COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 43 (2019) e12769

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12769

Information Integration in Modulation of Pragmatic Inferences During Online Language Comprehension

Rachel Ryskin,[a,b] Chigusa Kurumada,[c] Sarah Brown-Schmidt[d]

[a]*Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology*
[b]*Department of Speech, Language, and Hearing Science, Boston University*
[c]*Department of Brain and Cognitive Sciences, University of Rochester*
[d]*Department of Psychology & Human Development, Vanderbilt University*

Received 27 October 2017; received in revised form 3 June 2019; accepted 5 June 2019

Abstract

Upon hearing a scalar adjective in a definite referring expression such as "*the big*...," listeners typically make anticipatory eye movements to an item in a contrast set, such as a big glass in the context of a smaller glass. Recent studies have suggested that this rapid, contrastive interpretation of scalar adjectives is malleable and calibrated to the speaker's pragmatic competence. In a series of eye-tracking experiments, we explore the nature of the evidence necessary for the modulation of pragmatic inferences in language comprehension, focusing on the complementary roles of top–down information - (knowledge about the particular speaker's pragmatic competence) and bottom-up cues (distributional information about the use of scalar adjectives in the environment). We find that bottom-up evidence alone (e.g., the speaker says "the big dog" in a context with one dog), in large quantities, can be sufficient to trigger modulation of the listener's contrastive inferences, with or without top-down cues to support this adaptation. Further, these findings suggest that listeners track and flexibly combine multiple sources of information in service of efficient pragmatic communication.

Keywords: Language comprehension; Pragmatics; Eye-tracking

1. Introduction

The ability to refer—"my dog," "that rainbow," "her idea"—is one of the central building blocks of natural language. The apparent ease and speed at which humans comprehend language, however, belie the considerable challenges stemming from the many-

Correspondence should be sent to Rachel Ryskin, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 43 Vassar St., 46-3037, Cambridge, MA 12139. E-mail: ryskin@mit.edu

to-many mappings between what can be observed (e.g., a referential expression) and what is intended by the talker (e.g., a referent). That is, the same referent can be referred to with many different expressions (e.g., *the dog, he, Fido, a cocker spaniel*) and the same expression, for example, *the dog*, can be used for multiple referents (e.g., a pet dog, a stuffed animal, an unpleasant person). In communication, language users navigate this variability by integrating multiple sources of information such as properties of the intended referent (see Crain & Steedman, 1985; Roberts, 2003) as well as the context that the referent appears in (Olson, 1970; Osgood, 1971; Pechmann, 1989). In a context with a single dog, the bare noun phrase "the dog" would suffice to achieve reference. By contrast, in a situation with multiple dogs, the speaker would need to provide additional information, for example, through the use of a modified noun phrase, for example, "*the fluffy dog*" (see Brown-Schmidt & Konopka, 2011; Davies & Katsos, 2013; Nadig & Sedivy, 2002; Ryskin, Benjamin, Tullis, & Brown-Schmidt, 2015).

Recent studies have revealed fine-grained knowledge that listeners have about varying degree of contextual sensitivity exhibited by subclasses of modifiers. Experimental studies of referential form (Belke, 2006; Brown-Schmidt & Konopka, 2011; Brown-Schmidt & Tanenhaus, 2006; Nadig & Sedivy, 2002; Sedivy, 2003) show that scalar modifiers such as *tall, big,* and *skinny* are often prompted by the presence of a scalar contrast in the relevant context (e.g., a context with a big dog and a small dog) and are much less likely in contexts with only a single member of the class denoted by the noun (e.g., a context with a single dog). By contrast, color modifiers such as *green* or *aqua* are often used attributively, and thus are much more common than scalars in situations where mentioning color is not necessary to uniquely identify the referent (see Donnellan, 1966 for a discussion of the attributive use of definite reference). In addition, the meaning of color adjectives is thought to be less dependent on the context and the object it describes (see Sedivy, 2003 for discussion).

The fact that prenominal scalar adjectives are highly context sensitive makes them a good candidate for investigating moment-by-moment integration of lexical and contextual information in language comprehension. Along these lines, Sedivy et al. (1999) asked whether listeners use the presence of a scalar adjective in an unfolding noun phrase as a cue that the referent would be a member of a contrast set denoted by the adjective. Sedivy and colleagues evaluated this idea by examining the interpretation of scalar-modified noun phrases such as "*the tall glass*" in the following two contexts: The first context contained a pair of items matching the head noun that contrasted along the scalar dimension denoted by the adjective (e.g., a tall glass and a short glass), a competitor object that was consistent with the adjective but not in a contrast set (e.g., a large pitcher), and an unrelated item (e.g., a key). In the second context, the size-contrasting item (e.g., short glass) was replaced with an unrelated item (e.g., a file folder). Sedivy et al. found that when interpreting scalar-modified noun phrases, listeners looked at the intended referent (tall glass) more quickly when the size-contrasting object (short glass) was present in the display, compared to when it was replaced by an unrelated item (file folder). This finding demonstrates that interpretation of a prenominal scalar adjective is facilitated by the presence of a relevant scalar-contrast in the display.

While this finding suggests that scalar adjectives have predictive validity, it is not clear to what extent the contrast effect represents pragmatic inference about the speaker's referential intent. Grodner and Sedivy (2011) reasoned that if the contrast effect was contextually supported, it could be attenuated in a situation where the speaker deviated from the normal conversational usage of scalars and therefore was unlikely to use scalar adjectives with the intent to highlight a contextual contrast. If, on the other hand, the contrast effect is more automatic or directly tied to semantic properties of scalar adjectives, it should be impermeable to such circumstantial factors. To tease apart these two accounts, Grodner and Sedivy followed up on Sedivy et al. (1999), with a task in which participants heard instructions such as "Pick up the tall glass," produced either by a reliable or an unreliable speaker (manipulated between subjects). The reliable speaker condition replicated Sedivy et al. (1999), in which all instructions were given in a conventional manner. In the unreliable speaker condition, participants were told that the instructions were recorded by a speaker with "an impairment that caused language and social problems." Subsequently, in filler trials, the speaker mislabeled objects and referred to inappropriate locations, and was consistently over-informative in the use of size adjectives (e.g., "the tall glass" in a context with a single glass). Analysis of eye-gaze as participants interpreted *"the tall (glass)..."* showed that participants in the reliable speaker condition made more fixations to the intended referent (e.g., tall cup) when a contrast item (e.g., short cup) was present, replicating Sedivy et al. (1999), while listeners in the unreliable speaker condition did not. This result suggests that participants can suspend their contrastive interpretation when the current speaker is less likely to use a scalar adjective to signal a contextual contrast. Along with other findings (Sedivy, 2003), this work supports the idea that the contrast effect is a contextually sensitive pragmatic effect, subject to modulation based on what inferences are licensed in a given environment.

However, little is known about the nature of the evidence that elicits a change in the listener's inference process. Did the modulation come about primarily because the listener noticed repeated failures to use scalar adjectives appropriately or was it the outcome of a judgment elicited by the description of the speaker as having "language and social impairments"? Or do the two types of cues play complementary roles in modulating the listener's inferences? Answering these questions will shed light on how the language system assigns weights to different sources of information in the adjustment of language comprehension behaviors.

One way to approach this question is to regard the modulation of eye movements as the result of statistical learning, wherein the behavioral response to a scalar adjective (e.g., "large") is calibrated to its likelihood of signaling a contextual contrast (e.g., p (meaning = contrast | "large"), the probability of contrastive meaning given that the chosen adjective is "large"). Given the general sensitivity to input statistics attested in comprehension (e.g., Creel et al., 2008; Creel, 2014; Creel & Tumlin, 2011; Fine et al., 2013; Fine & Florian Jaeger, 2013; Wells et al., 2009) and acquisition (e.g., Aslin & Newport, 2014; Saffran et al., 1996; Smith & Yu, 2008; Wonnacott et al., 2008; Yurovsky et al., 2014), it is plausible for listeners to accrue the relevant statistics and suppress contrastive interpretation of adjectives to avoid misinterpretation. If so, the lower the probability that

an adjective signals a contextual contrast in a given discourse context (or experimental session), the fewer anticipatory eye movements listeners should make.

An important remaining question is about how such a statistical learning mechanism is influenced by the top–down instructions about the speaker. Recall, in Grodner and Sedivy's (2011) design, an explicit instructional manipulation was used to convey that the speaker's use of language may be idiosyncratic in unpredictable ways. If the primary mechanism of pragmatic adaptation is based on the bottom-up (statistical) input, such top–down information can be facilitatory, but not necessary. That is, top–down cues may draw attention to the idiosyncrasy of the talker's adjective use, or possibly provide scaffolding to support learning (e.g., Bransford & Johnson, 1972). Listeners should, however, in principle exhibit the same behavioral change without an explicit, top–down, cue. It is, in this light, interesting that Grodner and Sedivy (2011) in fact mention a follow-up experiment that did not use an explicit instruction and report a null effect. This suggests that the pragmatic modulation may be critically dependent on, or at least highly sensitive to, information about the speaker provided through multiple channels of communication. For instance, the top–down manipulation may make it clear that the observed idiosyncrasy is most plausibly attributed to the *speaker* rather than other possible causes (e.g., a technical problem with the stimulus presentation), making it unlikely that an otherwise plausible pragmatic interpretation (i.e., that scalars refer to a contrast) should apply to their choice of lexical items. A similar top-down manipulation was used to modulate listeners' interpretation of disfluencies (Arnold et al., 2007). Typically, a speaker's disfluencies prompt listeners to anticipate a word that the speaker might find difficult to utter, such as a referent that is unfamiliar. However, when listeners were told that the speaker had object agnosia, disfluencies ceased to trigger biased looks to unfamiliar objects. This suggests that statistical learning of p(meaning = contrast | "large") may be effectively conditioned on a speaker insofar as there is a plausible explanation as to *why* the speaker's production exhibits statistics deviating from what is ordinarily expected.

In this research, we delve into this issue by manipulating the types and amount of input given to the listener. In what follows, we present four experiments that build upon the findings of Grodner and Sedivy (2011), conceptually replicating their work using a computerized paradigm with a substantial number of observations ensuring sufficient statistical power. Thereby, we address the question of whether the bottom-up input alone can modulate pragmatic inferences during online language processing. To anticipate our results, we find that bottom-up evidence alone can trigger modulation of contrastive inferences but only following massive exposure. We highlight implications of our results and discuss possible mechanisms that support online pragmatic inferences in linguistic communication.

2. Experiment 1

The aim of Experiment 1 is to test whether listeners modulate their pragmatic inferences based on bottom-up exposure alone. We test this by examining eye fixations during the interpretation of scalar adjectives after exposure to a speaker using scalar cues to

contrast in an appropriately informative way (reliable pragmatic context) or a speaker using scalars over-informatively (unreliable pragmatic context).[1] In Experiment 1a, the speaker in the unreliable pragmatic context condition used scalar adjectives in situations where they were over-informative and not necessary to disambiguate the target referent from the immediate display (e.g., said "big circle" when there is only one circle, and three other big and small shapes). In Experiment 1b, the speaker's adjective use in the unreliable pragmatic context condition was not only over-informative but also sometimes inaccurate.

2.1. Method

2.1.1. Participants

A total of 108 students (40 in Experiment 1a, 68 in Experiment 1b) from the University of Illinois at Urbana-Champaign were given partial course credit upon participation in the experiment. All participants were fluent speakers of American English with normal or corrected-to-normal vision and normal hearing.

2.1.2. Procedure and materials

The experiments consisted of two phases: training (12 trials) and test (50 trials). In both phases, participants listened to auditory instructions (e.g., "Show me the circle.") and selected the target referent from a display of four pictures in a 2 × 2 grid (Fig. 1) by clicking on it with the computer mouse. Stimuli were presented on a 1,920 × 1,080 pixel display using the Psychtoolbox-3 extension (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997) for Matlab. During the test phase, participants' eye movements were tracked using an EyeLink-1000 desktop mounted eye-tracker, with a sampling rate of 1,000 Hz. The entire experimental session lasted about 20 min. See Appendix A for a summary of trials in all experiments.

2.1.2.1. Training phase: Before the start of the training phase, participants were instructed to imagine that "little Joe" and his mom are playing a game on her computer and their job was to listen to her speech and click on the pictures that she was talking about.

The training phase consisted of 12 trials in which participants saw four shapes (squares, triangles, or circles) with different combinations of colors (red, blue, or yellow), sizes (big or small), and patterns (checkers, dots, or stripes). Each display was accompanied by two sequential instructions[2], such as "Show me the blue triangle. Now, show me the small square." The details of visual and auditory stimuli for each trial are available at https://osf.io/5geba/. The order of the training trials and the location of pictures in the display were randomized for each participant within the Matlab code.

Participants were randomly assigned to one of two (between-subjects) training conditions: reliable pragmatic context vs. unreliable pragmatic context. The auditory stimuli were held constant across training conditions and reliability of the pragmatic context was

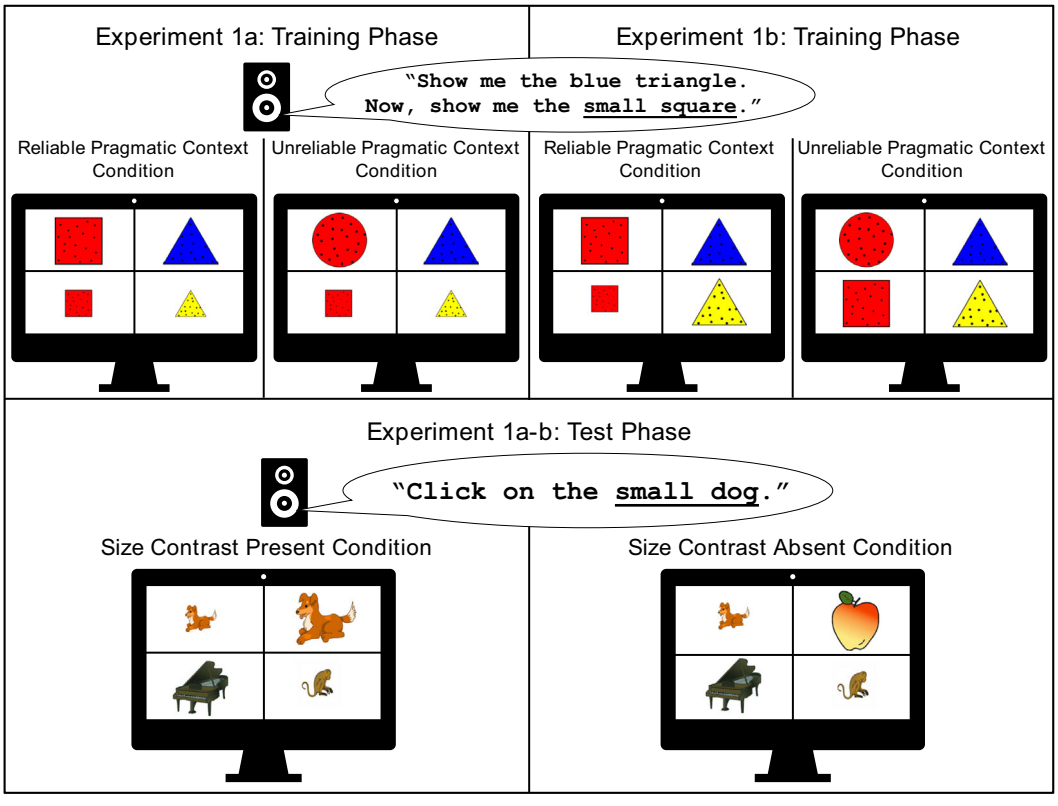

Fig. 1. Schematic of trials in Experiments 1a and b.

manipulated through the visual display, by switching one (or two) of the shapes. The details of this manipulation differ between Experiment 1a and 1b.

2.1.2.2. Experiment 1a: In all 12 training trials, two of the shapes in the display were big and two were small. In the reliable pragmatic context condition, all 12 trials contained felicitous use of scalar adjectives (e.g., "Show me the blue triangle. Now, show me the small square.") when a size contrast was present in the display (e.g., small red square vs. large red square; see Fig. 1). In the unreliable pragmatic context condition, all 12 trials contained infelicitous scalar adjectives that were over-informative because the target item was uniquely identifiable without referring to a size. The target was either a singleton (as in Fig. 1) or distinguished by color but not size (e.g., "Point to the triangle. Now, point to the small red circle." when there are two circles and both of them are small).

2.1.2.3. Experiment 1b: The reliable pragmatic context condition trials were identical to Experiment 1a. In the unreliable pragmatic context condition, there were three types of

trials: (a) Six of the trials included over-informative pre-nominal scalar adjectives (e.g., "Now, show me the small circle" when the display contains a small red circle, a small blue triangle, a small yellow triangle, and a small blue square); (b) three trials contained a post-nominal over-informative scalar adjective (e.g., "Show me the circle that's large" with a display that contains a large red circle, large yellow square, large blue triangle, large red square); (c) three trials contained a pre-nominal scalar adjective that didn't match the size of any of the four shapes (see Fig. 1).

2.1.2.4. Test phase: Before the test phase, the experimenter conducted a 9-point eye-tracker calibration and validation procedure. Participants were then informed that they would be hearing more instructions from the same speaker. The test phase consisted of 50 trials (20 critical and 30 filler trials) modeled after Grodner and Sedivy (2011), with a critical difference being that we implemented it in a computer-based paradigm while the original study used 3D props. Participants from both training conditions (reliable vs. unreliable pragmatic context) saw the exact same test trials.

On critical trials, two of the pictures in the display were big and two were small and the instruction contained a pre-nominal scalar adjective (e.g., "Click on the small dog."). Half of the critical trials contained a pair of pictures that differed only in size (size contrast present condition), and the other half contained four unique pictures (size contrast absent condition; see Fig. 1).

There were also three types of filler trials: (a) Ten of the filler trials did not include a scalar adjective (e.g., "Click on the dog" paired with an unambiguous display, a small dog, a large flag, large scissors, small scissors). (b) Ten of the filler trials included a pair of items in a non-scalar contrast (e.g., Material: "Point to the leather jacket" with a leather jacket, a rain jacket, a large hydrant, and a small dollar). (c) The last ten filler trials did not include a scalar (e.g., "Point to the bike"), but the display did contain a pair of items in a non-scalar contrast (e.g., a small bike, a large glass, a lead pencil, and a coloring pencil).

Two counterbalancing lists were created to allow target items (e.g., small dog) to appear both in the size contrast present and size contrast absent conditions across subjects (available at https://osf.io/5geba/). Participants were randomly assigned to a counterbalancing list. Target items were never repeated for a given participant, but they could reappear as distractor items.

2.2. Results

Interpretation of the scalar adjective was indexed by the proportion of eye movements that participants made to the target item as they interpreted the critical instructions, which consisted of a scalar adjective and a noun (e.g., *Click on the small dog*). A fixation was coded as a target fixation if the x,y fixation-coordinates landed on the target object (e.g., *the small dog*), or on the white space in the quadrant of the screen surrounding it (this buffer space did not overlap with any other object). The full time-courses of target fixations by conditions in Experiments 1a-b are shown in Figure 2a-b. Target fixations were analyzed in

two ways: (a) as average proportions (duration of target fixations divided by total duration of all fixations) across the time window of interest, and (b) as a binary measure (fixations to the target vs. not) for every 10 milliseconds (ms) within the time window.

2.3. Average target fixation proportions

The critical time window began 200 ms after the onset of the adjective (e.g., *small*) and ended 200 ms after the offset of the noun (e.g., *dog*). The mean duration of this time window was 726 ms. The 200 ms delay was included to account for the time needed to program and launch an eye movement (Hallett, 1986). The average proportions of target fixations within this time-window across pragmatic context (training) and size contrast presence conditions are shown in Figure 3a–b. Figure S2.1 (Material S2) shows average target fixation proportions for the adjective window (adjective onset + 200 ms to noun

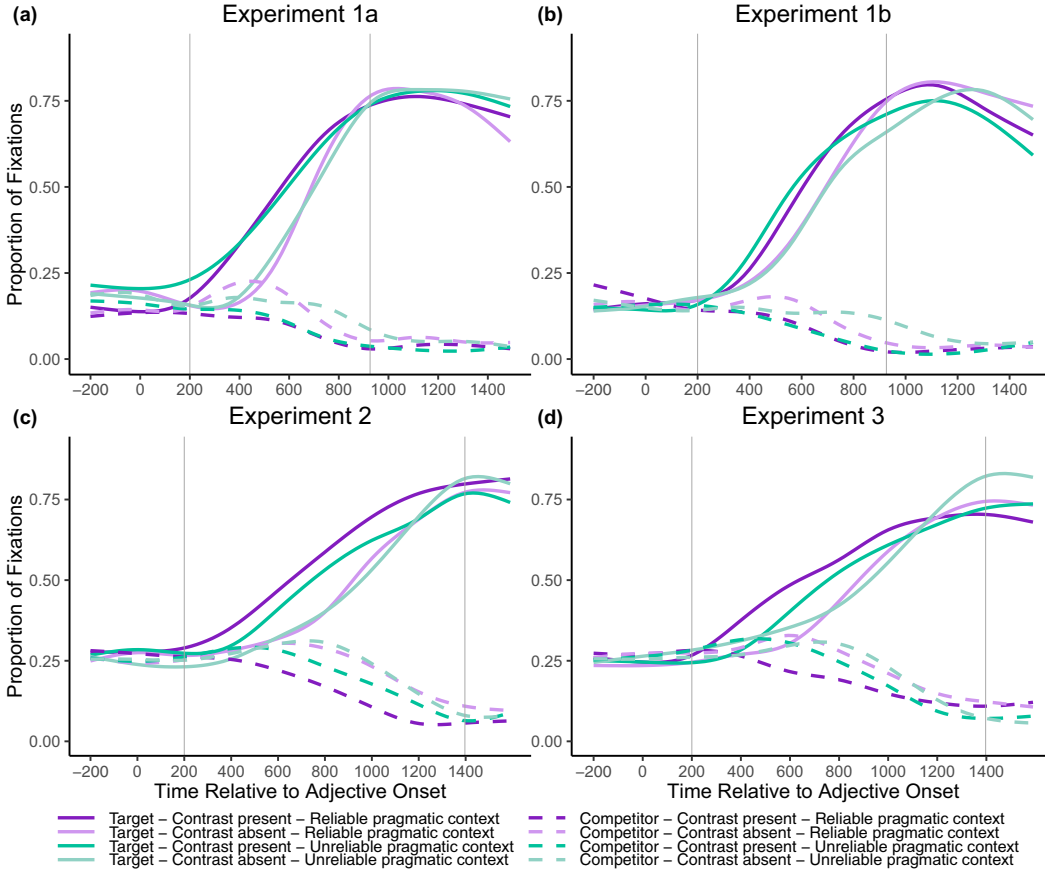

Fig. 2. Time-course of proportions of fixations to target and competitor images for all experiments during instructions (e.g., *Click on the small dog*) by pragmatic context (reliable vs. unreliable) and size contrast (present vs. absent) conditions. Vertical lines indicate the approximate time window used for analyses, starting 200 ms after the onset of the adjective and ending 200 ms after the offset of the noun.

onset + 200 ms) and noun window (noun onset + 200 ms to noun offset + 200 ms) separately.

The trial-level proportions of target fixation durations were first transformed using the empirical logit transformation and then analyzed in a multilevel linear regression, using the *lme4* software package in R (Bates, Maechler, Bolker, & Walker, 2015), as well as the *lmerTest* package in R (Kuznetsova et al., 2016). Pragmatic context and size contrast along with their interaction were entered as fixed effects with participants and items as random effects. We included random by-participants random slopes for size contrast and by-items random slopes for pragmatic context, size contrast, and their interaction. All fixed effects were deviation coded (size contrast condition: size contrast absent = −0.5, size contrast present = 0.5; pragmatic context condition: reliable pragmatic context = −0.5, unreliable pragmatic context = 0.5).

The full model estimates can be found in Table 1. There was an expected main effect of size contrast presence, such that participants made more target fixations when a size

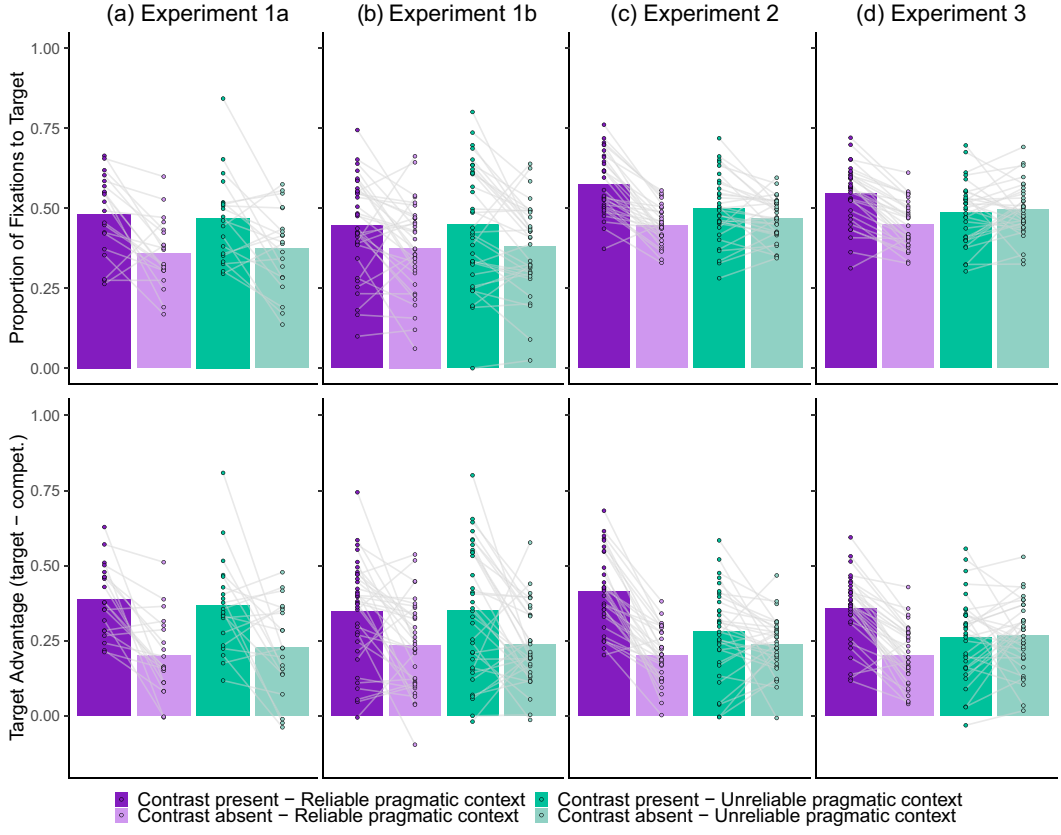

Fig. 3. Top row: Average proportions of target fixations during the interpretation of the scalar adjective and noun (e.g., *Click on the small dog*) by size contrast presence and pragmatic context conditions. Bottom row: Target advantage (proportion of target fixations minus proportion of competitor fixations) for the same time window. Points represent individual subject means.

contrast was present (E1a: $t = 4.01$; E1b: $t = 3.27$). The main effect of pragmatic context was not significant (E1a: $t = 0.14$; E1b: $t = -0.01$), and there was no significant interaction of size contrast and pragmatic context (E1a: $t = -0.03$; E1b: $t = 0.37$). Models for the adjective and noun windows separately are reported in Appendix S2: the patterns are consistent with these main analyses (though size contrast condition is only marginally significant in the adjective window for Experiment 1b).

2.4. Binary target fixations

In order to obtain a more fine-grained measure of the target fixation behavior, we analyzed these data with autoregressive logistic mixed-effect models (Cho, Brown-Schmidt, & Lee, 2018). We conducted this analysis to ensure that the observed presence and absence of effects reported above (Table 1) were not due to the aggregation of eye movements over a large time-window. In the current and following experiments, we use both analytic approaches for increased confidence in our statistical inferences drawn from the observed data. The critical time window began 200 ms after the onset of the adjective and ended 750 ms later (the average duration of the adjective and noun).

The binary fixation data (a 0 or 1 for whether there was a target fixation in each 10 ms bin in each trial for each participant) were analyzed in a multilevel logistic regression, accounting for the autocorrelation between adjacent time bins (i.e., if a participant

Table 1
Experiment 1a–b: Results of the linear mixed-effects model of empirical logit-transformed target fixation proportions over the critical time window (adjective onset + 200 ms to noun offset + 200 ms)

	Experiment 1a				Experiment 1b			
	β	*SE*	*t*	*p*	β	*SE*	*t*	*p*
Fixed effects								
Intercept	−0.17	0.05	−3.61	< .001	−0.19	0.05	−4.06	< .001
Size contrast condition	0.24	0.06	4.01	< .001	0.15	0.05	3.27	< .005
Pragmatic context condition	0.01	0.08	0.14	0.89	0.00	0.08	−0.01	0.99
Size contrast × pragmatic context	−0.00	0.13	−0.03	0.98	0.04	0.10	0.37	0.71
		SD				*SD*		
Random effects								
Participants: Intercept		0.18				0.27		
Participants: Size contrast		0.15				0.13		
Items: Intercept		0.14				0.12		
Items: Size contrast		0.11				0.11		
Items: Prag. context		0.17				0.06		
Items: Size contrast x Prag. context		0.19				0.08		
Residual		0.59				0.59		
		Observations: 800; Items: 20; Participants: 40				Observations: 1,360; Items: 40; Participants: 68		

is fixating the target at time, *t*, there is a higher chance that they will be fixating the target at $t + 1$ than if they are not fixating the target at *t*). Pragmatic context, size contrast, previous time step binary target fixation (fix_{t-1}), and the interaction of pragmatic context and size contrast were entered as fixed effects with participants and items as random effects. Fixed effects of interest were deviation coded (size contrast condition: size contrast absent= −0.5, size contrast present = 0.5; pragmatic context condition: reliable pragmatic context = −0.5, unreliable pragmatic context = 0.5). We included random by-participants random slopes for size contrast and fix_{t-1} and by-items random slopes for pragmatic context, size contrast, their interaction, and fix_{t-1} in the initial models. The procedure, throughout this paper, was to start by fitting the GLMM as in Cho et al. (2018) with the "bobyqa" optimizer (Powell, 2009). When it did not converge, we compared the fit across all possible optimizers (using the allFit function) and only simplified the model if none of the others reached convergence or the fixed effect estimates were not the same (to four decimal points) across optimizers. We then iteratively removed random slopes which appeared to capture the least variance according to the incomplete model fits (starting with interactions).

The full model estimates can be found in Table 2. Target fixations in the preceding time window significantly predicted target fixations (E1a: $z = 71.96$; E1b: $z = -115.27$). As in the proportion data, there was an expected main effect of size contrast presence, such that participants made more target fixations when a size contrast was present (E1a: z = 2.06; E1b: z = 2.85). The main effect of pragmatic context was not significant (E1a: $z = -0.48$; E1b: $z = -0.46$), and there was no significant interaction of size contrast and pragmatic context (E1a: $z = -0.33$; E1b: $z = 1.23$). Separate models for the adjective and noun windows are reported in Appendix S2. The patterns are broadly consistent with these main analyses, though the effect of size contrast was significant in the adjective windows (E1a: $z = 3.85$; E1b: $z = 3.27$) but not the noun windows (E1a: $z = -0.18$; E1b: $z = 1.77$).

2.5. Discussion

During the interpretation of scalar adjectives, participants made more fixations to targets that were in a size contrast set, consistent with results from Sedivy et al. (1999) and Grodner and Sedivy (2011). The pragmatic context manipulation, on the other hand, had no significant effect. One possible source of this null effect may be the fact that the unreliable pragmatic context was instantiated primarily using over-informative instructions (exclusively so in Experiment 1a). It has been reported that naturalistic adjective use contains a large amount of instances in which an adjective is not strictly necessary with respect to the goal of unique reference (Brown-Schmidt & Konopka, 2011). Over-informative adjectives may not impair on-line language processing (Arts, Maes, Noordman, & Jansen, 2011; Davies & Katsos, 2013; Levelt, 1989; Rubio-Fernández, 2016; cf., Engelhardt et al., 2011) and may reflect natural properties of utterance formulation (Belke, 2006; Pechmann, 1989). Indeed, post hoc norming (see Appendix B for details) of the naturalness of instructions paired with their corresponding displays (on a scale of 1–5)

Table 2
Experiment 1a–b: Results of the autoregressive generalized linear mixed-effects models of binary target fixations over the critical time window (adjective onset + 200 ms to adjective onset + 750 ms)

	Experiment 1a				Experiment 1b			
	β	*SE*	*z*	*p*	β	*SE*	*z*	*p*
Fixed effects								
Intercept	−3.98	0.06	−65.38	< .001	−4.00	0.06	−70.08	< .001
Size contrast condition	0.16	0.08	2.06	0.04	0.18	0.06	2.85	< .001
Pragmatic context condition	−0.04	0.09	−0.48	0.63	−0.05	0.10	−0.46	0.64
Fix_{t-1}	9.02	0.13	71.96	< .001	8.92	0.08	115.27	< .001
Size contrast × pragmatic context	−0.06	0.18	−0.33	0.74	0.16	0.13	1.23	0.22
		SD				*SD*		
Random effects								
Participants: Intercept		0.18				0.36		
Participants: Size contrast		–				0.11		
Participants: Fix_{t-1}		0.35				0.17		
Items: Intercept		0.16				0.08		
Items: Size contrast		–				0.10		
Items: Pragmatic context		0.12				0.04		
Items: Fix_{t-1}		0.24				0.09		
Items: Size contrast × Pragmatic context		–				0.07		
		Observations: 58,400; Items: 20; Participants: 40				Observations: 99,280; Items: 20; Participants: 68		

revealed that participants rated over-informative instructions from Experiment 1a (M = 4.83) as equally natural compared to optimally informative instructions (M = 4.88).

Furthermore, the use of scalar adjectives that were over-informative for a given visual display may have been attributed to a looser definition of the comparison class. For example, the comparison class for a triangle may be other shapes in general (e.g., "*small* square" to contrast with a larger circle). For this reason, Experiment 1b contained scalar adjectives that were over-informative relative to a larger comparison class—all the shapes in the immediate display—or plainly inaccurate (e.g., "the small square" when there are only big shapes in the display). However, these additional cues did not appear to have an effect on contrastive interpretations of adjectives during the test phase of the experiment.

These results may suggest that listeners do not discount scalar adjectives as a cue to a contextual contrast solely based on the exposure to infelicitous uses of these adjectives. Alternatively, it is possible that the experimental design used in Experiments 1a-b may not have provided suitable circumstances to observe modulation of online interpretation of scalar adjectives. Because the three shapes (circles, square, triangle) were repeated across training trials, the use of scalars that were over-informative for a given visual display may have been attributed to a tendency to lexically differentiate the currently observed shapes from those seen on previous trials (Van Der Wege, 2009; Yoon & Brown-Schmidt, 2014). We address these possibilities in Experiment 2.

It may also be the case that participants altered their pragmatic inferences during the training phase, but this did not transfer to the test phase because the context change was too abrupt. Even though we tried to ensure the continuity of the two phases by instructing the participant that the speaker remained the same across the training and test phases, these phases differed in two important ways: (a) the training phase consisted of trials with colored geometric shapes while the test phase consisted of trials with more complex images, and (b) participants' eye movements were monitored during the test phase but not the training phase (and as a result, a calibration occurred in between the two phases). Memory retrieval is contextually sensitive, and changes in context from learning to test can impair retrieval of previously learned information (Godden & Baddeley, 1975; Smith & Vela, 2001; cf. Eich, 1985, Mulligan, 2011). As a result, the contextual changes between training and test might have limited transfer of learning, mitigating whatever effect of training there was in the first place.

Finally, the amount of evidence that the pragmatic context was unreliable may have been insufficient to elicit long-lasting changes of interpretations. Across the entire experimental session, the proportion of infelicitous uses of a pragmatic cue was only 36% in the unreliable pragmatic context conditions.[3] Perhaps the large number of informative trials (including 30 fillers) at test counteracted any adaptation that resulted from the training phase. Experiment 2 was conducted to address these concerns.

3. Experiment 2

The goal of Experiment 2 is to provide a more suitable environment for observing changes in the interpretation of scalar adjectives. To achieve this, the contexts during training (exposure) and test were made more similar in multiple ways: (i) exposure and test trials were randomly intermixed, rather than there being separate training and test phases, (ii) eye-tracking occurred during all trials and the calibration was done at the very beginning of the experimental session, (iii) the images used in exposure and test were taken from a larger set of pictures of animals or objects. Additionally, in the unreliable pragmatic context condition in Experiment 3, over- and under-informative sentences constitute the vast majority (93%) of what the participant is exposed to.

3.1. Method

3.1.1. Participants

Sixty-four undergraduate students at the University of Illinois at Urbana-Champaign participated in this experiment in exchange for partial course credit. Participants had normal or corrected-to-normal vision and spoke English fluently.

3.1.2. Procedure and materials

Prior to beginning the experiment, participants were told to listen to the instructions (e.g., "Click on the big pickle"), then click on the item in the four-picture display that

best matched what the speaker said and, if they were not sure, to just use their best guess.[4] This instruction was included because some of the sentences were globally ambiguous (e.g., "Click on the pickle" when there are two pickles).

Stimulus display and eye-tracking were performed with the same setup as described in Experiment 1. Visual stimuli were assembled from images used in Brady, Konkle, Alvarez, and Oliva (2008, 2013). Auditory stimuli were recorded using Praat (Boersma & Weenink, 2016) in a sound isolation booth by the first author (available at https://osf.io/5geba/).

The experimental session consisted of 300 trials presented as a single phase and lasted about 30 min. Each participant saw 80 scalar exposure trials, 180 non-scalar exposure trials, and 40 test trials in an individually randomized order (with the constraint that the first three trials were non-scalar exposure trials). Participants were randomly assigned to the reliable pragmatic context or unreliable pragmatic context condition. The exposure trials differed by pragmatic context condition (between-subjects). See Appendix A for a summary of trials in all experiments. As in Experiment 1, the auditory stimuli were held constant across conditions and reliability of the pragmatic context was manipulated through the visual display, by switching one of the four images in exposure trials (see Fig. 4).

3.1.2.1. Primary exposure trials: In all 80 scalar exposure trials, two of the shapes in the display were big and two were small. In the reliable pragmatic context condition, half (40) of the scalar exposure trial instructions contained a scalar adjective (e.g., "Click on the big briefcase") and half (40) did not (e.g., "Click on the briefcase"). The presence or absence of a size adjective was felicitous: The target was a member of a size contrast pair (e.g., a big briefcase and a small briefcase) when the size adjective was present and the target was a singleton (e.g., only one briefcase in the display) when the size adjective was absent (see Fig. 4).

In the unreliable pragmatic context condition, half (40) of the scalar exposure trial instructions contained a scalar adjective (e.g., "Click on the big briefcase") and half (40) did not (e.g., "Click on the briefcase"). Critically, the presence or absence of a size adjective was *always infelicitous*: The target was a singleton (e.g., only one briefcase in the display) when the size adjective was present and the target was a member of a size contrast pair (e.g., a big briefcase and a small briefcase) when the size adjective was absent (see Fig. 4). Note that when the target item is a member of a size-contrasted pair (e.g., a big briefcase and a small briefcase) and no size adjective is provided to uniquely identify the referent (e.g., "Click on the briefcase"), participants are obliged to click on one of the pair members at random (they were told they should make their "best guess" in the instructions).

3.1.2.2. Secondary exposure trials: Three types of non-scalar exposure trials were used to make targets of scalar exposure and test trials somewhat less predictable. In the reliable pragmatic context condition, the presence or absence of (non-scalar) adjectives was felicitous.[5] In the unreliable pragmatic context condition, the presence or absence of (non-scalar) adjectives was *always infelicitous*.[6]

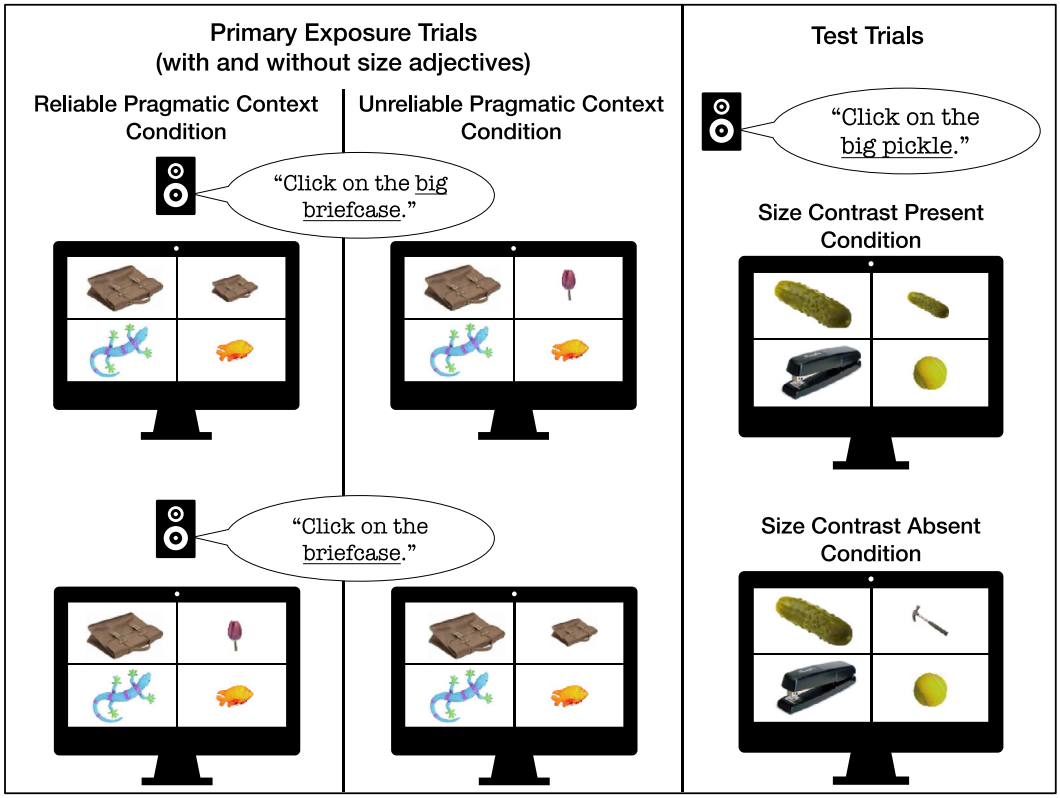

Fig. 4. Schematic of primary exposure and test trials in Experiments 2–3. (Secondary exposure trials not shown.) Unlike in Experiment 1, test trials were interspersed with exposure trials and appeared throughout the experiment.

3.1.2.3. Test trials: Test trial instructions always contained a scalar adjective (e.g., "Click on the big pickle"). Half (20) of the test trials contained a target item that was a member of a size-contrasted pair (e.g., a big pickle and a small pickle; size contrast present condition) and two distractor items (one big, one small; see Fig. 4). In the other half (20) of the test trials, the target item (e.g., big pickle) did not have a pair in the display (size contrast absent condition). Instead, there were three distractors (Fig. 4).

Target items (e.g., big pickle) were never repeated for a given subject, but they could reappear as distractor items. Two lists were created to allow target items to appear both in the Contrast and No Contrast conditions across subjects (available at https://osf.io/5geba/). However, due to a counterbalancing error, participants in the reliable pragmatic context were always assigned to list one and participants in the unreliable pragmatic context only received list two. Although we do not expect there to be any meaningful differences between lists, this does affect the analyses that can be conducted. We discuss this in more detail in the Results section.

3.2. Results

Fixations were coded and analyzed in the same way as in Experiment 1. A plot of the full time-course of fixations by conditions can be seen in Fig. 2c.

3.2.1. Average target fixation proportions

The critical time window began 200 ms after the onset of the adjective (e.g., *big*) and ended 200 ms after the offset of the noun (e.g., *pickle*). The mean duration of this time window was 1,197 milliseconds.[7] The average proportions of target fixations within this time window across the pragmatic context (training) and size contrast presence conditions are shown in Fig. 3c. Appendix S2 shows average target fixation proportions for the adjective window (adjective onset + 200 ms to noun onset + 200 ms) and noun window (noun onset + 200 ms to noun offset + 200 ms) separately.

As in Experiment 1a–b, the trial-level proportions of target fixation durations were first transformed using the empirical logit transformation and then analyzed in a multilevel linear regression. Pragmatic context and size contrast along with their interaction were entered as fixed effects with participants and items as random effects. We included random by-participants random slopes for size contrast and by-items random slopes for pragmatic context.[8] All fixed effects were deviation coded (size contrast condition: size contrast absent = −0.5, size contrast present = 0.5; pragmatic context condition: reliable pragmatic context = −0.5, unreliable pragmatic context = 0.5).

The full model estimates can be found in Table 3. There was an expected main effect of size contrast presence, such that participants made more target fixations when a size contrast was present ($t = 5.53$). The main effect of the pragmatic context was not significant ($t = -1.37$), but there was a significant interaction between size contrast and pragmatic context ($t = -3.24$). A follow-up analysis setting the reliable pragmatic context condition as reference level revealed a large size contrast effect ($\beta = 0.27$, $SE = 0.04$, $t = 7.47$, $p < .001$) in the reliable pragmatic context condition, and a reduction of this effect in the unreliable pragmatic context condition ($\beta = -0.20$, $SE = 0.06$, $t = -3.24$, $p < .002$). Models for the adjective and noun windows separately are reported in Appendix S2: The patterns are consistent with the main analysis. There is a significant effect of size contrast (adjective window: $t = 6.77$, noun window: $t = 4.20$), and it interacts significantly with pragmatic context (adjective window: $t = -2.09$, noun window: $t = -2.80$).

3.3. Binary target fixations

We analyzed the binary fixation data (a 0 or 1 for whether there was a target fixation in each 10 ms bin in each trial for each participant) in a multilevel logistic regression accounting for autocorrelation in fixations. The critical time window began 200 ms after the onset of the adjective and ended 1,200 ms later (the average duration of the adjective and noun rounded to the nearest 10 ms).

Table 3
Experiment 2: Results of the linear mixed-effects model of empirical logit-transformed target fixation proportions over the critical time window (adjective onset + 200 ms to noun offset + 200 ms)

	β	*SE*	*t*	*p*
Fixed effects				
Intercept	−0.01	0.02	−0.32	0.75
Size contrast condition	0.17	0.03	5.53	< .001
Pragmatic context condition	−0.06	0.04	−1.37	0.18
Size contrast × Pragmatic context	−0.20	0.06	−3.24	.002
	SD			
Random effects				
Participants (N = 63): Intercept	0.13			
Participant: Size contrast	0.10			
Items (N = 40): Intercept	0.06			
Items: Pragmatic context	0.10			
Residual	0.55			
Observations: 2520				

Pragmatic context, size contrast, fix_{t-1} (whether there was a target fixation in the preceding timebin), and the interaction of pragmatic context and size contrast were entered as fixed effects with participants and items as random effects. We included random by-participants random slopes for size contrast and fix_{t-1} and by-items random slopes for pragmatic context and fix_{t-1}, but the maximal model did not converge. All fixed effects of interest were deviation coded (size contrast condition: size contrast absent = −0.5, size contrast present = 0.5; pragmatic context condition: reliable pragmatic context = −0.5, unreliable pragmatic context = 0.5).

The full model estimates can be found in Table 4. Target fixations in the preceding time window significantly predicted target fixations (z = −255.93). As in the proportion data, there was an expected main effect of size contrast presence, such that participants made more target fixations when a size contrast was present (z = 3.25). The main effect of pragmatic context was not significant (z = −1.10), but there was a marginally significant interaction of size contrast and pragmatic context (z = −1.87). Models for the adjective and noun windows separately are reported in Appendix S2. In the adjective and noun windows individually, there were significant effects of size contrast (adjective window: z = 6.36, noun window: z = −2.84 note: in the opposite direction) but no significant interactions with pragmatic context (adjective window: z = −0.90, noun window: z = −0.76).

3.4. Effect of exposure over time

In order to test if this modulation is robust over time, we examine the data patterns across the two halves of the experiment by adding experiment half (first = −0.5 vs. second = 0.5) as a predictor into the models (both predicting average proportions and

Table 4
Experiment 2: Results of the autoregressive generalized linear mixed-effects models of binary target fixations over the critical time window (adjective onset + 200 ms to adjective onset + 1,200 ms)

	β	*SE*	*z*	*p*
Fixed effects				
Intercept	−4.08	0.03	−152.08	< .001
Size contrast condition	0.11	0.03	3.25	< .001
Pragmatic context condition	−0.05	0.05	−1.10	0.27
Fix_{t-1}	8.80	0.03	−255.93	< .001
Size contrast x Pragmatic context	−0.14	0.08	−1.87	0.06
	SD			
Random effects				
Participants (N = 63): Intercept	0.12			
Participants: Size	0.05			
Items (N = 40): Intercept	0.06			
Items: Prag. Context	0.07			
Observations: 302400				

binary fixations).[9] In each half of the experiment, participants encountered approximately 40 training trials, 20 test trials, and 90 fillers. Full summaries of the models are reported in Appendix S3. As in the main analysis, there was a main effect of size contrast presence (t = 5.48; z = 3.19) and a (marginally) significant interaction of size contrast presence and pragmatic context condition (t = −3.27; z = −1.88). There was also a marginally significant three-way interaction (though only for the analysis of average target fixation proportions) between size contrast condition, pragmatic context condition, and experiment half (t = 2.01; z = 1.61), such that the interaction between size contrast and pragmatic context may be reduced in the second half of the experiment. This may be largely due to the fact that target fixations overall (marginally) decrease in the second half of the experiment (t = −1.91; z = −2.74), perhaps as participants lose motivation or interest in the task. The larger effect in the first half of the experiment suggests that the large token count of input, 300 trials in total, was not necessary to trigger the modulation. Most likely it was the large proportion of training items (93% over- or under-informative use of scalar adjectives) that was responsible for the interaction between the size contrast and pragmatic context conditions.

3.5. Discussion

The contrast-contingent target preference was diminished when participants were exposed to a speaker who used adjectives infelicitously, in the absence of any explicit information about the speaker. These results suggest that listeners can discount scalar

adjectives as a cue to a contextual contrast based on the bottom-up information alone. Experiment 1 and 2 together suggest that such modulation of online interpretation of scalar adjectives requires a consistent context as well as a great deal of evidence that the speaker is unlikely to use an adjective in an informative manner.

Of interest is that the effect of size contrast presence was persistent, and it was observed even in an unreliable pragmatic context despite overwhelming evidence that pragmatic cues were being used infelicitously (93% infelicitous sentences). Put another way, when the speaker used a size adjective, 75% of the time she was referring to an item not in a contrast set (see Appendix A). Yet listeners continued to anticipate a referent in a contrast set to a certain degree. In Experiment 3, we test whether explicitly attributing the infelicity to the speaker's lack of pragmatic competence might remove any lingering pragmatic inference.

4. Experiment 3

The goal of Experiment 3 is to conceptually replicate the effects observed in Experiment 2 (pre-registration available at osf.io/bt3ct/) and test whether the addition of a top–down cue facilitates the adjustment of pragmatic inferences. To achieve this, we added an explicit characterization of the speaker, modeled on Grodner and Sedivy (2011).

4.1. Method

4.1.1. Participants

Sixty-six participants from the student community at Vanderbilt University participated in this experiment in exchange for partial course credit or a gift card worth $20. Participants had normal or corrected-to-normal hearing and vision and were native speakers of North American English.

4.1.2. Procedure and materials

The experiment was identical to Experiment 3, except that an additional description of the speakers was included. The instructions were as follows:

> Welcome to the experiment! You will see 4 objects on the screen. You will hear instructions telling you which object to click on. Just click on the item that best matches what the speaker said. If you are not sure, just use your best guess.
>
> The instructions were recorded from an individual who had to direct a listener through a sequence of object configurations. The experiment is designed to test how effectively speakers are able to convey instructions by observing listener responses. [**This particular speaker has an impairment that causes language and social problems.**] Please

let the experimenter know if you have any questions. If you have no questions, you may begin!

The bracketed sentence in the instructions above was only presented to participants in the unreliable pragmatic context condition (the actual instruction was presented without the brackets in italics without bolding or underlining). This sentence was not shown to participants in the reliable pragmatic context condition.

4.2. Results

Fixations were coded in the same way as Experiments 1 and 2. A plot of the full time-course of target fixations by conditions can be seen in Fig. 2d. The main analysis of average target fixation proportions was pre-registered at osf.io/bt3ct/.

4.2.1. Average target fixation proportions

As in Experiment 2, the critical time window began 200 ms after the onset of the adjective (e.g., *big*) and ended 200 ms after the offset of the noun (e.g., *pickle*). The mean duration of this time window was 1,197 ms. The average proportions of target fixations within this time-window across pragmatic context and size contrast presence conditions are shown in Fig. 3d. Figure S2.1 shows average target fixation proportions for the adjective window (adjective onset + 200 ms to noun onset + 200 ms) and noun window (noun onset + 200 ms to noun offset + 200 ms) separately.

As in Experiment 2, the trial-level proportions of target fixation durations were first transformed using the empirical logit transformation and then analyzed in a multilevel linear regression. Pragmatic context and size contrast along with their interaction were entered as fixed effects with participants and items as random effects. We included random by-participants random slopes for size contrast and by-items random slopes for pragmatic context, size contrast, and their interaction. All fixed effects were deviation coded (size contrast condition: size contrast absent = −0.5, size contrast present = 0.5; pragmatic context condition: reliable pragmatic context = −0.5, unreliable pragmatic context = 0.5).

The full model estimates can be found in Table 5. There was a main effect of size contrast presence, such that participants made more target fixations when a size contrast was present ($t = 3.91$). The main effect of pragmatic context was not significant ($t = -0.32$), but there was a significant interaction between size contrast and pragmatic context ($t = -3.32$). A follow-up analysis setting the reliable pragmatic context condition as reference level revealed a large size contrast effect ($\beta = 0.20$, $SE = 0.04$, $t = 5.36$, $p < .001$) in the reliable pragmatic context condition, and a reduction of this effect in the unreliable pragmatic context condition ($\beta = -0.22$, $SE = 0.07$, $t = -3.32$, $p < .005$). Models for the adjective and noun windows separately are reported in Appendix S2. In the adjective window, the patterns are consistent with these main analyses. There is a significant effect of size contrast ($t = 5.23$), and it interacts significantly with pragmatic

Table 5
Experiment 3: Results of the linear mixed-effects model of empirical logit-transformed target fixation proportions over the critical time window (adjective onset + 200 ms to noun offset + 200 ms)

	β	*SE*	*t*	*p*
Fixed effects				
Intercept	−0.01	0.02	−0.55	0.59
Size contrast condition	0.09	0.02	3.91	< .001
Pragmatic context condition	−0.01	0.04	−0.32	0.75
Size contrast × Pragmatic context	−0.22	0.07	−3.32	< .005
	SD			
Random effects				
Participants (N = 66): Intercept	0.14			
Participants: Size contrast	0.04			
Items (N = 40): Intercept	0.07			
Items: Size contrast	0.03			
Items: Pragmatic context	0.03			
Items: Size contrast × pragmatic context	0.17			
Residual	0.56			
Observations: 2,640				

context (adjective window: $t = -3.29$). There were no significant effects in the noun window.

4.2.2. Binary target fixations

We analyzed the binary fixation data (a 0 or 1 for whether there was a target fixation in each 10 ms bin in each trial for each participant) in a multilevel logistic regression accounting for autocorrelation in fixations. The critical time window began 200 ms after the onset of the adjective and ended 1,200 ms later (the average duration of the adjective and noun).

Pragmatic context, size contrast, fix_{t-1}, and the interaction of pragmatic context and size contrast were entered as fixed effects with participants and items as random effects. A model with the maximal random slopes structure did not reach convergence. Per the procedure outlined above, we arrived at a model with by-participant random slopes for size contrast and by-item random slopes for pragmatic context and fix_{t-1}. All fixed effects of interest were deviation coded (size contrast condition: size contrast absent= −0.5, size contrast present = 0.5; pragmatic context condition: reliable pragmatic context = −0.5, unreliable pragmatic context = 0.5).

The full model estimates can be found in Table 6. Target fixations in the preceding time window significantly predicted target fixations (z = 265.17). There were no main effects of size contrast presence (z = 1.48) nor of pragmatic context (z = 0.07), but there was a significant interaction of size contrast and pragmatic context ($z = -2.16$). An analysis setting the reliable pragmatic context condition as the reference level indicates that the size contrast effect was present in the reliable pragmatic context condition (β = 0.13,

SE = 0.05, $z = 2.58$, $p = 0.01$) and significantly reduced in the unreliable pragmatic context ($z = -2.16$). Models for the adjective and noun windows separately are reported in Appendix S2. In the adjective and noun windows individually, there were significant effects of size contrast but in opposite directions (adjective window: $z = 5.66$, noun window: $z = -5.35$) but no significant interactions with pragmatic context (adjective window: $z = -1.15$, noun window: $z = -0.03$). There was also a significant effect of pragmatic context in the noun window (z = 2.17), suggesting that target fixations were overall more likely in the unreliable pragmatic context condition than the reliable one.[10]

4.2.3. Effect of exposure over time

In order to test if this modulation is robust over time, we examine the data patterns across the two halves of the experiment by adding experiment half (first = −0.5 vs. second = 0.5) as a predictor into the models (both predicting average proportions and binary fixations).[11] In each half of the experiment, participants encountered approximately 40 training trials, 20 test trials, and 90 fillers. Full summaries of these models are reported in Appendix S3. There was a main effect of size contrast presence for average fixation proportions ($t = 3.89$) but not binary fixations ($z = 1.44$) and a significant interaction of size contrast presence and pragmatic context condition ($t = -3.30$; $z = -2.16$). The size contrast effect on average fixation proportion decreased in the second half of the experiment ($t = -2.06$), and there was also a significant interaction between size contrast condition, pragmatic context condition, and experiment half ($t = 2.11$), such that the interaction between size contrast and pragmatic context was reduced in the second half of the experiment. This may be largely due to the fact that the proportion of target fixations

Table 6
Experiment 3: Results of the autoregressive generalized linear mixed-effects models of binary target fixations over the critical time window (adjective onset + 200 ms to adjective onset + 1,200 ms)

	β	*SE*	*z*	*p*
Fixed effects				
Intercept	−4.12	0.03	−149.30	< .001
Size contrast condition	0.05	0.03	1.48	0.14
Pragmatic context condition	0.003	0.05	0.07	0.95
Fix_{t-1}	8.78	0.03	265.17	< .001
Size contrast x Pragmatic context	−0.16	0.07	−2.16	0.03
	SD			
Random effects				
Participants (*N* = 66): Intercept	0.018			
Participants: Size contrast	0.001			
Items (*N* = 40): Intercept	0.004			
Items: Pragmatic context	0.000			
Observations: 319,440				

overall decreases in the second half of the experiment ($t = -2.53$; $z = -3.28$), perhaps as participants lose motivation or interest in the task.

4.2.4. Effect of the top–down cue

To test whether the top-down cue facilitated the modulation of pragmatic inference, in an exploratory analysis, we compared the results of Experiment 2 and Experiment 3 directly (see Appendix C for full model summaries). As in the main analyses of average proportion of target fixations and binary target fixations, there was a main effect of size contrast presence ($t = 6.26$; $z = 3.49$) and a significant interaction of size contrast and pragmatic context ($t = 3.95$; $z = -2.39$). The three-way interaction between size contrast, pragmatic context, and experiment (E2 = −0.5 vs. E3 = 0.5) was not significant ($t = -0.27$; $z = -0.19$). However, the effect of size contrast presence on average fixation proportion did vary by experiment ($t = -2.12$; the effect on binary fixations did not: $z = -1.47$), such that the main effect of size contrast was larger in Experiment 2 than in Experiment 3, perhaps because in Experiment 3, the size contrast effect is driven primarily by the reliable pragmatic context condition. This result is suggestive of an additional effect of the top–down cue beyond the distributional information alone, though a direct test of this hypothesis is needed to draw any firm conclusions, as the presence of the top–down cue is confounded with other potential discrepancies between Experiment 2 and Experiment 3 (e.g., populations from different universities) and only manifests in the fixation proportions (not the binary fixation measure).

4.3. Discussion

As in Experiments 1 and 2, during the interpretation of scalar adjectives, participants made more fixations to targets that were in a contrast set, replicating Sedivy et al., (1999). In Experiment 3, this contrast-contingent target preference was attenuated when participants were exposed to a speaker who used scalar adjectives in a contextually infelicitous manner and who was known to have social and language difficulties. This result constitutes, to our knowledge, the first pre-registered replication of Grodner and Sedivy's (2011) finding. When provided with a top–down cue and bottom-up evidence of a speaker's pragmatic infelicity, listeners do appear to modulate their contrastive inferences. Whether this modulation was more robust than in Experiment 2, where no top-down cues were present, remains an empirical question that awaits further experimentation.

5. General discussion

In a series of eye-tracking experiments, we explored the nature of evidence necessary for listeners to modulate contrastive scalar inferences, which are indexed by rapid, anticipatory looks to sets of size-contrasted items in the presence of a size adjective (Sedivy et al., 1999). In previous work (Grodner & Sedivy, 2011), listeners suppressed these inferences when faced with a speaker who was introduced to be pragmatically

incompetent and prone to linguistic errors. This finding points to the intriguing possibility that the rapid, anticipatory, eye movements are contextual in nature as opposed to heuristically determined based on the identity of an adjective. This malleability of the anticipatory mechanism may be beneficial in navigating the variability and ambiguity inherent in much of reference resolution. If online inferences are adapted to the likelihood with which the speaker formulates linguistic expressions optimally given the context, listeners can avoid making superfluous eye movements when the linguistic information (e.g., scalar adjective) is unlikely to lead to successful reference resolution. However, these findings leave open the question of what is the nature of the evidence used by the listener during this pragmatic modulation.

In this work, we examined the relative contributions of bottom-up and top-down information to comprehenders' ability to modulate pragmatic inferences based exclusively on the speaker-context. Listeners were exposed to speakers who either used size adjectives felicitously (e.g., "the big dog" when a small dog and a big dog were present) or infelicitously (e.g., "the big dog" when only one dog was present). In one experiment (Experiment 3), the speaker was also described as having a language impairment. When top–down and bottom–up evidence were present, listeners readily adapted their pragmatic inferences. We also found that while bottom-up cues alone were sufficient to trigger modulation of contrastive inferences, this required tremendous amounts of evidence and a seamless transition between the learning and testing environment.

5.1. The role of prior assumptions

A noteworthy result from this work was the surprising persistence of the contrast-contingent anticipatory fixations. Even when exposed to an overwhelming 93% infelicitous use of scalar adjectives, participants fixated the target significantly more when it was part of a contrast set. This effect was reduced relative to the case when instructions were mostly felicitous but the interaction effect was fragile (e.g., not always detectable in smaller time windows). The contrastive inference derived from size adjectives appears to be very resilient to new evidence that it is no longer valid in the present context. This persistence of the inference may be due to the expectations that participants bring to the experiment. In everyday language use, listeners experience mostly felicitous adjective use, as well as occasional instances of over-informativeness. Scalars in particular are unlikely to be used infelicitously (Brown-Schmidt & Konopka, 2011; Nadig & Sedivy, 2002; Ryskin et al., 2015; Tarenskeen, Broersma, & Geurts, 2015).[12] The instances of infelicitous adjective use experienced in the experimental session was not sufficient to fully counter a lifetime of experience with felicitous scalar adjectives. While the top–down characterization of the speaker as pragmatically "impaired" may have scaffolded the learning from distributional cues, perhaps by altering these prior assumptions, the (albeit smaller) contrastive inference still persisted in this context.

This view of contextualized pragmatic adaptation generates a number of testable hypotheses as to how a priori assumptions listeners might have about different classes

of adjectives influence the rate and degree of pragmatic modulation. We can generate at least two, opposing, predictions. First is that inferences linked to more variably used adjectives should be more susceptible to modulation. In particular, color adjectives are often used even when not strictly necessary for disambiguation (Brown-Schmidt & Konopka, 2011; Sedivy, 2003). Exposure to consistently infelicitous color adjectives may lead to complete suppression of a contrastive inference in the presence of a color adjective, because that inference is not as robust to begin with. Similarly, prosodic cues to contrast (e.g., L + H*; Ito & Speer, 2008; Watson et al., 2008) may be more variable by virtue of reflecting each speaker's realization of the intonational contour. Indeed, Kurumada, Brown, et al. (2014) find that listeners suppress their contrastive interpretations of the L + H* accent after being exposed to a speaker who uses contrastive pitch accenting inappropriately (for similar results see Roettger & Franke, 2019).

On the other hand, one could also argue that listeners' familiarity with more variable uses of color adjectives may slow down the process of pragmatic modulation. In other words, listeners may not learn easily that a given environment is more likely to contain infelicitous adjective use because some amount of infelicity is consistently present in everyday language use. On this account, the exposure to an infelicitous use of a color adjective (e.g., *Point to the orange car* when there is only one car in sight) is unlikely to trigger a large error signal and therefore leads to only moderate learning (e.g., Chang et al., 2006). This is in line with the previous observation (Pogue et al., 2016) that over-informative utterances, which are more prevalent in the input and hence less surprising, are less likely to trigger speaker-specific modulation of pragmatic assumptions compared to under-informative utterances. Understanding the interplay of these hypothesized processes will require a systematic investigation of the prior assumptions that listeners have about different types of contrastive cues and what role these play in the modulation process.

5.2. Semantic tuning and pragmatic generalization

Another question that arises from the current results concerns the extent to which they reflect a truly *pragmatic* process. The results are compatible with an account in which listeners *tune* their semantic notions of the scalars, "big" and "small." In response to an overwhelming amount of evidence that these words do not highlight contextual contrast, listeners may begin to assume that the semantic meaning, not the usage, is altered. To argue that the observed changes in eye movements are pragmatic in nature, one must show that the exposure to infelicitous scalars generalizes not only to the same lexical items but to other pragmatic uses of language by the same speaker.

Recent studies have taken a step toward addressing these questions by testing *generalization* of pragmatic infelicity across lexical items (Bott & Chemla, 2016; Pogue, Kurumada, & Tanenhaus, 2016). Pogue et al. (2016) first exposed listeners to two speakers who gave instructions in displays where scalar adjectives were necessary (e.g., a display with a large and a small chair, and two unrelated items). One speaker used scalar

adjectives informatively (e.g., "Click on the big chair"), while the other consistently failed to use scalars even when needed (e.g., "Click on the chair"). No explicit commentary about the pragmatic capabilities of the speakers was provided. Participants were subsequently asked to make an explicit judgment about which speaker was likely to have uttered a given sentence in a particular display. Pogue et al. found that participants were more likely to attribute under-informative color-modified expressions (e.g., "Click on the red car" in a display with a large and small red car) to the previously under-informative speaker. This finding lends support to the idea that listeners are able to track how different speakers use, or do not use, adjectives informatively and generalize this learning beyond directly experienced items. This generalization is not predicted if the speaker is modifying their semantic expectations for given lexical items directly observed in the input.

5.3. What are we learning? Inferences about the pragmatic competence of speakers

Consideration of the *generalization* of pragmatic modulation opens up a number of new research directions with respect to how listeners accommodate variability across speakers in their pragmatic language use. One of the core questions concerns what is being learned in the face of unexpected, seemingly uncooperative, use of linguistic elements, such as prenominal adjectives. Grodner and Sedivy (2011) concluded that listeners suspended contrastive inferences in the face of an unreliable speaker. However, we can sketch out at least three different classes of inferences that listeners might make.

One possibility, first laid out by Grodner and Sedivy (2011), is that the listener learns that the speaker is indifferent to the Gricean Cooperative Principle (Grice, 1975). If so, this predicts that the observed learning should extend to other domains of pragmatic language use, such as relevance of the speaker's comments, or the quantity of information that the speaker provides in other domains. A second possibility, also discussed by Grodner and Sedivy (2011), is that the learning is specifically about, say, prenominal modifiers and their mapping to context, such that what is learned is distributional information about how this speaker uses this adjective type in context. As a variant of this view, a learner might assume that the speaker's idiosyncrasy is restricted to adjective use and does not extend to other types of modifiers such as quantifiers.

A third class of possibilities is that the listener preserves the assumptions that the speaker is Gricean, but infers that their perspective or assumed common ground is distinct and/or mismatched. For example, the listener may assume that the speaker is seeing something different (e.g., one of the objects visible to the listener is occluded in the speaker's display) or having a different experience of the world than the addressee. Perhaps, speakers may be drawing a contrast between something being currently discussed and something experienced in the past (e.g., "This movie is much better.") on the mistaken assumption that their interlocutor recalls what they had previously discussed. In these cases, infelicity is not diagnostic of the speaker's overall adherence to Gricean maxims. Given that speech acts and implicature usually arise based on the assumption that speakers flout, but do not simply ignore Gricean

Principles, it is plausible that listeners may first search for an incidental (as opposed to speaker-internal) explanation for an observed sign of pragmatic infelicity. It is, however, possible that a particular sign of pragmatic infelicity (e.g., consistent failure in perspective-taking) can be indicative of more pervasive difficulties across levels of pragmatic language use (e.g., formulating effective deictic expressions, scalar implicature). Thus, testing pragmatic adaptation and generalization can help elucidate listeners' underlying beliefs about whether and when listeners may fail to observe Gricean Maxims as well as about how likely these failures will recur in the speaker's subsequent language use.

6. Conclusion

Grice's Cooperative Principle states that human listeners are unlikely to assume that the speaker is being unreliable or infelicitous. Exploiting this tendency provides the speaker with a variety of powerful means to convey their intentions and social meanings (e.g., jokes, lies, sarcasm). The persistence of contrastive inferences observed in all the experiments reported here is in line with this basic assumption that speakers rarely misuse scalar modifiers. Nevertheless, when given an overwhelming amount of evidence (Experiment 2) and, when also given an explicit instruction to counteract the usual expectations (Experiment 3), the listener can flexibly modify their online language comprehension so as not to be led astray. What emerges from these observations is an adaptive mechanism of language comprehension that can integrate expectations based on prior experiences as well as distributional statistics from the recent exposure. An important goal for future research will be to investigate how these different types of information are evaluated and combined to support efficient and effective pragmatic communication.

Acknowledgments

This material is based on work supported by the National Science Foundation Grants NSF 12-57029 and NSF 15-56700 to Sarah Brown-Schmidt. We thank Sarah Bibyk for help with stimulus recording.

Notes

1. A secondary goal of Experiment 1 was to test another question relating to the modulation of online interpretations of scalar adjectives based on prosodic cues. For the sake of clarity, and because those results were equivocal, we do not discuss those data in the main text of this paper, but a summary can be found in the supplemental material S1 available online.

2. The sets of sequential instructions paired with the same display were originally intended to be minimally different from another set of training trials aimed at testing contrastive prosody. See Supplementary Material S1 for more information.
3. The training phase consisted of 12 infelicitous trials and the test phase contained 10 infelicitous trials (No Contrast trials), for a total of 22 infelicitous trials out of a grand total of 62 trials across training and test (22/62 = 0.355).
4. As opposed to Experiment 1, there was no cover story about where the sentences came from. We reasoned that the background story might bias listeners to invoke unique assumptions about the properties of child-directed speech (e.g., frequent non information seeking questions with pedagogical intentions). We chose to keep the focus here on examining whether listeners learn about the pragmatic competence of a speaker in a situation where the default assumption is that the speaker will be engaging in cooperative interactions.
5. There were three types of reliable pragmatic context condition secondary exposure trials. (1) One-third (60) of the non-scalar exposure trial instructions included a non-scalar adjective, such as, "Click on the glazed doughnut.", where the display contained a target item (e.g., a glazed doughnut), a contrasting item of the same category as the target (e.g., a powdered doughnut), and two distractors (e.g., a big muffin pan and a small saddle). (2) Another third (60) of the non-scalar exposure trials consisted of a bare noun phrase (e.g., "Click on the cake.") and a singleton target item (e.g., a big cake), along with three distractors which included a pair of size-contrasted items (e.g., a toothbrush, a big burger, and a small burger). (3) The remaining third (60) of the non-scalar exposure trials also consisted of a bare noun phrase (e.g., "Click on the couch"), and a singleton target item (e.g., a small couch), along with three distractors which included a pair of items in a non-scalar contrast (e.g., a big cheese plate, a glazed doughnut, and a powdered doughnut). The purpose of the paired distractors in these latter two trial types was to mitigate participants learning that a pair of related images in a display will always include the target.
6. There were three types of unreliable pragmatic context condition secondary exposure trials. (1) One third (60) of the non-scalar exposure trial instructions included a non-scalar adjective, such as "Click on the glazed doughnut." The display contained a target item (e.g., a glazed doughnut), but no contrasting item of the same category as the target. Instead, there were three distractors (e.g., a medium rug, a big muffin pan, and a small saddle). (2) Another third (60) of the non-scalar exposure trials consisted of a globally ambiguous bare noun phrase (e.g., "Click on the cake.") and a target item in a pair (e.g., a big cake and a small cake), along with a pair of size-contrasted distractors (e.g., a big burger and a small burger). (3) The remaining third (60) of the non-scalar exposure trials also consisted of a globally ambiguous bare noun phrase (e.g., "Click on the couch"), and a target item in a pair (e.g., a small couch and a big couch), along with two distractors which included a pair of items in a non-scalar contrast (e.g., a glazed doughnut and a powdered doughnut).

7. This duration differs from the one in Experiments 1 and 2 because the auditory stimuli for this experiment were expanded and recorded by a different speaker (the first author).
8. Due to a counterbalancing error, items only appeared in one of the two contrast conditions in each pragmatic context condition. In other words, for a given item, contrast condition and pragmatic context are perfectly correlated. In order for the model to be identified, we cannot include both of these as random by-item slopes. We chose pragmatic context to be included as the random by-item slope because model comparison indicated that this provided a better fit.
9. We also modeled exposure over time as continuous using trial order. We observed no significant interactions of trial order with the other predictors ($ps > 0.05$). Note that the order of trials was random for each participant so the amount of exposure to the pragmatic context differs between participants at each trial order position, which may have dampened any potential effects.
10. This pattern was not predicted, but we can speculate that perhaps participants in the reliable pragmatic context condition, having identified the target earlier, begin to look away, whereas those in the unreliable pragmatic context condition arrive at the target later, and so, in this later time window, those in the unreliable condition are more likely to still look at the target than those in the reliable condition.
11. We also modeled exposure over time as continuous using trial order. We observed no significant three-way interactions of trial order with the other predictors ($ps > 0.05$). Note that the order of trials was random for each participant, so the amount of exposure to the pragmatic context differs between participants at each trial order position, which may have dampened any potential effects.
12. In real-world settings, outside of the laboratory, scalars can be used when no contrast set is visually co-present (e.g., "Hand me that big coffee mug."); this may not constitute an infelicitous use of the modifier because the contrast might be derived from prior experiences with items similar to the referent (e.g., this particular coffee mug is over-sized relative to all previously seen mugs).

References

Arnold, J. E., Kam, C. L. H., & Tanenhaus, M. K. (2007). If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 914–930. https://doi.org/10.1037/0278-7393.33.5.914.

Arts, A., Maes, A., Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics*, *43*, 361–374. https://doi.org/10.1016/j.pragma.2010.07.013.

Aslin, R. N., & Newport, E. L. (2014). Distributional language learning: Mechanisms and models of category formation. *Language Learning*, *64*, 86–105. https://doi.org/10.1111/lang.12074.

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (2015). Fitting linear mixed-Effects models using {lme4}. *Journal Of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01.

Belke, E. (2006). Visual determinants of preferred adjective order. *Visual Cognition*, *14*, 261–294. https://doi.org/10.1080/13506280500260484.

Boersma, P., & Weenink, D. (2016). Praat: doing phonetics by computer [Computer program]. Version 6.0.56 [retrieved June 20, 2019]. Available at http://www.praat.org/.

Bott, L., & Chemla, E. (2016). Shared and distinct mechanisms in deriving linguistic enrichment. *Journal of Memory and Language*, *91*, 117–140. https://doi.org/10.1017/CBO9781107415324.004.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences, USA*, *105*(38), 14325–14329.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2013). Real-world objects are not represented as bound units: Independent forgetting of different object details from visual memory. *Journal of Experimental Psychology: General*, *142*(3), 791–808.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, *11*, 717–726. https://doi.org/10.1016/S0022-5371(72)80006-9.

Brown-Schmidt, S., & Konopka, A. E. (2011). Experimental approaches to referential domains and the on-line processing of referring expressions in unscripted conversation. *Information*, *2*, 302–326.

Brown-Schmidt, S., & Tanenhaus, M. K. (2006). Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language*, *54*, 592–609. https://doi.org/10.1016/j.jml.2005.12.008.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*, 234–272. https://doi.org/10.1037/0033-295X.113.2.234.

Cho, S. J., Brown-Schmidt, S., & Lee, W. Y. (2018). Autoregressive generalized linear mixed effect models with crossed random effects: An application to intensive binary time series eye-tracking data. *Psychometrika*, *83*, 751–771.

Crain, S., & Steedman, M. (1985). On not being led up the garden path: the use of context by the psychological parser. In D. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural Language Parsing: psychological, computational, and theoretical perspectives* (pp. 320–358). Cambridge, UK: Cambridge University Press. https://doi.org/10.1080/10643389.2012.728825.

Creel, S. C. (2014). Preschoolers' flexible use of talker information during word learning. *Journal of Memory and Language*, *73*, 81–98. https://doi.org/10.1016/j.jml.2014.03.001.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: the role of talker variation in lexical access. *Cognition*, *106*, 633–664. https://doi.org/10.1016/j.cognition.2007.03.013.

Creel, S. C., & Tumlin, M. A. (2011). On-line acoustic and semantic interpretation of talker information. *Journal of Memory and Language*, *65*, 264–285. https://doi.org/10.1016/j.jml.2011.06.005.

Davies, C., & Katsos, N. (2013). Are speakers and listeners 'only moderately Gricean'? An empirical response to Engelhardt et al. (2006). *Journal of Pragmatics*, *49*, 78–106. doi:10.1016/j.pragma.2013.01.004.

Donnellan, K. S. (1966). Reference and definite descriptions. *The Philosophical Review*, *75*, 281–304.

Eich, E. (1985). Context, memory, and integrated item/context imagery. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 764–770.

Engelhardt, P. E., Bariş Demiral, Ş., & Ferreira, F. (2011). Over-specified referring expressions impair comprehension: An ERP study. *Brain and Cognition*, *77*, 304–314. https://doi.org/10.1016/j.bandc.2011.07.004.

Fine, A. B., & Florian Jaeger, T. (2013). Evidence for implicit learning in syntactic comprehension. *Cognitive Science*, *37*, 578–591. https://doi.org/10.1111/cogs.12022.

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE*, *8*, e77661. https://doi.org/10.1371/journal.pone.0077661.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: on land and underwater. *British Journal of Psychology*, *66*, 325–331.

Grice, P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics*. Vol. *3* (pp. 41–58). New York: Academic Press.

Grodner, D. J., & Sedivy, J. C. (2011). The effect of speaker-specific information on pragmatic inferences. In E. A. Gibson & N. J. Pearlmutter (Eds.), *The processing and acquisition of reference*. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/9780262015127.001.0001.

Hallett, P. E. (1986). Eye movements. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance* (pp. 10.1–10.112). New York: Wiley.

Ito, K., & Speer, S. R. (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language*, *58*, 541–573. https://doi.org/10.1016/j.jml.2007.06.013.

Kleiner, M., Brainard, D., & Pelli, D. (2007). "What's new in Psychtoolbox-3?" Perception 36. ECVP Abstract Supplement.

Kurumada, C., Brown, S., Bibyk, S., Pontillo, D., & Tanenhaus, M. K. (2014). Rapid adaptation in online pragmatic interpretation of contrastive prosody. Proceedings of the 37th Annual Meeting of the Cognitive Science Society.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0-33 [retrieved July 2, 2019]. Available at https://CRAN.R-project.org/package=lmerTest.

Levelt, W. M. (1989). *Speaking : from intention to articulation* (p. c1989). Cambridge, MA: MIT Press.

Mulligan, N. W. (2011). Conceptual implicit memory and environmental context. *Consciousness and Cognition*, *20*, 737–744.

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of perspective-taking constraints in children's on-line reference resolution. *Psychological Science*, *13*, 329–336. https://doi.org/10.1111/1467-9280.00460.

Olson, D. R. (1970). Language and thought: Aspects of a cognitive theory of semantics. *Psychological Review*, *77*, 257–273.

Osgood, C. E. (1971). Where do sentences come from? In D. D. Steinberg & L. A. Jakobovits (Eds.), *Semantics: An interdisciplinary reader in philosophy, linguistics and psychology* (pp. 497–529). Cambridge, MA: Cambridge University Press.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, *27*, 89–110.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442. https://doi.org/10.1163/156856897X00366.

Pogue, A., Kurumada, C., & Tanenhaus, M. K. (2016). Talker-specific generalization of pragmatic inferences based on under- and over-informative prenominal adjective use. *Frontiers in Psychology*, *6*, 1–18. https://doi.org/10.3389/fpsyg.2015.02035.

Powell, M. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. Report DAMTP 2009/NA06, University of Cambridge.

Roberts, C. (2003). Uniqueness in definite noun phrases. *Linguistics and Philosophy*, *26*, 287–350.

Roettger, T., & Franke, M. (2019). Evidential strength of intonational cues and rational adaptation to (un-) reliable intonation.

Rubio-Fernández, P. (2016). How redundant are redundant color adjectives? an efficiency-based analysis of color overspecification. *Frontiers in Psychology*, *7*, 153. https://doi.org/10.3389/fpsyg.2016.00153.

Ryskin, R. A., Benjamin, A. S., Tullis, J., & Brown-Schmidt, S. (2015). Perspective-taking in comprehension, production, and memory: An individual differences approach. *Journal of Experimental Psychology: General*, *144*, 898–915. https://doi.org/10.1037/xge0000093.

Saffran, J., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *621*(35), 606–621.

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, *32*, 3–23.

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

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568. https://doi.org/10.1016/j.cognition.2007.06.010.

Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: a review and meta-analysis. *Psychonomic Bulletin & Review*, *8*, 203–220.

Tarenskeen, S., Broersma, M., & Geurts, B. (2015). Overspecification of color, pattern, and size: Salience, absoluteness, and consistency. *Frontiers in Psychology*, *6*, 1703. https://doi.org/10.3389/fpsyg.2015.01703.

Van Der Wege, M. M. (2009). Lexical entrainment and lexical differentiation in reference phrase choice. *Journal of Memory and Language*, *60*(4), 448–463. https://doi.org/10.1016/j.jml.2008.12.003.

Watson, D. G., Tanenhaus, M. K., & Gunlogson, C. A. (2008). Interpreting pitch accents in online comprehension: H* vs. L+H*. *Cognitive Science*, *32*, 1232–1244. https://doi.org/10.1080/03640210802138755.

Wells, J. B., Christiansen, M. H., Race, D. S., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology*, *58*, 250–271. https://doi.org/10.1016/j.cogpsych.2008.08.002.

Wonnacott, E., Newport, E. L., & Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology*, *56*(3), 165–209. https://doi.org/10.1016/j.cogpsych.2007.04.002.

Yoon, S. O., & Brown-Schmidt, S. (2014). Adjusting conceptual pacts in three-party conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 919–937. https://doi.org/10.1037/a0036161.

Yurovsky, D., Fricker, D. C., Yu, C., & Smith, L. B. (2014). The role of partial knowledge in statistical word learning. *Psychonomic Bulletin & Review*, *21*(1), 1–22. https://doi.org/10.3758/s13423-013-0443-y.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Appendix S1. Additional conditions in Experiment 1

Appendix S2. Figures and analyses split into adjective and noun time windows (Experiments 1, 2, and 3)

Appendix S3. Analyses from Experiments 2 and 3 with experiment half (first vs. second) as a predictor

Appendix A

Table A1
Summary of numbers of trials in different categories and percentage of infelicitous trials across experiments. Asterisks indicate infelicitous trials

		Reliable pragmatic context		Unreliable pragmatic context	
Experiment 1a-b		Adjective present (e.g., "Show me the big dog")	Adjective absent (e.g., "Show me the dog")	Adjective present (e.g., "Show me the big dog")	Adjective absent (e.g., "Show me the dog")
Training phase	Size contrast present	12	0	0	0
	Size contrast absent	0	0	12*	0
Test phase	Size contrast present	10	0	10	0
	Size contrast absent	10*	0	10*	0
Filler trials (presented during test phase)	Size contrast absent	0	10	0	10
	Semantic contrast present	10	0	10	0
	Semantic contrast absent	0	10	0	0
Total count trials:		62		62	
Total count (and %) infelicitous trials:		10 (16.13%)		22 (35.48%)	

		Reliable pragmatic context		Unreliable pragmatic context	
Experiments 2 & 3		Adjective present (e.g., "Show me the big dog")	Adjective absent (e.g., "Show me the dog")	Adjective present (e.g., "Show me the big dog")	Adjective absent (e.g., "Show me the dog")
Primary exposure trials	Size contrast present	40	0	0	40*
	Size contrast absent	0	40	40*	0
Secondary exposure trials	Semantic contrast present	60	0	0	60*
	Semantic contrast absent	0	60	60*	0
	Size contrast present	0	0	0	60*
	Size contrast absent	0	60	0	0
Test trials	Size contrast present	20	0	20	0
	Size contrast absent	20*	0	20*	0
Total count trials:		300		300	
Total count (and %) infelicitous trials:		20 (6.67%)		280 (93.33%)	

Appendix B

To assess whether the Infelicitous training trials in Experiment 1 were in fact perceived as infelicitous by the participants, we collected ratings of how "natural" the sentences sounded paired with the corresponding displays.

Participants

We collected data from 49 participants recruited through the Amazon Mechanical Turk platform. Participants were told that they should only participate if they are native speakers of American English who started learning English at age 3 or earlier.

Procedure and materials

Stimuli were presented using the Qualtrics online survey platform. After reading through the informed consent, participants read the following instructions: "For this hit, you will be listening to some sentences, looking at pictures, and judging if they make sense together. Little Joe and his mom are playing a game on mom's computer. *Your job is to listen to mom's speech and 1) click on pictures that she is talking about; and 2) rate whether they made sense to you. Let's start with an example. Are you ready?*"

The instructions were followed by a practice trial and then 12 trials. On each trial, participants heard two auditory instructions and saw a display consisting of four pictures (different colored shapes). The audio instructions and accompanying visual stimuli were taken from the training task used in Experiment 1. Each participant was randomly assigned to see either the 12 Reliable trials or the 12 Unreliable trials. During each trial, participants followed 5 steps that were indicated on the screen. 1. Listen to the first sentence by clicking play on the sound file (e.g., "Show me the blue triangle). 2. Click on a picture (e.g., the blue triangle) 3. Listen to the second sentence (e.g., "Now, show me the small square."). 4. Click on a picture (e.g., the small square). 5. Rate on a 5-point scale how good the second sentence was (1 = completely odd, 2 = relatively odd, 3 = just fine, 4 = relatively good, 5 = perfectly good).

Before finishing the study, participants answered two final questions: "Did you notice anything strange about the things the speaker said? Say 'No' if not." and "This survey is part of our communication research. If we told you that some people were listening to a speaker who has a language impairment and difficulty in saying the right thing, what would you say about the speaker you had? (1 = No Sign of impairment at all – 5 = Clear sign of impairment)."

Results

Participants listening to reliable materials rated them as relatively good to perfectly good ($M = 4.76$, $SD = 0.37$). Those listening to unreliable materials also did not notice any problems with them ($M = 4.65$, $SD = 0.44$). A two-tailed t-test indicated that there was no significant effect of Reliability condition ($t(47)=0.95$, $p = 0.35$). Participants do not appear to perceive the infelicitous sentences in Experiment 1 as such.

Appendix C

Table C1
Results of the linear mixed-effects model of empirical logit-transformed target fixation proportions for adjective + noun time window (adjective onset + 200 ms to noun offset + 200 ms) comparing experiments 2 and 3

Predictors	*b*	*SE*	*t*	*p*
Intercept	−0.01	0.02	−0.56	0.575
Size contrast	−0.04	0.03	−1.16	0.249
Pragmatic context	0.13	0.02	6.26	< **0.001**
Expt. 2 vs. Expt. 3	−0.01	0.03	−0.21	0.834
Size contrast × pragmatic context	−0.21	0.05	−3.95	< **0.001**
Pragmatic context × Expt. 2 vs. Expt. 3	0.04	0.06	0.79	0.432
Size contrast × Expt. 2 vs. Expt. 3	−0.07	0.03	−2.12	**0.036**
Size contrast × pragmatic context × Expt. 2 vs. Expt. 3	−0.02	0.07	−0.27	0.789
Random effects variances				
Participants: Intercept	0.02			
Participants: Size contrast	0.00			
Items: Intercept	0.00			
Items: Pragmatic context	0.01			
Items: Expt. 2 vs. Expt. 3	0.00			
Items: Pragmatic context × Expt. 2 vs. Expt. 3	0.00			
Residual	0.31			
Observations	5,160			

Table C2
Results of autoregressive generalized linear mixed-effects model of binary target fixations over the critical time window (adjective onset + 200 ms to adjective onset + 1,400 ms) comparing experiments 2 and 3

Predictors	*b*	*SE*	*z*	*p*
Intercept	−4.10	0.02	−193.22	< **0.001**
Size contrast	0.08	0.02	3.49	< **0.001**
Pragmatic context	−0.02	0.03	−0.74	0.461
Expt. 2 vs. Expt. 3	−0.04	0.03	−1.27	0.204
Fix_{t-1}	8.79	0.02	368.73	< **0.001**
Size contrast × Pragmatic context	−0.15	0.06	−2.39	**0.017**
Size contrast × Expt. 2 vs. Expt. 3	−0.07	0.05	−1.47	0.143
Pragmatic context × Expt. 2 vs. Expt. 3	0.05	0.06	0.84	0.398
Size contrast × Pragmatic context × Expt. 2 vs. Expt. 3	−0.02	0.09	−0.19	0.847
Random effects variances				
Participants: Intercept	0.02			
Participants: Size contrast	0.00			
Items: Intercept	0.01			
Observations	621,840			