pickering & Frisson, 2001; Pylkkänen et al., 2006; Rabagliati & Snedeker, 2013), what exactly this underspecified representation amounts to remains a matter of contention (Falkum & Vicente, 2015; Vicente, 2018). In the case of maximum standard absolute adjectives, such an underspecified representation would have to exclude the requirement for an argument to possess the maximal degree of a property (i.e., the requirement to be *precise*), considering that it would have to allow both precise and imprecise interpretations.

To sum up, while the semantic literature on MSAAs broadly posits that precision is part of the meaning of these adjectives, work on the mental representation of word meanings suggests that MSAAs could be underspecified regarding their degree of precision, with this information being

filled in by the relevant context. One way of testing these hypotheses is to examine the relative comprehension cost of interpreting MSAAs.

### 1.2 The cost of deriving precise and imprecise interpretations

From the views described above, it is possible to derive a processing account that sees precision as part of the encoded 'literal' meaning of MSAAs. According to this 'semantic view', understanding MSAAs imprecisely should generally bring about a comprehension cost relative to understanding them precisely. A plausible alternative to this view can be derived from the literature on underspecified lexical entries. According to such an 'underspecification view', precision is not part of the lexical meaning of MSAAs and should instead be contextually derived. Comprehension costs associated with precise and imprecise readings will depend on their contextual availability.

Two studies investigating the comprehension of precise and imprecise interpretations of MSAAs lend credence to the 'semantic view'. Consider first the landmark study from Syrett et al. (2010), who introduced a *presupposition assessment* task to directly investigate how both children (ages 3-5) and adults interpret sentences containing MSAAs. In the first of their three experiments, participants heard a request that presupposed the existence or the uniqueness of a referent (e.g., *Show me the full one*, when said of a jar). This put the participants in a position to choose one of three options – to choose between one of two jars, which varied with respect to their fullness, or else a third option indicating neither. In one 'felicitous use' condition, there was a referent that cohered with a precise reading (a full jar). In another 'infelicitous use' condition, there were instead two unsatisfactory referents (two half-empty jars, one fuller than the other). The results showed that both children and adults correctly picked the right referent in the felicitous condition, thus displaying an ability to derive precise interpretations. In the infelicitous condition, adults

tended to reject both referents (88% rejection rate in Experiments 1 & 3) while the children did so only 40% of the time. The children were more likely to select the referent closest to the precise standard, e.g., they chose the jar with the most liquid at least 60% of the times in all three experiments. Upon closer inspection, the authors noted that children took significantly longer to select a referent in the 'infelicitous' relative to the 'felicitous' condition. This led the authors to formulate an explicit linking hypothesis regarding the relationship between the derivation of an imprecise interpretation of an MSAA and picture selection time. They argued that understanding imprecision requires a contextually-motivated assessment of how much deviation from the maximal standard will be tolerated. This assessment involves reasoning that goes past the computation of semantic content, and as such, the authors argue, it should come with an added cost relative to deriving a precise interpretation.

The second relevant paper, from Bambini et al. (2013), compared the processing profile of three different types of pragmatic enrichment (metaphor, metonymy and approximation) to their respective literal counterparts via timed sensicality judgements. They found that participants took longer to react to sentences with an approximative interpretation of an adjective (*That land is flat*) relative to sentences with a more precise, 'literal' interpretation (*That spatula is flat*). Participants also displayed lower accuracy in their sensicality judgements in the approximate relative to the literal condition. Within the approximation group, the study included imprecise usages of MSAAs such as *straight* and *full*. Their study also included approximate usages of color adjectives (*black, red, pink, yellow*, etc.) as well as of several other adjective types (*silent, blinding, long, unchanged, boiling*, etc.).

Given the findings of Bambini et al. (2013) and Syrett et al. (2010), one might argue that precision is part of the lexical meaning of MSAAs, with imprecision requiring a pragmatic adjustment that brings additional effort. However, we see it as premature to jump to that

conclusion. Here we make five observations, four with respect to the Syrett et al. (2010) findings and one more with respect to Bambini et al. (2013), that give us pause. First, Syrett et al. (2010) investigated only two MSAAs across their experiments: *full* (Experiments 1 and 2) and *straight* (Experiment 3). Second, in each of their experiments, their participants only saw a single trial facilitating an imprecise interpretation. Third, given that a majority of adults in the study were quite intolerant of imprecision, it could not be determined – for lack of sufficient data – whether the reaction time delays for imprecision found for children would also hold for adults. It is an open question as to whether this pattern would persist if adults saw a visual referent described imprecisely that was closer to the precise maximal standard. Fourth, and as far as Bambini et al. (2013) is concerned, it is difficult to make strong claims from this study since MSAAs made up only a small part of their item set and MSAAs are different in their semantics relative to the other adjectives that they tested (see, e.g., Kennedy, 2007). A fifth comment on previous studies pertains to the marked differences in interpretation. While Syrett et al. (2010) found that adults were rather intolerant of imprecision (rejecting imprecise interpretations 80% of the time), Bambini et al. (2013) found that their participants accepted approximate interpretations at a rate of 87%. In addition, a study on how imprecise adjectives generate contrastive reference effects suggest that imprecise interpretations are readily available (Leffel et al., 2016).

How can these incompatible findings be accounted for? One possibility is that they arise because of differences in the degree of pragmatic slack (Lasersohn, 1999) that participants were asked to tolerate (e.g., interpreting a half-empty jar as *full* vs. an only slightly bent line as *straight*). This makes it critical to consider the effect of a referent's conceptual distance from the precise standard on both comprehension rate (i.e., tolerance of imprecision) and comprehension effort.

A second critical factor to consider is the role of context when it elicits different expectations regarding the level of precision. Previous research has shown that reasoning about a speaker's

intended level of precision affects how people answer questions (Potts, 2012) and acquire the precise meanings of MSAAs from imprecise input (Lee & Kurumada, 2021). Additionally, speaker-specific contextual information can be used as a cue to decide whether to accept or reject an imprecise interpretation of a numeral expression (Beltrama & Schwarz, 2022). Information stemming from the visual context also constrains comprehension. Leffel et al. (2017) found that the degree to which an image matched the precise standard of an expression involving an MSAA (e.g., the number of stripes on a shirt described as a plain t-shirt) determined the acceptance of the picture as an accurate referent of the expression, with participants only tolerating a minimal amount of imprecision. Aparicio et al. (2016) argued that the context-sensitivity of MSAAs is due to pragmatic reasoning about the threshold for imprecision. In an eye-tracking Visual World study, they found that (precisely used) MSAAs generate contrastive reference effects, but do so later than relative adjectives (such as *tall* or *big*). They conclude that MSAAs carry an additional processing cost (compared to relative adjectives) because of the need to contextually set the precision threshold.

These results point to the specific context-sensitivity of MSAAs, but, critically, do not speak to the potential differences in how contextual expectations help comprehenders reach precise and imprecise interpretations of MSAAs. This is necessary to elucidate considering some of the claims made in the literature on semantic underspecification. Here, it has been reported that when specific contextual expectations are raised, the processing effort of polysemous words is reduced, so that no single sense of a word can be said to be more salient than another. Frisson and Pickering (2007) make this claim for the case of metonymy. They report that, while unfamiliar metonymies (*She often read Needham*) are read more slowly than literal counterparts (*She often met Needham*), this effect disappears when sentences are embedded in an appropriate context. Frisson (2009) sees this as evidence for an underspecified lexical representation: Even when a

name has never been heard before and a metonymic interpretation should, *prima faciae*, require additional effort, this effort disappears when contextual expectations are met.

An 'underspecification view' of MSAA comprehension would therefore posit that when contextual expectations are given, precise and imprecise readings of MSAAs should not differ in their comprehension effort. This is because precise and imprecise senses of an MSAA should be derived from a single lexical meaning that is underspecified regarding precision (and, contrary to the 'semantic view', there is no requirement that the argument possess the maximal degree of a relevant property).

### 1.3 The present work

Our study has two overarching goals. First, we aim to build on previous empirical studies on imprecise MSAAs (e.g., Bambini et al., 2013; Syrett et al., 2010) to understand whether discrepancies regarding participants' tolerance for imprecision can be explained by the varying effect of a referent's conceptual distance from a precise standard. This will additionally help us understand how different degrees of precision are comprehended and how this influences comprehension effort. This issue is addressed in Experiment 1. Second, we aim to better understand the processing cost associated with deriving contextually licensed precise and imprecise interpretations of MSAAs. To do this, we tested how expectations of precision and a referent's conceptual distance from a precise maximal standard interact to mediate the comprehension rate and effort of precise and imprecise MSAAs. In turn, this should help us better understand the encoded lexical meaning of MSAAs by testing the 'semantic' and 'underspecified' views of MSAAs. This is the focus of Experiment 2. In Experiment 3, we present a more direct test of the two hypotheses by investigating the processing cost of precise and imprecise interpretations during self-paced reading.

## 2. Experiment 1

The goal of Experiment 1 was to investigate how a referent's varying degree of conceptual distance from the precise standard impacts comprehension for a range of MSAAs (*straight, round, clean, empty, full*, and *closed)*. Presumably, the more uneven a line is, the less likely it is that it will be accepted as a straight line. However, it is uncertain whether this will translate to a delay in comprehension time for all degrees of distance from the precise standard. It is also uncertain whether such a pattern would generalize to different MSAAs in various adjective-noun combinations. For our investigation, we adopted and extended the paradigm of Syrett et al. (2010) to include both (a) different instances of maximum standard absolute adjectives and (b) varying degrees of imprecision. The predictions of Experiments 1, 2 and 3 were pre-registered. All materials, data, scripts and pre-registration forms can be found on the project's OSF repository: https://osf.io/6gwdq/ .

### 2.1 Norming Study 1

The first step was to create visual depictions of objects that would be perceived as systematically varying in distance from the precise maximal standard of the described property (e.g., a line with different degrees of 'straightness'). For this purpose, we made one precise and one imprecise picture for each of our critical items. We then asked a group of 30 participants to state whether each of the images depicted the intended adjective. For example, participants read the sentence *show me the straight line* and were asked to select the target picture, a distractor picture, or to state that neither of the pictures showed a straight line. We then examined the results averaged by each critical item. Depending on how often each imprecise image was accepted as the target to the sentence, we created 2 additional versions of each image, so that each item would have 3 different degrees of imprecision. In other words, if one of the target

pictures in Figure 1 had an average rating of ~50%, we created one image with a higher level of imprecision and one with a lower level of imprecision in order to have three degrees of imprecision per item, in addition to a precise condition, and one unrelated control condition (depicting the correct property but applied to an incorrect referent), creating all of the images shown in Figure 1. If an image had an average rating of ~80%, we created two images that were incrementally more imprecise, and if an image had an average rating of ~30%, we created two incrementally more precise images. After creating two new pictures per item, we conducted a pilot study with 90 participants in order to examine whether the average ratings in each one of the three categories ('Imprecise 1', 'Imprecise 2' and 'Imprecise 3') would match our expectations (we also collected reaction times for this pilot study in order to calculate the required number of participants for Experiment 1, as we explain below in the 'Participants' section). The pilot study confirmed that, on average, our images captured the intended degrees of imprecision. Pictures in the 'control' condition were accepted as correct in 4% of all trials. Pictures in the 'Imprecise 3' condition were accepted 27% of the time, pictures in the 'Imprecise 2' condition 57% of the time, and pictures in the 'Imprecise 1' condition 78% of the time. Finally, pictures in the 'precise' condition were accepted 100% of the time. We take this as confirmation that the images were perceived as displaying distinct levels of imprecision.

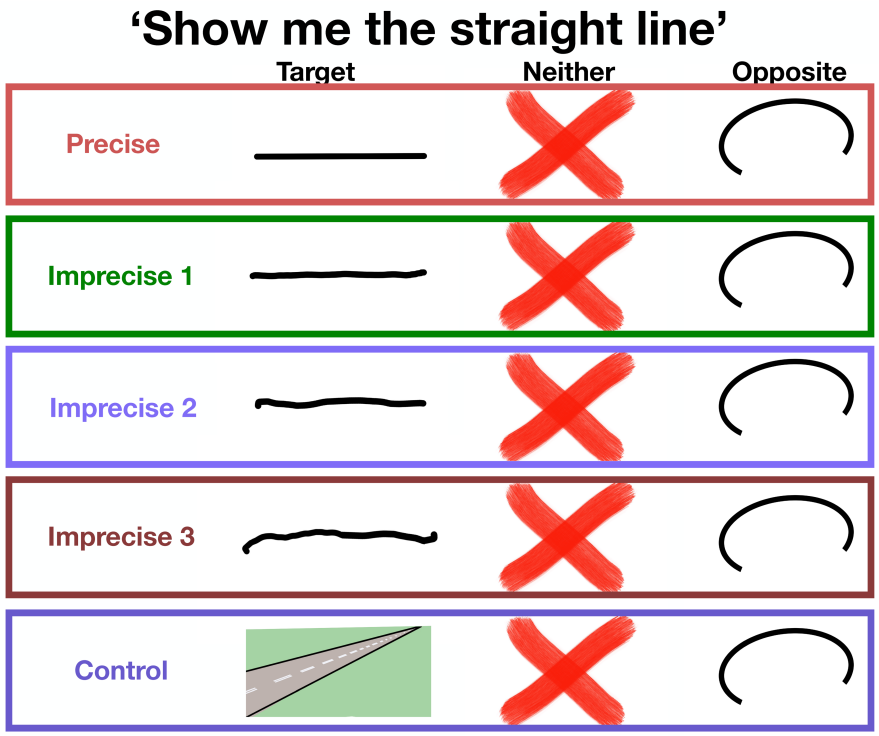


*Figure 1.* Example of a target trial in Experiment 1 in the five conditions

**2.2 Materials and design**

Experiment 1 was programmed using the PCIbex software (Zehr & Schwarz, 2018) and was run via the internet. It was based on the seminal task developed by Syrett et al. (2010), introducing various changes to it to adapt the task to a web-based paradigm and to best allow for the use of multiple MSAAs. (and was identical to that of our pilot study). In each trial, participants saw three images. One image was always a red X. The other two were the Target and Distractor images. Participants could only select the images using the J, K, and L keys of their keyboard, representing the picture in the left, middle, and right side of their screens respectively. The Distractor picture displayed an object with a mismatched property (e.g., a completely curved line). The Target picture changed according to the experimental condition. It was either a perfectly straight line (‘Precise’ condition), a slightly squiggly line (‘Imprecise 1’ condition), an

even squigglier line ('Imprecise 2' condition), a very squiggly line ('Imprecise 3' condition) or a picture depicting an unrelated object that could also be felicitously described as *straight* (e.g., a road). We introduced this final, 'control', condition so that there would be cases in which participants should unequivocally select the red X, i.e., reject both Target and Distractor images. This means that Experiment 1 had one factor with five levels. Participants read a sentence in each trial instructing them to select one of the pictures (*Show me the straight line*). The sentence was always shown above the three images. Position of the images was pseudo-randomized: The red X always appeared in the middle, and the position of the target and distractor images was randomized between left and right. This way, participants knew that the potential referent could only appear in one of two locations, and that if neither of them was appropriate they had to always press the K key. An example of the visual display and corresponding sentence in each of the five conditions is shown in Figure 1.

We created a total of 12 critical items. Each item consisted of a combination of a sentence (e.g., *Show me the straight line* in Figure 1) and the possible pictures in each condition. The sentence was identical across conditions. The 12 critical items used 6 different maximum standard absolute adjectives (*straight, round, clean, empty, full*, and *closed*), meaning that there were two items per absolute adjective (e.g., *straight line* as well as *straight arrow, round cookie* and *round sun*). The 5 versions of each critical item were distributed across five lists in a Latin square fashion, so that each participant would only see a single version of each item (the lists of each experiment are available on the OSF repository). Additionally, we created 16 filler trials. The filler trials had the same structure as the critical ones but used different types of adjectives, such as gradable adjectives (e.g., 'show me the long snake'; all adjectives were *big*, *tall, long, old, fat, skinny),* color adjectives (e.g., 'show me the pink pants'; all adjectives were *green*, *pink, blue, yellow, purple) and emotional adjectives* ('show me the happy friends'; all adjectives were *angry,*

*happy, surprised, disgusted)*. Half of the fillers required selecting the red X as the correct response (e.g., the sentence was 'show me the skinny dog' but the two pictures showed cats and not dogs), while the other half had a unique correct target image.

### 2.3 Procedure

Participants were told that they would read a sentence asking them to select an image. They were told that if it was not possible to select an image based on that instruction, they should select the red X instead. Prior to the experiment, they were shown an image depicting how their right hand should remain poised above the response keys (J, K, and L) and how they should only use their right index to press the J key, their middle finger to press the K key and their ring finger to press the L key. They were further told that it was paramount to use this response method only and to select an image as quickly as possible. They saw 4 practice trials (involving color and relative adjectives only) before beginning the experiment, two of which required selecting the red X as the correct response. After the practice trials, instructions were repeated, emphasizing the importance of response speed and of the required position of the hand. All critical and filler items were then presented in a pseudo-randomized order, making sure that there was at least one filler item in between critical trials. Participants took 8 minutes on average to complete the study.

### 2.4 Participants

For Experiment 1, we recruited 200 right-handed monolingual speakers of American English between the ages of 18 and 35. They were recruited using the Prolific recruitment portal and were given the equivalent of 0.7 dollars as compensation. This number was determined after conducting an a priori power analysis via simulations using the R package SIMR (Green & MacLeod, 2016), based on the results of the pilot study. To start, we used the parameters of the linear model fitted to the reaction times of the pilot data to simulate 1000 new data sets. The only

parameter we changed was the effect size found in the pilot for the difference in picture selection time between 'Precise' and 'Imprecise 1' conditions (Cohen's D = 0.38). We replaced it with a more conservative estimate (Cohen's D = 0.2) given how small scale experiments tend to exaggerate the estimate of the magnitude of an effect (Button et al., 2013; Ioannidis, 2008). The results showed that with 200 participants we would expect to have 80% power to detect a true effect that is at least as big as the assumed estimate of Cohen's D = 0.2.

**2.5 Predictions**

In Experiment 1 we examined both the picture selection rate (i.e., did they choose the target image, or the red X?) and the picture selection time (i.e., how long did it take participants to accept the target picture as the referent?). First, in terms of picture selection rate, we hypothesized that acceptance of a target picture as the intended referent would gradually decrease as the depicted property moves away from the standard. However, in contrast to the findings of Syrett et al. (2010) for adults, we predicted that participants would accept imprecise pictures in the 'Imprecise 1' condition as adequate referents to the adjective at a level higher than chance. This prediction is based on the observation that people routinely use and understand adjectives imprecisely.

In terms of picture selection time, we predicted that - in the absence of a discourse context allowing participants to adjust their expectations - participants would be significantly faster to select the target picture as the referent in the 'Precise' condition compared to all other conditions, in line with the findings of Syrett et al. (2010). This would suggest that imprecise interpretations of maximum standard absolute adjectives are derived at a cost relative to precise ones when participants do not have enough prior information to adjust their expectations. This prediction is in line with what a 'semantic view' of the encoded meaning of MSAAs would predict. Further,

we predicted that there would also be a gradient effect of condition on picture selection time, with reaction times in the 'Imprecise 1' condition being shorter than in the 'Imprecise 2' and 'Imprecise 3' conditions. The 'Imprecise 3' condition should show the overall longest reaction times. These findings would suggest that distance from a precise maximal standard is a determining factor in the comprehension of maximum standard absolute adjectives. More specifically, these findings would suggest that the precise maximal standard is the one that most closely aligns with the semantics of the maximum standard absolute adjectives, with the retrieval of a stored sense from the lexicon being less costly than adjusting the sense of the adjective to fit the depicted degree of imprecision.

**2.6 Analysis and results**

Prior to analysis, we removed participants who scored lower than 75% accuracy on the filler trials, as per our pre-registration. This reduced the total number of participants to 198. We then chose to deviate from our pre-registration (where we stated that RTs shorter than 200 ms and longer than 10000 ms would be removed) by not removing any outliers based on picture selection time. We decided on this given the evidence that outlier-removal strategies based on RTs can bias the means across conditions, a bias which increases with larger sample sizes (Miller, 1991). We maintained this deviation from the pre-registration for Experiment 2. Further, we deviated from our pre-registration (in this and in the subsequent experiments) by analyzing the picture selection rate using a treatment-contrast coding scheme. This was done to match the contrast coding scheme used to model reaction times, and to better examine the pattern of interactions in Experiments 2 and 3. All analyses for Experiments 1, 2 and 3 were conducted using R (R Core Team, 2020) and R-Studio (RStudio Team, 2020). For data processing, visualization and analysis, we used the following packages: ggplot2 (Wickham, 2016), lme4 (Bates et al., 2007),

Rmisc (Hope, 2013), MASS (Ripley et al., 2013), dplyr (Wickham et al., 2020), DoBy (Højsgaard, 2012), papaja (Aust & Barth, 2017), here (Müller, 2017), and afex (Singmann et al., 2015).

**2.6.1 Picture selection rate.** To examine the picture selection rate, we fitted a mixed-effects logistic regression model. Responses were coded as 1 if they selected the target, and 0 if they selected the red X. Trials in which participants selected the distractor image were removed prior to analysis, resulting in the exclusion of 25 trials or around 1% of the data. The model had a treatment-contrast coding scheme, with the 'Precise' condition coded as the intercept. This coding scheme allows us to evaluate tolerance for imprecision relative to a precise interpretation. The random effects structure of the model included random intercepts by items and by participants and a random slope by items, excluding random correlations between intercepts and slopes (in R syntax: selected ~ condition + (1|participant) + (1+ condition||Item) ). This was the maximal converging model, following Barr et al. (2013). The model shows that all imprecise conditions were accepted at a significantly lower rate relative to the 'Precise' condition (all p-values < 0.001). The 'Precise' condition showed an acceptance rate of 99.6%, followed by the 'Imprecise 1' (89%), the 'Imprecise 2' (50%), and the 'Imprecise 3' (18%) conditions. The 'Control' condition was rejected 97.3% of the time. The results can be visualized in panel A of Figure 2, and the output of the model is summarized in Table 1.

**2.6.2 Picture selection time.** To analyze reaction times, we focused on instances of accepting the target picture (52% of trials). Prior to analysis, we removed any instances of the 'Control' condition from the acceptance data set (~1% of trials). We then fitted a mixed-effects linear regression model. The model had a treatment contrast coding scheme, with the 'Precise' condition coded as the baseline. The model included random intercepts and slopes by items and

by participants (in R syntax: transformed_RT ~ condition + (1+ condition|participant) + (1+ condition|Item) ). For our dependent measure, we used an inverse square root transformation of picture selection times. We did this because the residuals of the model using raw reaction times were not normally distributed, and a box-cox procedure (Box & Cox, 1964) suggested this transformation as the optimal dependent measure.

The model showed a significant difference between the 'Precise' and 'Imprecise 1' conditions ($t = 8.45$, $p < 0.001$), with trials in the 'Precise' condition showing shorter reaction times. Reaction times in the 'Imprecise 2' and 'Imprecise 3' conditions were also significantly shorter than in the 'Precise' condition ($t = 5.48$, $p < 0.01$ and $t = 4.83$, $p < 0.01$, respectively. These results are shown in panel B of Figure 2 and the model output is shown in Table 2.

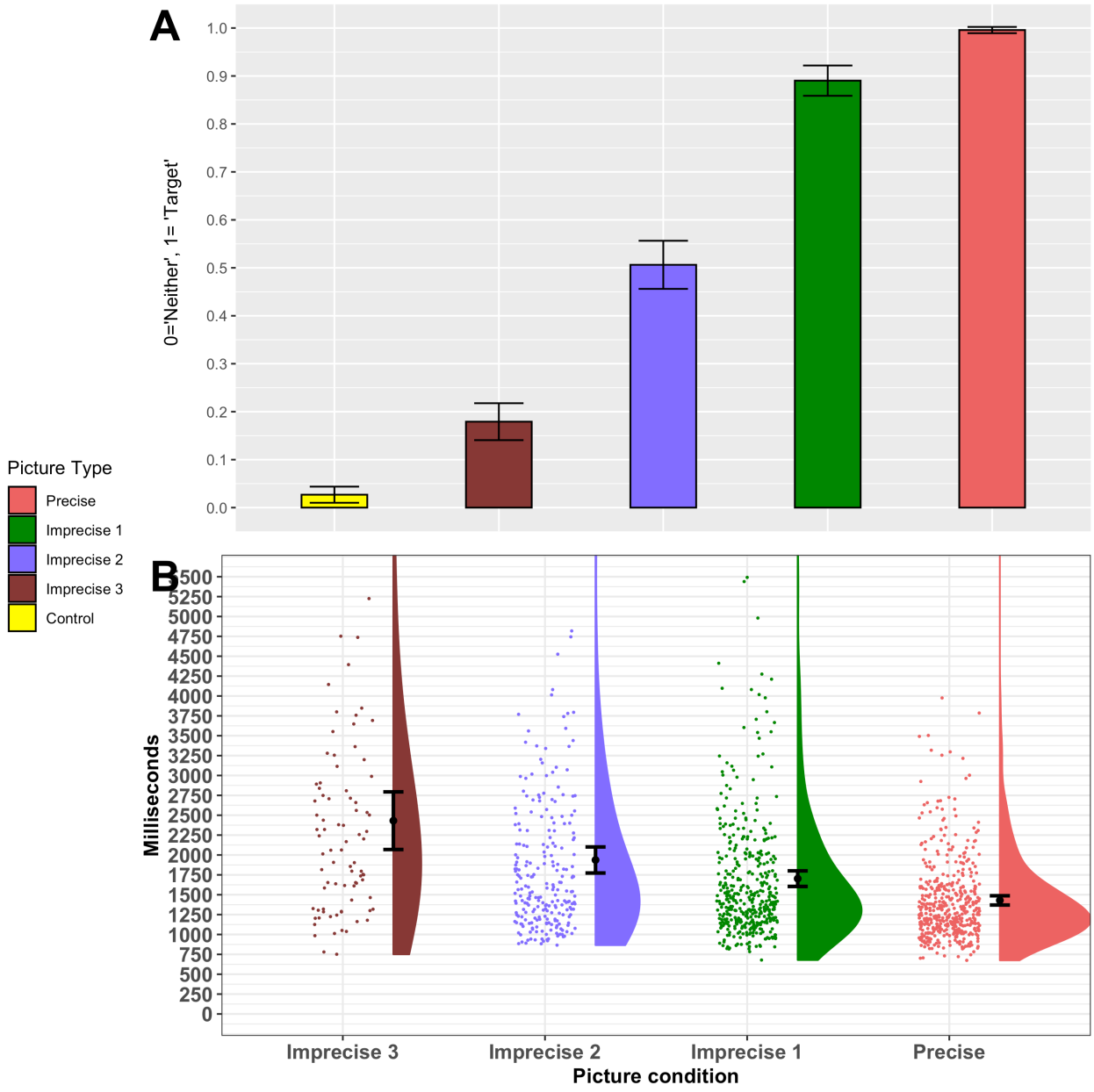


*Figure 2.* Target picture selection (panel A) and raw-reaction times (panel B) for Experiment 1. Error bars show confidence intervals.

Table 1:

*Logistic regression model of picture selection in Experiment 1*

| term | $\hat{\beta}$ | 95% CI | $z$ | $p$ |
|---|---|---|---|---|
| Precise vs. Imprecise 1 | -2.96 | [-4.71, -1.20] | -3.30 | .001 |
| Precise vs. Imprecise 2 | -6.14 | [-7.62, -4.66] | -8.13 | < .001 |
| Precise vs. Imprecise 3 | -8.21 | [-9.75, -6.67] | -10.46 | < .001 |
| Precise vs. Control | -10.69 | [-12.56, -8.82] | -11.22 | < .001 |

*Note.* Model used a treatment-contrast coding scheme

Table 2:

*Linear regression model of Target picture selection time in Experiment 1*

| term | $\hat{\beta}$ | 95% CI | $t$ | $df$ | $p$ |
|---|---|---|---|---|---|
| Precise vs. Imprecise 1 | 0.004 | [0.003, 0.005] | 7.57 | 11.21 | < .001 |
| Precise vs. Imprecise 2 | 0.006 | [0.004, 0.009] | 5.00 | 10.48 | < .001 |
| Precise vs. Imprecise 3 | 0.011 | [0.007, 0.014] | 5.90 | 8.79 | < .001 |

*Note.* Model used a treatment-contrast coding scheme

**2.7 Discussion**

Experiment 1 reports two main findings. First, we found that participants were overall very tolerant of imprecision: Pictures in the 'Imprecise 1' condition were accepted as correct referents around 90% of the time. However, as seen in Figure 1A, there was a steep decline in acceptance for the remaining imprecise conditions: Pictures in the 'Imprecise 2' condition were only accepted in 50% of all trials, and pictures in the 'Imprecise 3' condition were accepted in approximately 17% of trials. These differences in acceptance rate relative to Syrett et al. (2010) could be explained by the type of materials used: The degree of imprecision displayed by the target referent in the 'infelicitous' condition of Syrett et al. (2010) was arguably most similar to the referent in the present study's 'Imprecise 3' condition. Second, we found a delay in reaction times for deriving imprecise - relative to precise - interpretations. This confirms the findings of Syrett et al. (2010) - who found that children took longer to accept imprecise, relative to precise, interpretations in the absence of context - and extends it to an adult population and to 6 different maximum standard absolute adjectives.

These findings suggest that in the absence of any prior contextual expectations, participants readily tolerate a slight degree of imprecision (pictures in the 'Imprecise 1' condition) but nevertheless show a processing cost relative to a precise interpretation. This is in line with the findings and conclusions of Syrett et al. (2010), according to which MSAAs require their argument to possess a maximal degree of the relevant property (i.e., to be precise). Deviations from this standard require contextually adjusting the interpretation, resulting in additional processing cost. This speaks in favor of the 'semantic view' of the meaning of MSAAs.

That said, note that Experiment 1 took place in the absence of a supporting discourse context: Though there was a visual and sentential context for interpreting the MSAA, expectations of precision were not specified. It could be that if specific expectations for precision are set, the

processing cost will vary accordingly, similarly to the findings of Frisson and Pickering (2007) for metonymy. In Experiment 2, we investigate this by examining how an explicit discourse context influences the interpretation of the target sentences containing MSAAs.

## 3. Experiment 2

Experiment 1 suggests that, in the absence of an explicit discourse context, imprecise interpretations of maximum standard absolute adjectives come at a cost relative to precise interpretations. However, as discussed in the Introduction, it is likely that the derivation of precise and imprecise interpretations will be influenced by contextual expectations. Will imprecision be tolerated across the board when comprehenders have low contextual expectations of precision? Or will the differences in degrees of tolerated imprecision maintain the pattern found in Experiment 1? Importantly, how do contextual expectations affect the comprehension cost? Is a contextually licensed imprecise interpretation similarly costly to a contextually licensed precise interpretation? Experiment 2 addresses these questions.

### 3.1 Norming study 2

We created contexts that could generate opposing expectations regarding the relevance of precision for the interpretation of the critical items used in Experiment 1. For each of our critical items, we came up with single-sentence discourse contexts that would either make precision highly relevant for the interpretation (‘strict’ Context) or not relevant at all (‘loose’ Context). Figure 3 shows an example of each of the contexts for one critical item (all contexts can be found in the supplementary materials on the project’s OSF page). After this, we set out to confirm our intuitions by having a group of participants rate the expectation of precision raised by each context.

We conducted a rating study on PCIbex in which each participant read a sentence and rated it on a continuous slider scale ranging from 0-100. We recruited 50 participants on Prolific (who did not take part in any of the other studies) and assigned them to one of two experimental lists with an equal distribution of 'strict' and 'loose' contexts for each of our items. Their task was to rate the likelihood of the target referent having a high degree of precision of the intended property. For example, for the critical item in Figure 1 we asked them how *straight* they thought the line drawn by Jasmine was going to be. We additionally added 5 filler contexts that should unequivocally generate strong negative expectations, e.g., the context described a boy with a blue crayon and the question asked *how red do you think the crayon is going to be?*. Overall, participants confirmed our intuitions and rated items in the 'strict' context as generating expectations of precision (each 'strict' context rated 70 or higher, on average). All but 2 of the items in the 'loose' context were rated as not generating expectations of precision (i.e., ratings around 50 or lower). The contexts of these two items were changed to further reduce the relevance of precision. We then incorporated all contexts to the paradigm used in Experiment 1 to create Experiment 2.

| | |
|---|---|
| *Strict Context* | Jasmine carefully drew a line with a ruler on a piece of paper. |
| *Loose Context* | Jasmine rashly drew a line with her eyes closed on a piece of paper. |

*Figure 3.* Example of two context sentences for a critical item in Experiment 2.

### 3.2 Materials and design

Experiment 2 was similar to Experiment 1 with one addition: a discourse context was introduced for each item. This resulted in a 2x5 design, with the factors PICTURE TYPE (levels: 'Precise', 'Imprecise 1', 'Imprecise 2', 'Imprecise 3' and 'Control') and CONTEXT (levels: 'loose' and 'strict'). Since each item had 10 versions, they were distributed across 10 lists, so that one participant would see only one version of each item.

The context sentences for critical items generated expectations regarding whether a precise amount of the relevant property (to which the adjective in the target sentence refers to) would be relevant (or not) for the interpretation of the target sentence. For example, In Figure 3, the 'strict context' leads the reader to expect a perfectly straight line, while the 'loose context' does not. The context sentences in the filler items did not highlight precision but instead provided information that was critical for selecting the correct picture. This was done to make sure that participants paid attention to the context sentence as well as to the pictures. As a result, we changed half of the filler items of Experiment 1 so that the target sentence would refer back to the context sentence in these cases, i.e., selecting the correct picture required remembering the context (e.g., the context says that Luca's friends are happy, and the target sentence says 'show me Luca's friends', see the OSF repository for a full list of critical and filler items).

### 3.3 Procedure

In each trial, participants first read the one-sentence context, and were instructed to press the SPACEBAR after reading it. The context sentence was displayed for a mandatory minimum of 2 seconds. Immediately after pressing the SPACEBAR the target sentence appeared together with the three pictures ('Target', 'Distractor' and a red X), analogously to Experiment 1. Participants took 12 minutes on average to complete the task.

### 3.4 Participants

As with Experiment 1, we conducted a power analysis via simulations based on pilot data (N = 200) to determine the number of participants needed for Experiment 2. The goal was to determine the number of participants needed to find an interaction effect between the levels 'Precise' and 'Imprecise 1' of the factor PICTURE TYPE and the two levels of the factor CONTEXT (i.e., 'loose' and 'strict'). Crucially, we assumed a more conservative effect size (Cohen's D = 0.2) for this interaction than the one found in the pilot study (Cohen's D = 0.32). The power analysis showed that with 360 participants we should have over 80% power to find a more conservative effect size than the one found in the pilot for said interaction. We therefore recruited 360 right-handed monolingual speakers of American English between the ages of 18-35 on the Prolific recruitment platform. Participants received the equivalent of 1 dollar as compensation for their participation in the study.

### 3.5 Predictions

Regarding picture selection, we had two main predictions. First, as in Experiment 1, we hypothesized that acceptance of the target picture as a referent for the absolute adjective would decrease as the depicted property moved further away from the precise standard. Second, we expected this to be mediated by the type of context participants read prior to the target picture. When the context elicits a 'loose' expectation, we expected that pictures in the 'Imprecise 1' condition (i.e., only slightly imprecise referents) would be accepted at comparable rates to pictures in the 'Precise' condition. We expected the remaining two imprecise conditions ('Imprecise 2' and 'Imprecise 3') to be accepted at lower rates, reflecting the effect of conceptual distance from the threshold found in Experiment 1. Conversely, we expected there to be very little tolerance for imprecision in the 'strict' context conditions, given how only the 'Precise'

condition is compatible with the contextual expectations. We therefore expected all imprecise conditions to be at chance (around 50% acceptance) or below.

The response times are of importance to the question of the encoded meaning of MSAAs. If precision is part of the encoded meaning of MSAAs and imprecise interpretations are necessarily reached with added effort only (even if contextual expectations of precision are low, as they are in the 'loose' contexts), acceptance times in the 'Precise' condition should be faster than in the 'Imprecise 1' condition regardless of context. This would be in line with a 'semantic' view. Alternatively, if the lexical representation of MSAAs is underspecified for precision (with no advantage for precise interpretations), response times should be wholly dependent on context, with the 'Precise' condition being faster than the 'Imprecise 1' following 'strict' contexts, and the 'Imprecise 1' condition delivering shorter response times than the 'Precise' condition following 'loose' contexts. This would be in line with an 'underspecification' view.

### 3.6 Analysis and results

#### 3.6.1 Picture selection rate.

Picture selection rates were analyzed similarly to Experiment 1. First, participants with an accuracy lower than 75% on the filler items were removed from further analysis. We also removed the data from 3 participants who completed the experiment despite not meeting the recruitment criteria. This reduced the total number of participants to 351. We then excluded all trials in which participants selected the 'Distractor' image as the referent, which amounted to removing 125 trials (3% of the data). We fitted a logistic regression model to the remaining data. The model had two predictors: PICTURE TYPE, and CONTEXT, both coded using treatment contrast coding, with the 'loose context - Precise' condition as the baseline. The dependent variable was the type of picture selected, with 0 signifying choosing the red X, and 1 signifying choosing the Target picture.

The model included random intercepts by items and by participants as well as random slopes for both factors and their interaction by items, a random slope term for CONTEXT by participants, and no random correlations between intercepts and slopes (R syntax: selected ~ PICTURE TYPE * CONTEXT+ (1+ CONTEXT ||participant) + (1+ PICTURE TYPE * CONTEXT ||Item)

The results can be visualized in panel A of Figure 4, and Table 3 shows the output of the model. Figure 4A shows that there was a step-wise decrease in acceptance rate of imprecise referents in the 'loose' contexts. However, as seen in Table 3, acceptance in the 'Loose - Imprecise 1' condition was significantly higher than in the 'Loose - Precise' condition (z-value = 2.23, $p < 0.05$). This suggests that the contextual manipulation led participants to accept a slightly imprecise interpretation at a higher rate than a precise interpretation of the maximum standard absolute adjectives when contextual expectations of precision were low. There was also an interaction between the 'top' two PICTURE TYPE conditions ('Precise' and 'Imprecise 1') and CONTEXT (z-value = 5.37, $p < 0.001$): the preference for imprecise interpretations in the 'loose' context was reversed in the 'strict' context. Further, differences in expectations for precision had an impact on all imprecise conditions: There were interactions between all levels of imprecision (relative to the 'precise' condition) and CONTEXT.

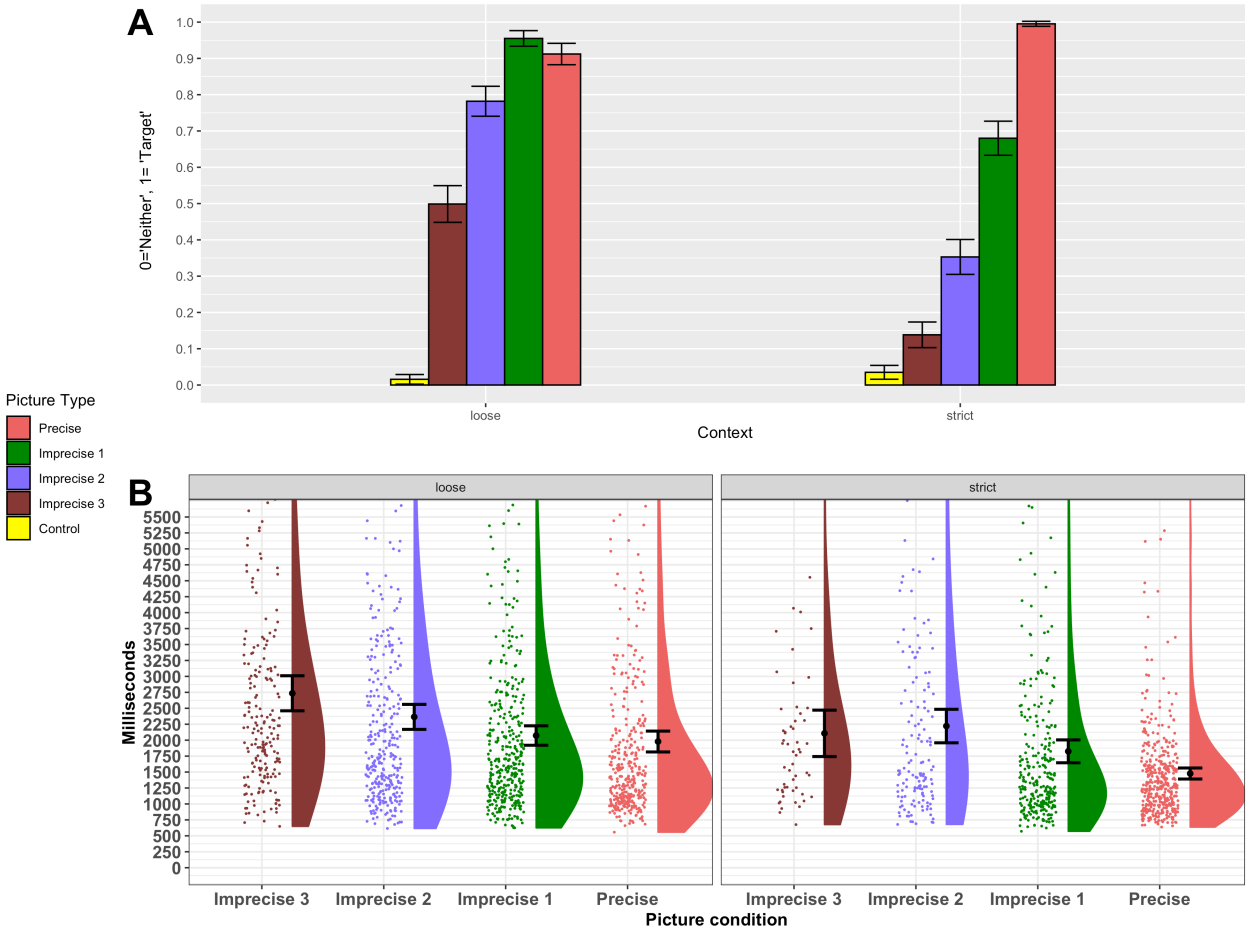


*Figure 4.* Target picture selection (panel A) and raw-reaction times (panel B) for Experiment 2. Error bars show confidence intervals.

Table 3:

*Logistic regression model of picture selection in Experiment 2, loose conditions*

| term | $\hat{\beta}$ | 95% CI | $z$ | $p$ |
|---|---|---|---|---|
| Loose - Precise vs. Loose – Imprecise 1 | 0.69 | [0.08, 1.30] | 2.23 | .026 |
| Loose - Precise vs. Loose – Imprecise 2 | -1.25 | [-1.71, -0.80] | -5.37 | < .001 |
| Loose - Precise vs. Loose – Imprecise 3 | -2.68 | [-3.22, -2.14] | -9.74 | < .001 |

| term | $\hat{\beta}$ | 95% CI | $z$ | $p$ |
|---|---|---|---|---|
| Loose - Precise vs. Loose - Control | -7.56 | [-8.89, -6.23] | -11.15 | < .001 |
| Loose - Precise vs. Strict - Precise | 3.51 | [1.95, 5.06] | 4.42 | < .001 |
| (Precise vs. Imprecise 1)*CONTEXT Interaction | -5.90 | [-7.52, -4.29] | -7.17 | < .001 |
| (Precise vs. Imprecise 2)*CONTEXT Interaction | -5.98 | [-7.55, -4.40] | -7.44 | < .001 |
| (Precise vs. Imprecise 3)*CONTEXT Interaction | -6.09 | [-7.76, -4.42] | -7.15 | < .001 |

*Note.* Both factors used a treatment contrast coding scheme.

Table 4:

*Linear regression model of target selection time in Experiment 2*

| term | $\hat{\beta}$ | 95% CI | $t$ | $df$ | $p$ |
|---|---|---|---|---|---|
| PICTURE TYPE | 0.22 | [0.15, 0.29] | 5.96 | 9.01 | < .001 |
| CONTEXT | 0.09 | [-0.01, 0.18] | 1.81 | 11.27 | .097 |
| PICTURE TYPE*CONTEXT Interaction | -0.24 | [-0.37, -0.11] | -3.63 | 9.72 | .005 |

*Note.* Model used a sum-contrast coding scheme. RTs were log-transformed.

Table 5:

*Pairwise comparisons of picture acceptance time in Experiment 2*

| contrast | *ΔM* | *95% CI* | *t* | *df* | *p* |
|---|---|---|---|---|---|
| Loose – Imprecise 1 vs. Loose - Precise | 0.10 | [-0.06, 0.25] | 1.90 | 9.15 | .291 |
| Loose – Imprecise 1 vs. Strict – Imprecise 1 | -0.03 | [-0.25, 0.18] | -0.49 | 10.51 | .959 |
| Loose – Imprecise 1 vs. Strict - Precise | 0.30 | [0.17, 0.44] | 6.82 | 9.59 | < .001 |
| Loose - Precise vs. Strict – Imprecise 1 | -0.13 | [-0.35, 0.09] | -1.79 | 10.35 | .332 |
| Loose - Precise vs. Strict - Precise | 0.21 | [0.07, 0.34] | 4.72 | 9.34 | .004 |
| Strict – Imprecise 1 vs. Strict - Precise | 0.34 | [0.19, 0.49] | 6.86 | 10.09 | < .001 |

*Note.* RTs were log-transformed. P-values corrected for multiple comparisons

**3.6.2 Picture selection time.** As in Experiment 1, we focused only on trials in which the target picture was selected. We first log-transformed the response times following the results of a box-cox test[1]. After this, we decided to deviate from our pre-registration in two ways. First, we did not exclude any outliers based on RTs being too short or too long (as in Experiment 1). Second, we decided to focus only on the 'Precise' and 'Imprecise 1' conditions to test our main prediction, namely whether there is an interaction between CONTEXT and these two levels of the factor PICTURE TYPE in picture acceptance times. We did this because there were so few

[1] Note that this is a different transformation from that used in Experiment 1, resulting from the differences in the distribution of residuals of the linear model (judging by the results of the box-cox test). The choice of transformation does not affect the overall pattern of results.

instances of accepting target pictures in the ‘strict context - Imprecise 3’ and ‘strict context – Imprecise 2’ conditions, that it was not possible to fit a model with an appropriate random effects structure. We therefore removed these few instances from the statistical analysis.

We fitted a mixed-effects linear model to the remaining data, using a sum-contrast coding for both factors. The random effects structure included random intercepts and slopes for both factors and their interaction by items and participants. We followed up on this model by performing pairwise comparisons using the R package emmeans (Lenth et al., 2019). We corrected for multiple comparisons using the Tukey adjustment. This type of analysis allowed us to both test for main effects and interactions, as well as to examine the comparisons between all individual conditions, which are both relevant for testing our hypotheses.

These results can be visualized in panel B of Figure 4. Table 4 shows the relevant output of the linear model and Table 5 shows the pairwise comparisons. There was a main effect of PICTURE TYPE ($t = 5.96$, $p < 0.001$) and an interaction between PICTURE TYPE and CONTEXT ($t = 3.63$, $p < 0.01$). The pairwise comparisons show that the ‘Strict - Precise’ was significantly faster than all other conditions (all p-values < .005). Most importantly, this condition was significantly faster than the ‘Loose - Imprecise 1’ ($t = 6.8$, $p < .001$), which represents a direct comparison of the two felicitous usages (an imprecise referent in a loose context vs. a precise referent in a strict context). It is also important to note that we failed to find a significant difference between the ‘Loose - Imprecise 1’ and the ‘Loose - Precise’ conditions ($t = 1.9$, $p = 0.29$), though the ‘Loose - Precise’ was numerically shorter.

### 3.7 Discussion

Experiment 2 examined the comprehension of sentences containing precise and imprecise usages of maximum standard absolute adjectives when these are embedded in simple discourse

contexts that differ in the type of elicited expectations of precision. The contextual manipulation had a global effect on both the acceptance rate and measures of comprehension effort, which led to a series of important findings. First, in terms of picture selection rate, contextual expectations clearly mediated the interpretation of the adjectives: Pictures in the 'Imprecise 1' condition were accepted significantly more often than pictures in the 'Precise' condition when an imprecise interpretation was compatible with the discourse context ('loose' context). Further, context improved the likelihood of acceptance of the visual referent for all degrees of imprecision, as evidenced by the interactions between all PICTURE TYPE conditions and CONTEXT reported in Table 3. Changes in the absolute percentages of tolerated imprecision are also noteworthy: Following the 'strict' context, the 'Imprecise 2' and 'Imprecise 3' pictures were accepted as referents 35% and 13.8% of the time, respectively. Following the 'loose' contexts, these rates went up to 78% and 49.8%. This surge in acceptance illustrates the extent to which comprehenders are flexible during the interpretation of sentences containing MSAAs. There are, however, limits to this flexibility: Pictures in the 'Imprecise 1' condition were accepted at a fairly high rate in the 'strict' contexts (~68%), suggesting that small degrees of imprecision might be highly conventionalized and can hardly be eliminated by a biasing discourse context.

Central to our investigation are the results regarding picture selection time. Here, we found an interaction in picture selection time between the top levels of PICTURE TYPE ('Imprecise 1' and 'Precise') and CONTEXT. This finding is in principle compatible with the 'underspecification view': Comprehension effort is mediated by contextual expectations, suggesting that both precise and imprecise interpretations of MSAAs require a similar contextual adjustment that results in added cost. However, there was also a main effect of PICTURE TYPE: the 'precise' condition was overall faster than the 'Imprecise 1' condition, regardless of context. This result is in line with the 'semantic' view: Despite contextual expectations, precise

interpretations were generally derived at a lower cost relative to imprecise interpretations. This raises the question of the source of the interaction between both factors. A visual inspection of Panel B of Figure 4 suggests that the difference is not driven by changes in acceptance time of imprecise interpretations in the 'loose' contexts, but by speedier acceptance times of precise readings in the 'strict' contexts. This is supported by the pairwise comparisons: There was no significant difference between 'Loose context – Imprecise 1' and 'Strict context – Imprecise 1' ($t = 0.4$, $p = .9$) or between the 'Strict context - Imprecise 1' and the 'Loose context - Precise' ($t = 1.78$, $p = .33$), while the 'Strict context - Precise' was faster than all other conditions.

The pattern is overall more compatible with a 'semantic' view of the lexical representation of MSAAs: Raising the contextual expectations of precision facilitated the comprehension of the critical sentences when these aligned with the semantics of the maximum standard absolute adjectives, while lowering the expectations resulted in additional pragmatic work to derive an imprecise interpretation, even when such an interpretation was highly supported by the context. This accounts for both the main effect of PICTURE TYPE, and the interaction between both factors. The results would be harder to explain from an 'underspecification' perspective, which would posit that the source of the interaction should be a contextual facilitation when expectations align with the visual referent.

## 4. Experiment 3

Experiments 1 and 2 suggest that deriving imprecise interpretations comes at a cost relative to deriving precise ones. However, picture selection time does not only reflect processing the MSAA, but also integrating it with the sentence, the visual and discourse context, and the selection of one from three possibilities. This makes it difficult to evaluate exactly where the processing differences between precise and imprecise interpretations originate, since they could

arise in any of these steps. We designed Experiment 3 as a self-paced reading task in order to remove the decision-making process from the measurement of processing time. This final experiment thus presents a stronger test of the two accounts discussed so far on the lexical meaning of MSAAs.

### 4.1 Materials, design, procedure and predictions

Experiment 3 was similar to Experiment 2 but had a different sequence of events in each trial and the structure of the target sentence was modified. Participants first read the discourse context ('loose' or 'strict', contexts were identical to those used in Experiment 2) and simultaneously saw a single image (the Target image in the conditions 'Precise','Imprecise 1', 'Imprecise 2', 'Imprecise 3' or 'Control' of Experiments 1 and 2, see Figure 1). This 'context slide' was displayed for a mandatory minimum of 2 seconds, and participants had to press the spacebar to continue. Then, participants read a critical sentence (containing an MSAA) in a self-paced, moving-window manner. After this, participants had to indicate whether the picture seen at the beginning matched the story as a whole (i.e., the context and the target sentence) by using the F (did not match) or J (matched) keys. An example of a target sentence is shown in (1) below, and the procedure is illustrated in Figure 5. Dashes indicate the boundaries of the sentence regions shown to participants at a time.

1. *"/The line / was straight / and Jasmine / drew another / after that/".*

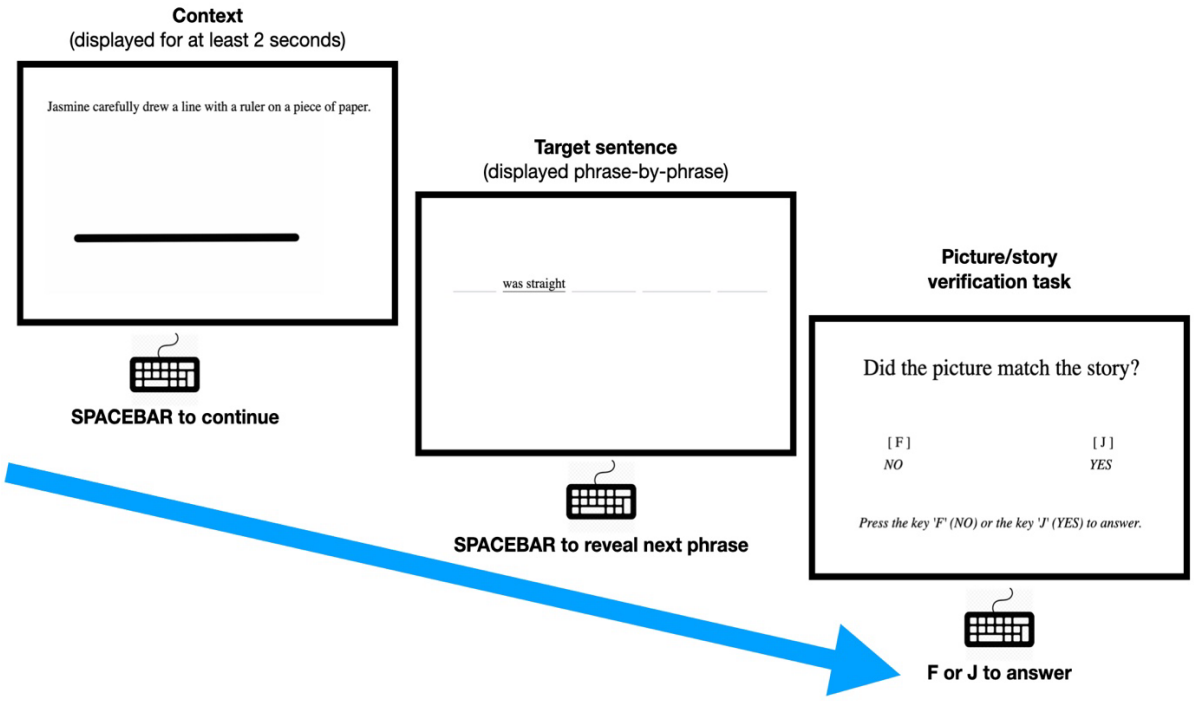


*Figure 5.* Example of the sequence of events in a single trial of Experiment 3

Participants pressed the SPACEBAR for the next region to appear on their screen. Sentences in critical items had five sentence regions, with the relevant MSAA always in the second region. We measured reading times of regions 2-5, i.e., from the point in which the adjective appears until the end of the sentence. We also adapted the 16 filler items of Experiment 2 in a similar way. Filler target sentences had between 4 and 6 sentence regions. In half of them (8 trials), the picture matched the story, while in the other half it did not. As in the previous experiments, participants had to score at least 75% accuracy in the filler trials (i.e., 12 trials) in order to be included in the analysis.

Our predictions relate to both the reading times and the story-picture verification task. Regarding reading times, if it is the case that precision is part of the encoded meaning of maximum standard absolute adjectives, sentences in the 'Strict context – Precise' condition

should be significantly faster to read overall than in the 'Loose context – Imprecise 1' condition. Further, there should either be no difference between 'Loose context – Imprecise 1' and 'Loose context - Precise' conditions (as was the case in Experiment 2), or the 'Loose context - Precise' condition should be faster than the 'Loose context – Imprecise 1' condition. This pattern of results would confirm the findings of Experiment 2 and suggest that despite contextual expectations, a precise interpretation of an absolute adjective is easier to derive than an imprecise interpretation. Alternatively, if it is the case that the encoded meaning of maximum standard absolute adjectives is not precise, there should be a 'full' interaction between context and picture type: The 'Loose context – Imprecise 1' condition should be faster than the 'Loose context - Precise', and there should be no difference between 'Loose context – Imprecise 1' and 'Strict context – Precise' conditions, i.e., between the two felicitous cases. Our predictions refer to the collapsed reading times of regions 2-5, to account for the possibility that effects might only become visible in one of the 'spill-over' regions after processing the adjective, as is commonly the case in self-paced reading paradigms. As a post-hoc measure, we also examined the reading times of individual regions to determine the earliest moment at which effects (if any) become visible.

Regarding the post-trial picture-story verification, we anticipated findings similar to those of Experiment 2. There, we found that picture selection of the target was mediated by contextual expectations: Acceptance of all imprecise conditions (Imprecise 1, 2, and 3) increased in the presence of 'loose' contexts relative to when they followed 'strict' contexts. In Experiment 3, we can examine whether this pattern holds when participants are merely asked to state whether the picture matched the story, instead of comparing the picture to the distractor. In sum, Experiment 3 presents a more stringent test of how contextual expectations mediate tolerance for imprecision by allowing us to examine how the sentences are processed independently of the potential effort

attributed to selecting a matching visual referent. If the findings of Experiment 2 represent how participants pragmatically adjust the meaning of maximum standard absolute adjectives to tolerate different degrees of imprecision, we should find a similar pattern of results in Experiment 3. Alternatively, if the results of Experiment 2 are influenced by the comparison process of the task, we might find a different pattern in Experiment 3.

### 4.2 Participants

Our goal was to match the number of participants of Experiment 2, which was 360. For this, we recruited 390 right-handed monolingual speakers of American English between the ages of 18-35 on the Prolific recruitment platform, assuming that some might not meet our inclusion criteria (i.e., achieve at least 75% accuracy on the post-trial story-picture verification task). The final number of participants was 371. Participants received 1 dollar as compensation for their participation in the study.

### 4.3 Analysis and results

#### 4.3.1 Picture-story verification.

We fitted a logistic regression model to analyze the picture-story verification data. For this, we used the full 5 × 2 design, as was done in Experiment 2 for the picture selection data. The dependent variable was the key pressed, with J coded as 1 and F coded as 0. The factors PICTURE TYPE and CONTEXT were coded with a treatment-contrast coding scheme (as was done in Experiment 2). The model had the 'Loose - Precise' condition as the baseline. The random effects structure included random intercepts and slopes for both factors and their interaction by items and by participants, suppressing the random correlations between intercepts and slopes (R syntax: selected ~ PICTURE TYPE * CONTEXT+ (1+ PICTURE TYPE * CONTEXT ||participant) + (1+ PICTURE TYPE * CONTEXT ||Item).

As seen in panel A of Figure 6, the results were similar to those of Experiment 2: there was a step-wise decrease in acceptance rate of imprecise referents in the ‘loose’ contexts. Further, as Table 6 shows, there was an interaction between the ‘top’ two PICTURE TYPE conditions (‘Precise’ and ‘Imprecise 1’) and CONTEXT (though, in contrast to Experiment 2, this interaction did not result in the ‘Imprecise 1’ condition having a higher acceptance rate than the ‘Precise’ condition in ‘loose’ contexts). In fact, there were interactions between all levels of imprecision relative to the ‘precise’ condition and CONTEXT, as was the case in Experiment 2.

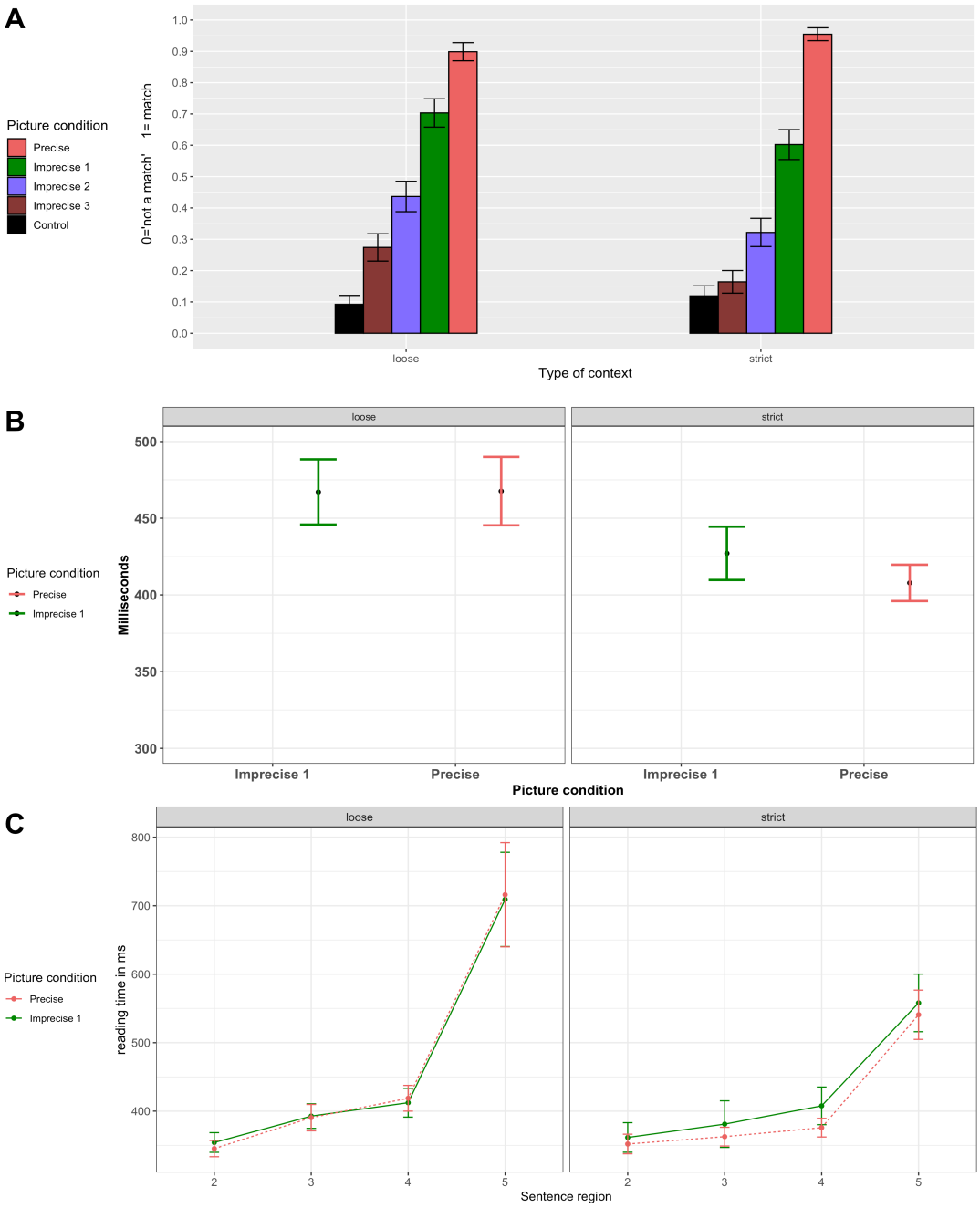


*Figure 6.* Picture-story verification accuracy (panel A) sentence raw-reading times (panel B) and raw-reading times by individual region (Panel C) for Experiment 3. Error bars show confidence intervals.

Table 6:

*Logistic regression model of picture-story verification in Experiment 3, loose conditions*

| term | $\hat{\beta}$ | 95% CI | $z$ | $p$ |
|---|---|---|---|---|
| Loose - Precise vs. Loose – Imprecise 1 | -1.36 | [-1.91, -0.80] | -4.79 | < .001 |
| Loose - Precise vs. Loose – Imprecise 2 | -2.78 | [-3.19, -2.38] | -13.33 | < .001 |
| Loose - Precise vs. Loose – Imprecise 3 | -3.65 | [-4.09, -3.21] | -16.09 | < .001 |
| Loose - Precise vs.  Loose - Control | -5.58 | [-6.54, -4.62] | -11.45 | < .001 |
| Loose - Precise vs. Strict - Control | 1.05 | [0.43, 1.68] | 3.31 | .001 |
| (Precise vs. Imprecise 1)*CONTEXT Interaction | -1.59 | [-2.42, -0.77] | -3.78 | < .001 |
| (Precise vs. Imprecise 2)*CONTEXT Interaction | -1.76 | [-2.43, -1.09] | -5.14 | < .001 |
| (Precise vs. Imprecise 3)*CONTEXT Interaction | -2.06 | [-2.87, -1.25] | -5.00 | < .001 |

*Note.* Both factors used a treatment contrast coding scheme.

Table 7:

*Linear regression model of reading times collapsed across regions in Experiment 3*

| term | $\hat{\beta}$ | 95% CI | $t$ | $df$ | $p$ |
|---|---|---|---|---|---|
| PICTURE TYPE | 0.05 | [0.02, 0.07] | 3.55 | 9.98 | .005 |
| CONTEXT | -0.07 | [-0.11, -0.03] | -3.43 | 11.17 | .006 |
| PICTURE TYPE*CONTEXT Interaction | 0.01 | [-0.03, 0.06] | 0.57 | 10.05 | .581 |

*Note.* Model used a sum-contrast coding scheme. RTs were log-transformed.

Table 8:

*Pairwise comparisons of reading times in Experiment 3*

| contrast | *ΔM* | *95% CI* | *z* | *p* |
|---|---|---|---|---|
| Loose – Imprecise 1 vs. Loose - Precise | 0.04 | [-0.01, 0.08] | 2.20 | .122 |
| Loose – Imprecise 1 vs. Strict - Imprecise 1 | 0.06 | [0.00, 0.13] | 2.45 | .067 |
| Loose – Imprecise 1 vs. Strict - Precise | 0.11 | [0.05, 0.18] | 4.84 | < .001 |
| Loose - Precise vs. Strict - Imprecise 1 | 0.02 | [-0.04, 0.09] | 0.97 | .766 |
| Loose - Precise vs. Strict - Precise | 0.08 | [0.02, 0.13] | 3.44 | .003 |
| Strict - Imprecise 1vs. Strict - Precise | 0.05 | [0.01, 0.10] | 2.86 | .022 |

*Note.* RTs were log-transformed. P-values corrected for multiple comparisons

**4.3.2 Reading time.** To analyze reading times, we first fitted a mixed-effects linear regression model to the log-reading times of the critical sentences, collapsing all regions. This model was fitted to the 'top' two levels of the factor PICTURE TYPE ('precise' and 'imprecise 1'), as was done in Experiment 2. This model included the factors PICTURE TYPE (levels: precise, imprecise 1) and CONTEXT (levels: loose and strict), together with their interaction. Both factors were sum-contrast coded. The model included random intercepts and slopes (for both factors and their interaction) by items and by participants, suppressing the random correlation between intercepts and slopes (R syntax: log_RT ~ PICTURE TYPE * CONTEXT+ (1+ PICTURE TYPE * CONTEXT ||participant) + (1+ PICTURE TYPE * CONTEXT ||Item).

Critically, only trials were included in the analysis for which participants derived precise interpretations in the precise conditions and imprecise interpretations in the imprecise conditions (judging by their responses in the picture-story verification task).

As per our pre-registration (and matching Experiment 2), we followed up on this model by performing pairwise comparisons of all four individual conditions, correcting for multiple comparisons using the Tukey adjustment.

The results, summarized in Table 7, show that there were main effects of PICTURE TYPE and CONTEXT, but no interaction. This suggests that precise interpretations (i.e., when the visual referent was closest to the maximum standard) were faster to read regardless of context, and that contexts where precision was not relevant for the interpretation resulted in overall longest reading times. The pairwise comparisons revealed that the 'Strict context – Imprecise 1' condition was slower than the 'Strict context – precise' condition, but we did not find a difference between 'Loose context – Imprecise 1' and 'Loose context – Precise' conditions, as seen in Table 8. To examine which sentence region was the first to show this pattern, we fitted the model to each individual region from region 2 to 5. The results showed that the two main effects first appear in region 3 (the first spill-over region) (effect of PICTURE TYPE: *t-value* = 2.09, $p < 0.05$; effect of CONTEXT: *t-value* = 3.44, $p < 0.005$). The overall pattern of results can be seen in Panels B and C of Figure 6.

### 4.4 Discussion

The results of Experiment 3 broadly confirm the findings of Experiment 2: In terms of picture-story verification, participants tolerated more imprecise interpretations when precision was not relevant for the interpretation ('loose' contexts) than when it was ('strict' contexts). The interaction between the top levels of the factor PICTURE TYPE and CONTEXT further suggests

that participants manage to adjust their precise and imprecise interpretations accordingly. However, the picture-story verification results must be interpreted carefully, seeing as the task itself was rather vague: We asked participants whether a picture 'matched' a story, instead of directly asking them to select the intended referent (as in Experiments 1 and 2).

Experiment 3 does allow for a better interpretation of the processing data relative to Experiments 1 and 2. Here, the reading times suggest that imprecise interpretations come at a cost across the board: It takes participants less time to read sentences eliciting precise readings vs. those eliciting imprecise ones. This effect is visible in the first spill-over region, i.e., the sentence region immediately after the MSAA was read. Together, the results from the picture-story verification and the self-paced reading tasks support our interpretation of the findings of Experiment 2 by suggesting that the advantage for precision is not exclusively related to the decision-making process of selecting a visual referent.

## 5. General discussion

The goal of the current work was to investigate how two distinct contextual factors - contextually-induced expectations of precision and a visual referent's conceptual deviation from a presumably precise maximal standard - work together to facilitate precise and imprecise interpretations of expressions containing maximum standard absolute adjectives such as *straight*. In Experiment 1 (which had no explicit discourse context), participants only tolerated the smallest deviation from the maximal standard ('Imprecise 1' condition, 90% acceptance rate), and did so at a cost: They selected precise referents at ceiling levels (100% acceptance) and took significantly less time to do so relative to all imprecise conditions. In Experiment 2, we showed that adding explicit discourse expectations (which manipulated the relevance of precision for an

interpretation) produced substantially different results. Here, when precision was made less relevant (i.e., 'loose' contexts), participants tolerated two out of three levels of deviation from the standard ('Imprecise 1' and 'Imprecise 2' provided an acceptance rate of 96% and 78%, respectively) and they accepted a slightly imprecise interpretation ('Imprecise 1') more often than a precise one ('Precise' condition, 91% acceptance). The acceptance rate of Experiment 1 thus seems to be somewhere in between the acceptance rates of the 'loose' and 'strict' contexts of Experiment 2. This suggests that, in the absence of context, participants in Experiment 1 assumed a moderately loose/strict context by default. Alternatively, there could have been a mix of precise and imprecise interpretations that resulted in the averages seen in Figure 2A. In either case, participants in Experiment 1 likely made ad hoc assumptions about the type of context needed in each trial. Given how those assumptions are outside of the experimenter's control, this speaks in favor of introducing specific contextual expectations in experiments that investigate context-sensitive expressions (such as MSAA are).

Furthermore, there was no significant difference in picture selection time between these two conditions. When precision was relevant for the interpretation ('strict' contexts), only the slightest degree of imprecision was tolerated ('Imprecise 1'), and it came at a cost (in terms of picture selection time) relative to deriving a precise interpretation. Experiment 3 additionally showed that in a self-paced reading task, participants read sentences with precise interpretations faster than the same sentences with imprecise interpretations, regardless of whether a high degree of precision was relevant for the overall interpretation.

Taken together, these results add substantially to those of previous studies on imprecision (e.g., Bambini et al., 2013; Syrett et al., 2010), while helping to draw a clearer picture concerning its comprehension. In Experiment 1, we replicated the findings of Syrett et al. (2010) and extended them to an adult population, to a richer list of maximum standard absolute adjectives and

different degrees of imprecision: without a discourse context, participants accept imprecise interpretations, and when they do, RTs are associated with a cost relative to accepting precise interpretations. Experiment 2 suggests that the relative delay in picture selection time found for imprecision in Experiment 1 can be mediated by contextual expectations such that Precise interpretations are faster to derive than imprecise ones following 'strict' contexts, but this effect is not present when following 'loose' contexts. Experiment 3 additionally showed that processing costs related to comprehending imprecision arise naturally during self-paced reading. This suggests that the findings of Experiment 2 were not exclusively caused by the decision-making process of selecting a target picture.

In Experiment 2, we found an interaction between context and type of visual referent on picture selection time. However, this interaction was not fully in line with what would be predicted by an 'underspecification' view: It was not the case that imprecise interpretations were facilitated by context (though 'loose' contexts significantly affected their acceptability). Instead, it was precise readings that were easier to understand when expectations of precision were high. This interaction was absent in Experiment 3, which arguably presented a stronger test of the accounts. In general, there was never an instance of imprecise interpretations being less costly in terms of comprehension time relative to precise interpretations. Precise readings were less costly than imprecise ones both in the absence of context (Experiment 1), with high expectations of precision (Experiment 2, 'strict' context), and in the absence of a picture-selection task (Experiment 3).

This processing asymmetry likely reflects how the communicated imprecise usage of an MSAA is asymmetrically entailed by the precise semantic meaning of it (see Lauer, 2012). This can also be taken to support a 'semantic' view on the encoded meaning of MSAAs: Though generally malleable by contextual expectations, comprehension effort is contingent on the proximity to the

maximal, precise standard, biasing towards a precise interpretation of an MSAA.

An important contribution of the present work pertains to the differences in comprehension elicited by the different degrees of visual deviation of a referent from the maximal standard (i.e., the factor PICTURE TYPE). First, they can help make sense of discrepancies between previous studies. For example, Syrett et al. (2010) might have found that adults were not very tolerant of imprecision because the pictures they used were in a range comparable to our 'Imprecise 2' or 'Imprecise 3' conditions, whereas the pictures used by Leffel et al. (2016) (who found that imprecise visual referents generated anticipatory contrastive inference effects) might have been closer to our 'Imprecise 1' images. Second, they reveal limits to what can be considered as *conventional imprecision*: Pictures in the 'Imprecise 1' condition were consistently tolerated, even when ostensively at odds with the contextual information of Experiments 2 and 3 ('strict' contexts).[2] This suggests that such a small degree of imprecision – or 'pragmatic slack' - is

---

[2] This would be in line with a third type of theoretical account, which is currently being explored within the relevance-theoretic framework, most notably in the work of Robyn Carston (see, e.g., Carston, 2019). Carston proposes that the meaning(s) of substantive words (nouns, verbs, adjectives; words that encode concepts) are stored as 'polysemous complexes', that is, an evolving set of interrelated senses/concepts which are conventionalized to some appreciable degree and are input to pragmatic processes that form occasion-specific senses. For MSAAs, it could be that the precise sense is typically the most conventionalized among the different senses in the polysemy complex, and therefore more accessible, but other degrees of imprecision could also be partly conventionalized as part of the polysemy complex.

routinely tolerated by language comprehenders regardless of contextual expectations (though it comes with an additional processing cost). The 'Imprecise 2' condition, on the other hand, represents a more radical case of ad hoc pragmatic meaning adjustment. As such, it is extremely context sensitive: the acceptance rate in Experiment 2 ranges from 35% following the 'strict' context to 79% following the 'loose' context (on average). Notably, this condition always displayed longer picture selection times in Experiment 2 relative to the precise and the 'Imprecise 1' condition, suggesting that constructing an interpretation that deviates to a large extent from the semantically encoded sense induces an additional processing cost.

It is possible to draw parallels from our experiments to those conducted on the processing of numerals, specifically to the ones reported in Helena Aparicio's dissertation (2018, ch. 5). Aparicio conducted a series of experiments on the processing of precise and imprecise readings of round numbers. The starting point for her investigation is the existing claim that round numbers are preferentially interpreted imprecisely (Krifka, 2002, 2007; Bastiaanse, 2011, i.a.). If this were the case, it would be expected for the processing of imprecise interpretations of round numerals to be less costly than that of precise interpretations. Critically, Aparicio's findings did not support this hypothesis and are more in line with the opposite claim, namely that precise readings are less costly than imprecise ones. Our findings are thus in line with Aparicio's. This further strengthens the view that imprecision is a pragmatic adjustment that brings about a cost. It also gives credence to the claim that imprecision in the adjectival and numeral domains may be of the same nature, as put forth by Aparicio (2018, p. 96).

### 5.1 Limitations and future directions

A limitation of the current study is the relatively ambiguous measure of picture selection time used in Experiments 1 and 2, given how it conflates processing the adjective and deciding between possible visual referents. This was overcome in Experiment 3, where we find a similar pattern of results as in Experiment 2 using only self-paced reading. However, self-paced reading brings with it other limitations. While Experiment 3 shows a processing advantage for precision across all sentence regions, the first region to on its own show an effect is the one after the MSAA (the so-called 'spill-over' region). This leaves the question open as to whether this effect reflects later stages in processing (because it does not appear immediately upon encountering the adjective), or whether it reflects a lag in the signal (because self-paced reading is a coarse measure of processing). It is therefore possible that using more time-sensitive measures (such as eye-tracking or EEG, for example) that tap into the earliest comprehension stages (i.e., as soon as the MSAA is integrated with the utterance), one would find no differences between precise and imprecise readings, in line with the 'underspecification view', as stated by Frisson (2009). While we agree that a stronger test of the 'underspecification view' requires a more fine-grained measure, the current results still bear relevance to this debate. First, we adopted the design of the seminal study by Syrett et al. (2010), which allows us to more readily compare our results to theirs (Experiments 1 and 2). Second (and though we remain agnostic regarding the precise locus of the effect), it is still the case that throughout the entire comprehension process, precise interpretations were less effortful than imprecise ones, even when processing was measured without an explicit decision making task (Experiment 3). This seems largely incompatible with underspecification views, which posit that speakers zero in on the contextually appropriate sense after first activating the underspecified meaning. One possible objection is that our 'loose' contexts were not sufficiently biasing towards the imprecise interpretations, and that more

naturalistic and/or elaborate contexts raising low expectations of precision might have reduced or even eliminated the additional processing effort found for imprecise compared to precise interpretations in our study. While we cannot completely rule out this possibility, all our 'loose' contexts were considered not to generate expectations of precision in our norming study, although the ratings were not as unequivocal as for the strict contexts.

Further, it is important to note that it is outside of the scope of the current investigation to spell out exactly what the lexical entries of MSAA would look like according to an 'underspecification' account. It is possible that such an account would see for the lexical representation of absolute adjectives to be like that of relative adjectives, as posited by Lassiter and Goodman (2013). In that account, the difference in interpretation between both adjective types is not given by the semantics of the adjectives, but by the prior expectations that interpreters have regarding the statistical properties of the relevant comparison class. If prior expectations are contextually shifted, imprecise interpretations could be preferred to precise interpretations. This is not entirely distinct from what we observe in our data, where we find that shifting the prior expectations of precision results in changes in the interpretation of the MSAAs. However, from a psycholinguistic perspective, it would be ideal for such an 'underspecification' account to make predictions regarding how processing cost of imprecision is mediated by context, which Lassiter and Goodman (2013) do not make. In the current manuscript we are concerned with formulating what such an account would predict in terms of processing, and not with the explication of the account per se.

A loose thread presents itself from a developmental perspective. In the study by Syrett et al. (2010), it was consistently found that children were more tolerant of imprecision compared to adults. Syrett et al. (2010) speculate that adults might have been more tolerant of imprecision had

the materials been closer to the precise standard. This intuition finds some support in the present Experiments, where even in the absence of context our adult participants accepted imprecise interpretations quite frequently (of type 1). But how would children perform in our task? We could assume that they would likely show remarkable tolerance for imprecision in Experiment 1. But would they be able to modulate this according to contextually-raised expectations of precision? We leave this question open for future studies.

## 6. Conclusion

The present study investigated how two contextual factors – distance from a precise maximum standard and expectations of precision raised by a discourse context – influence both the tolerance for imprecision and the comprehension effort associated with deriving precise and imprecise interpretations of maximum standard absolute adjectives. Our findings bridge the gap between previous studies by showing how a referent's visual distance from the precise standard mediates both comprehension rate and effort. More importantly, our results are in line with a 'semantic view' of maximum standard absolute adjectives, according to which precision (defined as the requirement for an argument to possess the maximal degree of a relevant property) is part of the encoded meaning of these adjectives, and speak against an 'underspecification view' in which precision comes about as a result of contextual adjustment.

## 7. Data availability statement

All data, analysis scripts and pre-registrations for the experiments in this article are available on the project's OSF repository: https://osf.io/6gwdq/

## 8. Competing interests statement

The authors have no completing interests to declare.

## 9. Ethics and consent statement

All experiments reported in this manuscript received ethical approval from Sikt - Norwegian Agency for Shared Services in Education and Research, ref. no. 596365.

## 10. Author contributions

The authors made the following contributions. Camilo R. Ronderos: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing, Formal Analysis, Visualization, Data curation, Resources, Methodology, Investigation, Project administration; Ira Noveck: Conceptualization, Writing - Review & Editing, Supervision; Ingrid Lossius Falkum: Conceptualization, Writing - Review & Editing, Supervision, Funding acquisition, Project administration.

## 11. Acknowledgements

We would like to thank the editor Petra Schumacher, the anonymous reviewers and Steven Verheyen for providing detailed and helpful feedback. We are also grateful to Nicole Gotzner for feedback on a previous version of this manuscript and to Henriette Johansen for assistance in creating the materials. We would also like to thank the members of the DEVCOM lab, members of the SPA lab and audiences of ELM 1, HSP 1 and Xprag.it 2020 for their comments and suggestions. This article is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program under grant agreement no. 853211 (ERC Starting Grant 2019), awarded to Ingrid Lossius Falkum.

**References**

Aparicio, H. (2018). *Processing context-sensitive expressions: the case of gradable adjectives and numerals*. Doctoral Dissertation, University of Chicago.

Aparicio, H., Xiang, M., & Kennedy, C. (2016). Processing gradable adjectives in context: A visual world study. *Semantics and Linguistic Theory*, *25*, 413–432.

Aust, F., & Barth, M. (2017). *Papaja: Prepare reproducible APA journal articles with R Markdown. R package version 0.1. 0.9997*.

Bambini, V., Ghio, M., Moro, A., & Schumacher, P. B. (2013). Differentiating among pragmatic uses of words through timed sensicality judgments. *Frontiers in Psychology*, *4*, 938.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bastiaanse, H. (2011). The rationality of round interpretation. In R. Nouwen, R. van Rooij, U. Sauerland, & H.-C. Schmitz (Eds.), *Vagueness in communication. International workshop, ViC 2009 held as part of ESSLLI 2009*, Bordeaux, France, July 20-24, 2009. Revised Selected Papers, (volume 6517 of Lecture Notes in Computer Science, pp. 37–50). Springer

Bates, D., Sarkar, D., Bates, M. D., & Matrix, L. (2007). The Lme4 package. *R Package Version*, *2*(1), 74.

Beltrama, A., & Schwarz, F. (2022). Imprecision, personae, and pragmatic reasoning. *Semantics and Linguistic Theory*, *31*, 122–144.

Beretta, A., Fiorentino, R., & Poeppel, D. (2005). The effects of homonymy and polysemy on lexical access: An MEG study. *Cognitive Brain Research*, *24*(1), 57–65.

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 211–252.

Burnett, H. (2014). A delineation solution to the puzzles of absolute adjectives. *Linguistics and Philosophy*, *37*(1), 1–39.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Carston, R. (2002). *Thoughts and utterances: The pragmatics of explicit* communication. Oxford: Blackwell.

Carston, R. (2010). XIII-Metaphor: Ad Hoc Concepts, Literal Meaning and Mental Images. *Proceedings of the Aristotelian Society (Hardback)*, *110*(3pt3), 295–321. https://doi.org/10.1111/j.1467-9264.2010.00288.x

Carston, R. (2013). Word meaning, what is said and explicature. In P. Carlo & D. Filippo (Eds.), *What is said and what is not* (pp. 175–204). CSLI Publications.

Carston, R. (2016). The heterogeneity of procedural meaning. *Lingua*, *175*, 154–166.

Carston, R. (2019). Ad hoc concepts, polysemy and the lexicon. In K. Scott, B. Clark, & R. Carston (Eds.), *Relevance, pragmatics and interpretation* (pp. 150-162).

Cambridge University Press. https://doi:10.1017/9781108290593.014

Dehaene, S., & Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition*, *43*(1), 1–29. https://doi.org/10.1016/0010-0277(92)90030-L

Falkum, I. L., & Vicente, A. (2015). *Polysemy: Current perspectives and approaches*.

Frisson, S. (2009). Semantic underspecification in language processing. *Language and Linguistics Compass*, *3*(1), 111–127. https://doi.org/10.1111/j.1749-818X.2008.00104.x

Frisson, S., & Pickering, M. J. (2001). Obtaining a figurative interpretation of a word: Support for underspecification. In *Metaphor and symbol* (pp. 149–171). Psychology Press.

Frisson, S., & Pickering, M. J. (2007). The processing of familiar and novel senses of a word: Why reading Dickens is easy but reading Needham can be hard. *Language and Cognitive Processes*, *22*(4), 595–613.

Gibbs, R., & Bryant, G. A. (2008). Striving for optimal relevance when answering questions. *Cognition*, *106*(1), 345–369.

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Højsgaard, S. (2012). The doBy package. *R Package Version*, *4*(3).

Hope, R. M. (2013). Rmisc: Ryan miscellaneous. *R Package Version*, *1*(5).

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, 640–648.

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, *30*(1), 1–45.

Kennedy, C., & McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language*, 345–381.

Krifka, M. (2002). Be brief and vague! And how Bidirectional Optimality Theory allows for verbosity and precision. In D. Restle & D. Zaefferer (Ed.), *Sounds and Systems: Studies in Structure and Change. A Festschrift for Theo Vennemann* (pp. 439-458). Berlin, New York: De Gruyter Mouton. https://doi.org/10.1515/9783110894653.439

Krifka, M. (2007). Approximate interpretations of number words: A case of strategic communication. In G. Bouma, I. Kramer, & J. Zwarts (Eds.), *Cognitive Foundations of Interpretation* (pp. 111–126). Humboldt-Universität zu Berlin, Philosophische Fakultät II.

Lasersohn, P. (1999). Pragmatic halos. *Language*, 522–551.

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. *Semantics and Linguistic Theory*, *23*, 587–610.

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

Lee, C., & Kurumada, C. (2021). Learning Maximum Absolute Meaning Through Reasoning About Speaker Intentions. *Language Learning*, *71*(2), 326–368. https://doi.org/10.1111/lang.12439

Leffel, T., Xiang, M., & Kennedy, C. (2016). Imprecision is pragmatic: Evidence from referential processing. *Semantics and Linguistic Theory*, *26*, 836–854.

Leffel, T., Xiang, M., & Kennedy, C. (2017). *Interpreting gradable adjectives in context: Domain distribution vs. Scalar representation*. Unpublished Manuscript.

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2019). *Package "emmeans"*.

Lewis, D. (1979). Scorekeeping in a language game. In *Semantics from different points of view* (pp. 172–187). Springer.

Miller, J. (1991). Short Report: Reaction Time Analysis with Outlier Exclusion: Bias Varies with Sample Size. *The Quarterly Journal of Experimental Psychology Section A*, *43*(4), 907–912. https://doi.org/10.1080/14640749108400962

Müller, K. (2017). *Here: A simpler way to find your files* [Manual].

Pickering, M. J., & Frisson, S. (2001). Processing ambiguous verbs: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(2), 556.

Pietroski, P. (2005). Meaning before truth. *Contextualism in Philosophy: Knowledge, Meaning, and Truth*, *255*, 302.

Potts, C. (2012). Goal-driven answers in the Cards dialogue corpus. *Proceedings of the 30th West Coast Conference on Formal Linguistics*, 1–20. Cascadilla Proceedings Project.

Pylkkänen, L., Llin'as, R., & Murphy, G. L. (2006). The representation of polysemy: MEG evidence. *Journal of Cognitive Neuroscience*, *18*(1), 97–109.

R Core Team. (2020). *R: A language and environment for statistical computing* [Manual]. R Foundation for Statistical Computing.

Rabagliati, H., & Snedeker, J. (2013). The truth about chickens and bats: Ambiguity avoidance distinguishes types of polysemy. *Psychological Science*, *24*(7), 1354–1360.

Recanati, F. (2010). *Truth-conditional pragmatics*. Clarendon Press.

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D., & Ripley, M. B. (2013). Package "mass." *Cran r*, *538*, 113–120.

RStudio Team. (2020). *RStudio: Integrated development environment for r* [Manual]. RStudio, PBC.

Ruhl, C. (1989). *On monosemy: A study in linguistic semantics*. SUNY Press.

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2015). *Package "afex"*.

Solt, S., Cummins, C., & Palmović, M. (2017). The preference for approximation. *International Review of Pragmatics*, *9*(2), 248–268. https://doi.org/10.1163/18773109-00901010

Sperber, D., & Wilson, D. (1986/1995). *Relevance: Communication and Cognition*. Blackwell.

Sperber, D., & Wilson, D. (2008). A deflationary account of metaphors. In R. Gibbs (Ed.), *The Cambridge Handbook of Metaphor and Thought* (pp. 84–105). Cambridge University Press. https://doi.org/10.1017/CBO9780511816802.007

Syrett, K., Kennedy, C., & Lidz, J. (2010). Meaning and context in children's understanding of gradable adjectives. *Journal of Semantics*, *27*(1), 1–35.

Van Der Henst, J., Carles, L., & Sperber, D. (2002). Truthfulness and relevance in telling the time. *Mind & Language*, *17*(5), 457–466.

Vicente, A. (2018). Polysemy and word meaning: An account of lexical meaning for different kinds of content words. *Philosophical Studies*, *175*(4), 947–968. https://doi.org/10.1007/s11098-017-0900-y

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *Dplyr: A grammar of data manipulation* [Manual].

Wilson, D., & Carston, R. (2006). Metaphor, Relevance and the 'Emergent Property' Issue. *Mind & Language*, *21*(3), 404–433. https://doi.org/10.1111/j.1468-0017.2006.00284.x

Wilson, D., & Carston, R. (2007). A unitary approach to Lexical Pragmatics: Relevance, Inference and Ad hoc concepts. In N. Burton-Roberts (Ed.), *Pragmatics* (pp. 230–259). London: Palgrave Macmillan UK. https://doi.org/10.1057/978-1-349-73908-0_12

Wilson, D., & Sperber, D. (2012). *Meaning and Relevance*. Cambridge University Press. https://doi.org/10.1017/CBO9781139028370

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