Imprecision is pragmatic: Evidence from referential processing

Ciyang Qing

Replication of Leffel, Xiang, & Kennedy, 2016

1 Introduction

This paper reports on a replication of (part of) Leffel et al., 2016. The issue under investigation concerns the semantics and pragmatics of *gradable adjectives*, e.g., *tall, big, full,* and *straight*. These adjectives have similar meanings in their comparative and superlative forms, and according to one prominent analysis, they denote scales (measure functions), which are functions that take an individual and return a degree (e.g., Kennedy, 2007).

(1) $[\![\text{tall}]\!] = \lambda x.\textbf{height}(x) \qquad [\![\text{full}]\!] = \lambda x.\textbf{fullness}(x)$

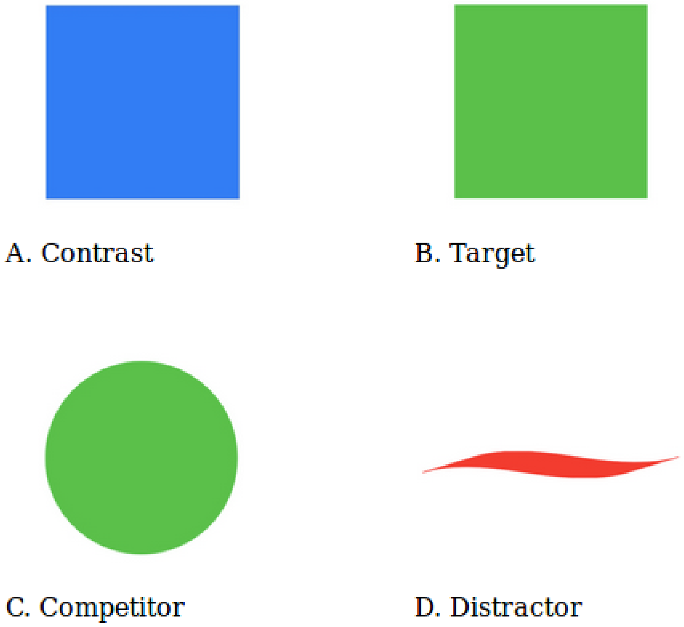
Under this approach, the positive form of a gradable adjective, e.g., *tall* in *John is tall*, is analysed as follows.

(2) $[\![\text{x is tall}]\!] = \textbf{height}(x) \geq \theta \qquad [\![\text{x is full}]\!] = \textbf{fullness}(x) \geq \theta$

However, it seems that there are (at least) two classes gradable adjectives wrt the property of the *threshold/standard of comparison* θ. On the one hand, *relative adjectives* such as *tall* and *big* have *vague* thresholds, e.g., people can be quite uncertain about how tall counts as tall. On the other hand, *(maximum) absolute adjectives* such as *full* and *straight* do not seem to be vague and the threshold θ is always the maximum degree on the scale, e.g., something is full iff it is maximally full.[1] This dichotomy is complicated by the fact that in reality, people do not always use language in a perfectly precise manner. For example, people may say that a glass is full even though there is still some room left. This suggests that *imprecision* as well as vagueness play a role in people's use of positive forms of gradable adjectives.

Leffel et al. (2016) considers two competing theories of the semantics and pragmatics of gradable adjectives. According to what they call the *semantic hypothesis* (**HS**), the threshold θ is always contextually determined based on the statistical distribution of the degrees in the comparison class (Lassiter & Goodman, 2013, 2015). Under this view, an absolute adjective has a threshold whose probability distribution is concentrated in somewhere close to the maximum degree, while a relative adjective has a threshold whose probability distribution has higher variance. An alternative approach, which (Leffel et al., 2016) call the *pragmatic hypothesis* (**HP**), holds that only relative adjectives have unspecified thresholds that need to be contextually determined. An absolute adjective has a lexicalized threshold that is the

[1] There is another class of absolute adjectives such as *bent* and *dirty*, the threshold θ of which is always the minimum degree of the scale. Following Leffel et al. (2016), I will not further discuss them in this paper.



A. Contrast

B. Target

C. Competitor

D. Distractor

Figure 1: Referential Contrast Effect (RCE) originally reported by Sedivy et al. (1999). When hearing *click on the green . . .* , participants are more likely to look at the target (the green square) if the contrast (the blue square) is present due to Gricean pragmatic reasoning. The example is from Leffel et al. (2016).

maximum degree of the scale, and an additional pragmatic mechanism is needed to account for imprecision (e.g., Kennedy & McNally, 2005; Kennedy, 2007).

It is not easy to tease the two theories apart in terms of empirical predictions, since two approaches can both capture people's truth/felicity judgments. As a result, Leffel et al. (2016) proposes that we test them based on how people process these two classes of gradable adjectives. Concretely, they consider the *referential contrast effect* (Sedivy, Tanenhaus, Chambers, & Carlson, 1999). An example is illustrated in Fig. 1. When hearing *click on the green . . .* , participants are more likely to look at the target (the green square) if the contrast (the blue square) is present due to Gricean pragmatic reasoning: if the target were the green circle, the speaker could have said *click on the circle*, which would be shorter but still unambiguous. This is called the *Referential Contrast Effect* (RCE).

According to Leffel et al. (2016), the two hypotheses of gradable adjectives make different predictions about the RCE for the two classes of adjectives. Concretely, the semantic hypothesis **HS** predicts that the effects would be the same for both classes, since they undergo the same mechanism to determine the threshold θ. On the other hand, the pragmatic hypothesis **HP** predicts that the effects would be different, because vagueness and imprecision involve different mechanisms.

2 Methods

2.1 Participants

4 native English speakers participated in the eyetracking experiment (2 female and 2 male). They are Linguistics/SymSys graduate students. 2 participants speak the British variety and the other 2 American East Coast.

2.2 Stimuli

Leffel et al. (2016) kindly provided the stimuli from the original experiment, which were used in the replication. There are 120 scenes, each consisting of 4 objects: Target, Competitor, Contrast/Distractor, Distractor. In particular, there are 30 critical scenes, 10 of which are associated with (maximum) absolute adjectives and the other 20 are associated with relative adjectives. Half of the scenes are in the Contrast condition, where there is a Contrast object that shares the head noun with the Target obejct but has a smaller degree wrt to the adjective. The other half of the scenes are in the NoContrast condition, where there are 2 Distractors and no Contrast object. Leffel et al. (2016) also manipulated the relationship between the degrees of the Target and Competitor, but for simplicity, in this replication, the degree of A-ness of the Target is always smaller than that of the Competitor (T<C).

2.3 Procedure

After the instruction and calibration, participants in each trial first clicked on a "+" symbol in the center of the screen (to ensure that the mouse cursor originated from the center) and fixated on it. After 1s of fixation, the 4 images in the scene of the current trial appeared and the participants were encouraged to familiarize with them. After 2s of exploration, the audio instruction started, telling the participant to click on the target object. The instruction is in the form *click on the (adj) noun*. (In some filler conditions there were no adjectives). After the participant clicked on an image, it is surrounded by a black border to mark the choice and then the whole scene disappeared after 600ms and the next trial began.

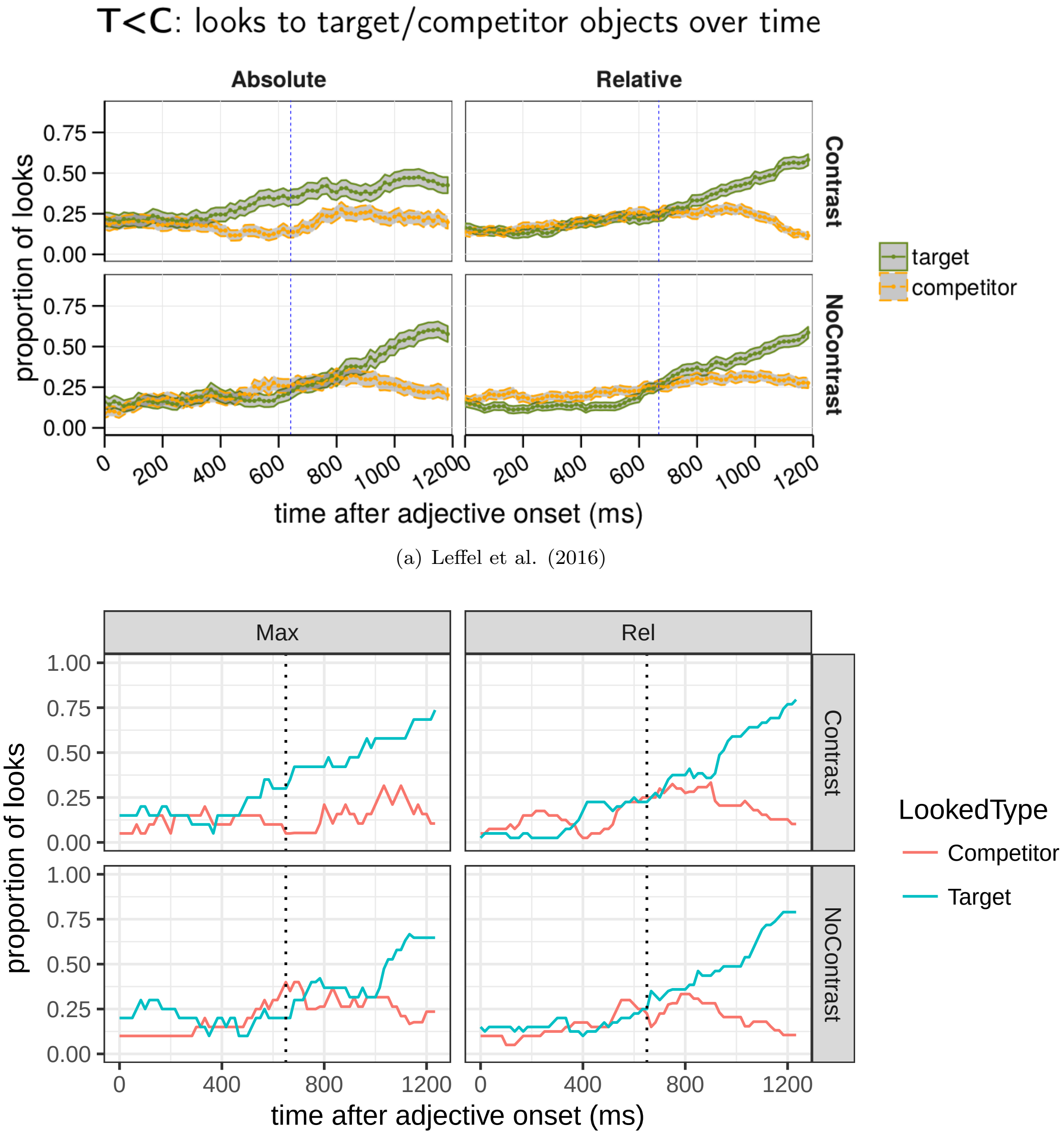
The sampling rate of the eye-tracker was 300Hz, but the raw data were later down-sampled to 60Hz in the analysis, to make it comparable with Leffel et al., 2016.

3 Results

3.1 Qualitative patterns

First, I compare the qualitative patterns of the data. Fig. 2 shows the original result in Leffel et al., 2016 and the replication and Fig. 3 shows the responses by each participant. We can see that even though there is a lot of variation in participants' responses, overall the qualitative trend seems comparable to what (Leffel et al., 2016) reports. Crucially, it seems that before the information about the head noun is available (650ms, which is 200ms after the average onset of noun: 450ms after the onset of the adjective), a preference for the target is observed only for those scenes with absolute adjectives and a contrast object.

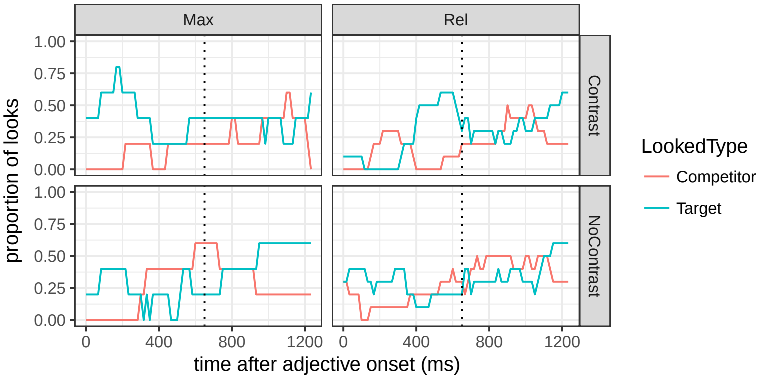
Therefore, the qualitative patterns reported by Leffel et al. (2016) seems to be replicated. Of course, given that there were only 4 participants in the replication, this conclusion is very tentative and should not been seen as very reliable.



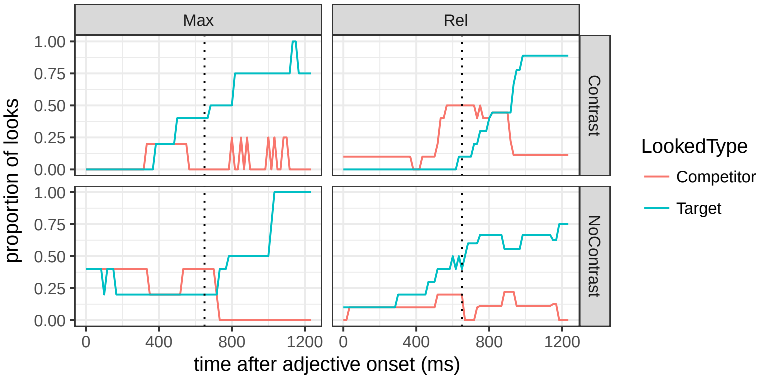
(a) Leffel et al. (2016)

(b) Replication ($n = 4$)

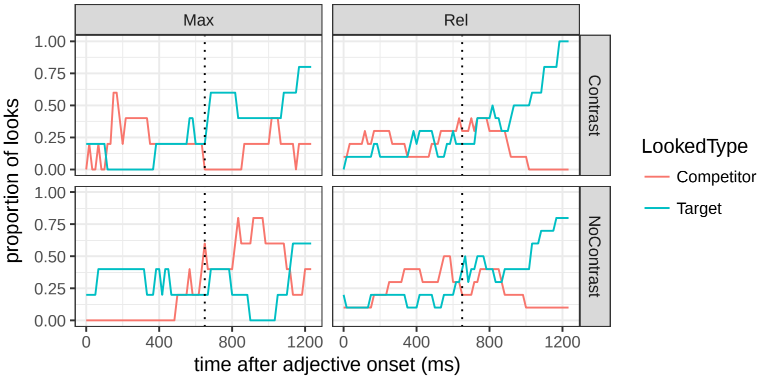
Figure 2: Proportions of looks after adjective onset in the original experiment and the replication. The information of the head noun is not available until 650ms (the dotted line).



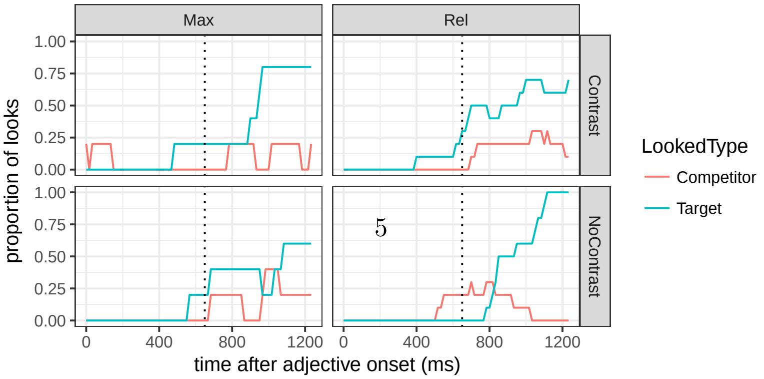
(a) Participant 1



(b) Participant 2



(c) Participant 3



(d) Participant 4

3.2 Statistical analysis

Again, given that there were only 4 participants, there are simply not enough data to fit a reliable statistical model. Therefore, the analysis below is mainly for illustrative purposes.

Following Leffel et al. (2016), I consider the proportions of looks on the Target vs the Competitor on several time windows of length 150ms. Starting with 200ms after the adjective onset, which is when the adjective information becomes available, I consider 5 consecutive windows: 200–350ms (window 0), 350–500ms (window 1), 500–650ms (window 2), 650–800ms (window 3), and 800–950ms (window 4). Leffel et al. (2016) only consider window 1 to 4 and they use ANOVA in the statistical analysis, which is inappropriate for proportion data. Therefore, I use a mixed-effect logistic regression instead, with AdjectiveType (Max vs Rel) and Condition (Contrast vs NoContrast) as fixed effects and participant and scene as random effects.

In most cases, the effects are not significant. This is expected since there are so few participants. Also, taking into account the random effects makes a difference (to the extent that the estimates are reliable given so few data).

As an example, Fig. 4 shows the results for Window 2 (500–650ms) and Window 3 (650–800ms).

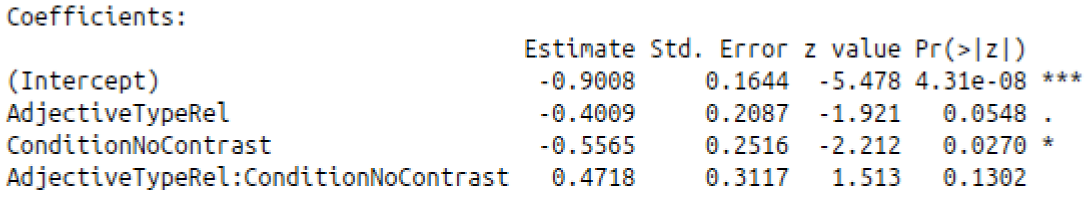
In Window 2, according to the mixed-effect model, there is no interaction between AdjectiveType and Condition. For both types of adjectives, when there is no contrast object present, the target is less likely to be looked at ($b_{\max} = -.856$, $b_{\max} = -.687$, $p < .05$ in both cases). In other words, the RCE has been observed for both absolute and relative adjectives in Window 2.

In Window 3, according to the mixed-effect model, there is a significant interaction between AdjectiveType and Condition ($b = -1.1, p < .01$). Relative adjectives lead to stronger RCE effects. Note that including random effects makes a difference, since a logistic regression model with only fixed effects has a marginally significant interaction term that suggests the opposite direction ($b = .45, p = .099$).

Again, as noted before, these results are likely not very reliable given the small number of participants, but they do suggest that participant and scene variability can be potential confounds. Indeed, from participants' click data and general feedback after the experiment, it appears that there are quite a few trials that are confusing to the participants. One such example is shown in Fig. 5. The participants were asked to *click on the straight line*. In this case, the Target is intended to be the slightly bent line (object B). However, some participants clicked on the perfectly straight arrow, which is understandable because there is indeed a perfectly straight line as a part of the arrow! Given that there are quite a few such items in the stimuli, interpreting the results become more difficult.

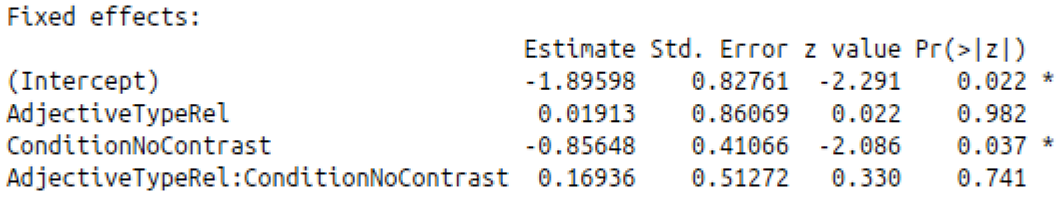
4 Discussion and conclusion

The replication of Leffel et al., 2016 is very inconclusive due to the small number of participants but still quite informative. On the one hand, the qualitative patterns of the original results seem to be replicated, suggesting that the findings are empirically robust. On the other hand, a careful examination of the stimuli and participants' click responses suggests that the original stimuli contain potentially confounding factors that complicate the interpretation of the results. This means that even if the empirical findings were highly robust, they might bear little on the theoretical issue under investigation. This suggests that



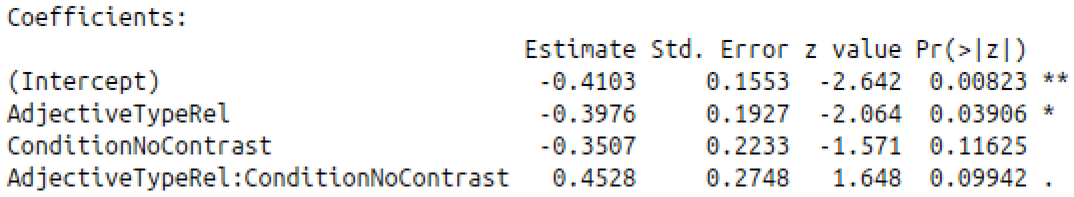
```
Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)                           -0.9008     0.1644  -5.478 4.31e-08 ***
AdjectiveTypeRel                      -0.4009     0.2087  -1.921   0.0548 .  
ConditionNoContrast                   -0.5565     0.2516  -2.212   0.0270 *  
AdjectiveTypeRel:ConditionNoContrast   0.4718     0.3117   1.513   0.1302    
```

(a) Window 2, only fixed effects



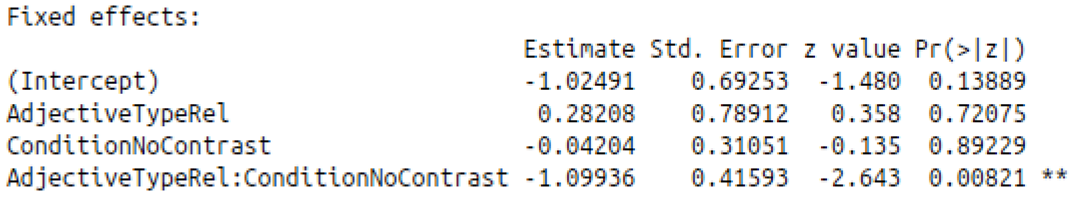
```
Fixed effects:
                                     Estimate Std. Error z value Pr(>|z|)  
(Intercept)                          -1.89598    0.82761  -2.291    0.022 *
AdjectiveTypeRel                      0.01913    0.86069   0.022    0.982  
ConditionNoContrast                  -0.85648    0.41066  -2.086    0.037 *
AdjectiveTypeRel:ConditionNoContrast  0.16936    0.51272   0.330    0.741  
```

(b) Window 2, with random effects



```
Coefficients:
                                     Estimate Std. Error z value Pr(>|z|)   
(Intercept)                           -0.4103     0.1553  -2.642  0.00823 **
AdjectiveTypeRel                      -0.3976     0.1927  -2.064  0.03906 * 
ConditionNoContrast                   -0.3507     0.2233  -1.571  0.11625   
AdjectiveTypeRel:ConditionNoContrast   0.4528     0.2748   1.648  0.09942 . 
```

(c) Window 3, only fixed effects



```
Fixed effects:
                                     Estimate Std. Error z value Pr(>|z|)   
(Intercept)                          -1.02491    0.69253  -1.480  0.13889   
AdjectiveTypeRel                      0.28208    0.78912   0.358  0.72075   
ConditionNoContrast                  -0.04204    0.31051  -0.135  0.89229   
AdjectiveTypeRel:ConditionNoContrast -1.09936    0.41593  -2.643  0.00821 **
```

(d) Window 3, with random effects

Figure 4: Logistic regression models to predict proportions of looks on Target vs Competitor in Window 2 (500–650ms) and Window 3 (650–800ms). AdjectiveType (Max vs Rel) and Condition (Contrast vs NoContrast) are fixed effects. Participant and scene are random effects.

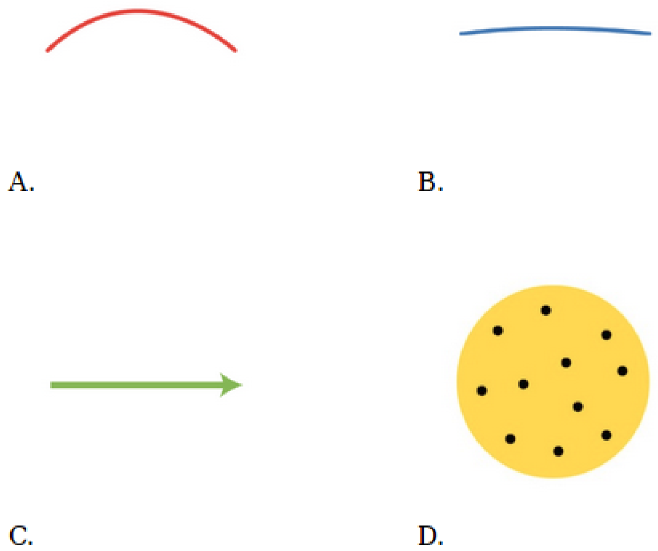


Figure 5: An example of confusing trials, in which participants were asked to *click on the straight line.*

future iterations should have corresponding offline measures to make sure that participants can reliably recognize the speaker's intended target object.

Finally, I also have concerns about the way in which the empirical predictions are derived from the competing theories. It seems to me that the empirical prediction based on Lassiter and Goodman's theory as formulated by Leffel et al. (2016) is largely a straw man. Absolute and relative adjectives using the same mechanism to contextually determine the threshold does not mean that there will be no difference in terms of processing time. As an analogy, the time it takes to run the same algorithm can be very different depending the input. Therefore, even if the stimuli were perfectly crafted to avoid all of the problems in the original experiment, we still need to reconsider what different empirical predictions the two theories really make.

References

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

Kennedy, C., & McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language*, *81*(2), 345–381.

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Proceedings of the 23rd semantics and linguistic theory conference (SALT 23).*

Lassiter, D., & Goodman, N. D. (2015). Adjectival vagueness in a Bayesian model of interpretation. *Synthese*.

Leffel, T., Xiang, M., & Kennedy, C. (2016). Imprecision is pragmatic: Evidence from referential processing. In M. Moroney, C. Rose Little, J. Collard, & D. Burgdorf (Eds.), *Proceedings of the 26th semantics and linguistic theory conference (SALT 26)* (pp. 836–854).

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109 - 147. doi: 10.1016/S0010-0277(99)00025-6