IS RETERIEVAL FLUENCY A HEURISTIC IN AUDIENCE DESIGN?

Kieran J. O'Shea, Caitlyn R. Martin and Dale J. Barr

–

EXPERIMENT 2

Summary:
This study attempts to test the *retrieval fluency* hypothesis for referential encoding: that attending to a referent with the goal of referential encoding elicits retrieval of previous referential expressions used for a particular referent, and that speakers use the strength/fluency of these memory signals as a cue to their informational adequacy in the current communicative situation. We derive the assumption that memory signals correlate with informational adequacy from the *encoding specificity principle* of episodic memory (Tulving & Thomson, 1973), whereby the strength of a memory signal is a function of the similarity between encoding and retrieval contexts. The retrieval fluency hypothesis assumes that speakers who experience strong retrieval fluency associated with a particular expression in a particular context will engage in less assessment of its contextual adequacy. It follows that speakers experiencing strong fluency will be less likely to notice a change in the communicative situation that invalidates the informational adequacy of the retrieved expression, leading them to misspecify referents at a higher rate than speakers who experience low fluency.

This is our second attempt to evaluate the hypothesis. For theoretical background, please see the documentation for the original study (https://osf.io/4akir/).

In our original study, speakers entrained upon using particular expressions to describe particular target letters ('the *little* L') in a recurring context (i.e., arrangement of letters). We attempted to manipulate retrieval fluency by varying a communicatively-irrelevant cue over the course of a block of training trials: whether the configuration of (irrelevant) distractors letters in a display varied their positions and colours across instances, or whether they remained relatively constant. At the end of a block of trials, there was a single test trial that reflected the prototype arrangement that was experienced during training. Our reasoning was that although speakers would have encoded the referent the same number of times in both conditions, they should experience stronger retrieval fluency for the established description in the Low Variability condition than in the High Variability condition. There was a critical change in the test trial that invalidated that established description (e.g., the bigger L was removed from the display and replaced with a different letter, such that the modifier 'little' was no longer necessary to designated the target). Our prediction was that speakers should fail to notice the change more often in the High Variability condition than in the Low Variability condition, and thus be more likely to misspecify the referent. However, this prediction was not borne out: the overall misspecification rate (N = 36 speakers) was similar in both the Low (18%) and High (17%) Variability conditions ($z = .77$, $p = .22$).

We did observe an unexpected result, however. One additional factor that we manipulated was whether ignoring the contextual change at test would lead to 'overspecification' (referring to the sole L as 'the little L') versus underspecification (referring to the smaller of two Ls as 'the L'). Unlike previous studies, we found that speakers were more likely to underspecify a referent (26%) than overspecify the referent (9%) in the test trial ($z = 4.47$, $p < .01$).

In this pre-registration document we outline details of our second attempt to document the existence of the retrieval fluency heuristic in language production. We have altered the stimuli, procedure and test phase of our experiment.

METHOD

Participants

Thirty-six Native English speakers from the University of Glasgow will participate in our study. Subjects will either be offered four 'participation credits' (course credits) for taking part in the study or will receive payment of £6 for their participation. Subjects will give written informed consent before beginning the experiment and will be fully debriefed afterwards. Our procedures fully comply with the ethical code of conduct of the British Psychological Association.

Experimental Setup and Task

Similarly to our first study, the experiment will be interactive with the participant playing the role of the 'Director' (the speaker) and the experimenter playing the role of the 'Matcher' (the listener). The Director and the Matcher will sit in different areas of the testing room and look at separate computer monitors throughout the experiment. Both will be seated facing in opposite directions so that they are unable to see each other's display. In each trial, the Director will be asked to describe a target object which appears highlighted on their monitor to the Matcher. The Matcher will then identify this object on his/her own screen and select it using a computer mouse. The target object will appear on the Director's screen within a grid among other 'filler' objects (see Figure 1). The Director will be informed that in each trial the listener will have the same objects on their monitor but that they may be arranged in a different format to the grid that appears on their screen (please see Appendix 1 for the instruction sheet which will be given to the participant).

Target and Competitor objects were normed beforehand by 68 Native English speaking volunteers. A number of items were updated or replaced based on our norming feedback. 4 entirely new stimuli pairs were added (please see Appendix 2 for a complete list of the Target and Critical objects used).

Design

There are two factors in the design, *Direction of Shift* (Singleton-to-Contrast versus Contrast-to-Singleton) and *Training-Test Consistency* (Consistent versus Inconsistent), forming a full-factorial 2x2 within-participant design.

Direction of Shift:

In each sequence, the target will appear with a "critical" object, whose identity forms the factor of *Direction of Shift*. This factor refers to whether speakers entrained upon descriptions for a target object in a context where modifiers were not required ("the car") and then tested in a context requiring a modifier ("the family car") or vice versa. In the former condition (Singleton-to-Contrast; see Figure 1), the critical object during the training phase was a non-competitor object, leading directors to entrain upon a bare noun phrase ("the car"). We refer to this non-competitor object as "the foil" as it was chosen to be perceptually similar (in shape and colour) to the competitor object used in the test trial, but clearly represented a different category of object (e.g., the computer mouse, which has the same shape and colour as the competitor car). For the test trial in this condition, the foil would be replaced with the competitor object, which was another object from the same category as the target (e.g., a car) but differing in some critical way (e.g., a sports car), thus requiring speakers to modify their descriptions ("the car" -> "the family car").

In the Contrast-to-Singleton condition (see Figure 2 for an example) this order will be reversed: the critical object during training will be the competitor (e.g., the “sports car”), leading speakers to entrain upon a modified expression during training. At test, the competitor will then be replaced with the foil, such that participants will be able to simplify their description of the target item (“the family car” -> “the car”).

In addition to the critical item, each display will contain other *filler* items. The relation of the arrangement of these items during training to their arrangement during test forms the critical manipulation of *Training-Test Consistency*.

The trials will be presented in blocked sequences with the order counterbalanced across participants. Unlike our previous experiment, all training trials presented will have a relatively stable arrangement during training; what we varied instead in this experiment was whether that training arrangement was similar or dissimilar to the arrangement at test. In the Consistent condition (previously the ‘Low Variability Context’ in Experiment 1) the configuration of items in the display at training will be highly similar to the configuration presented at test. In the Inconsistent condition, the configuration of items in the training displays will be highly dissimilar to test. The reasoning was that in attempting to referentially encode the target item at test, speakers in the Consistent condition should experience a stronger memory signal associated with the expression used in training, based on the higher similarity between training and test arrangements.

Across all training trials, the positions of the target and filler items was fixed, with the exception that the position of the critical item (Competitor or Foil) could swap with the position of one of the filler items. This was to prevent speakers from learning where to look to check for the presence of a competitor.

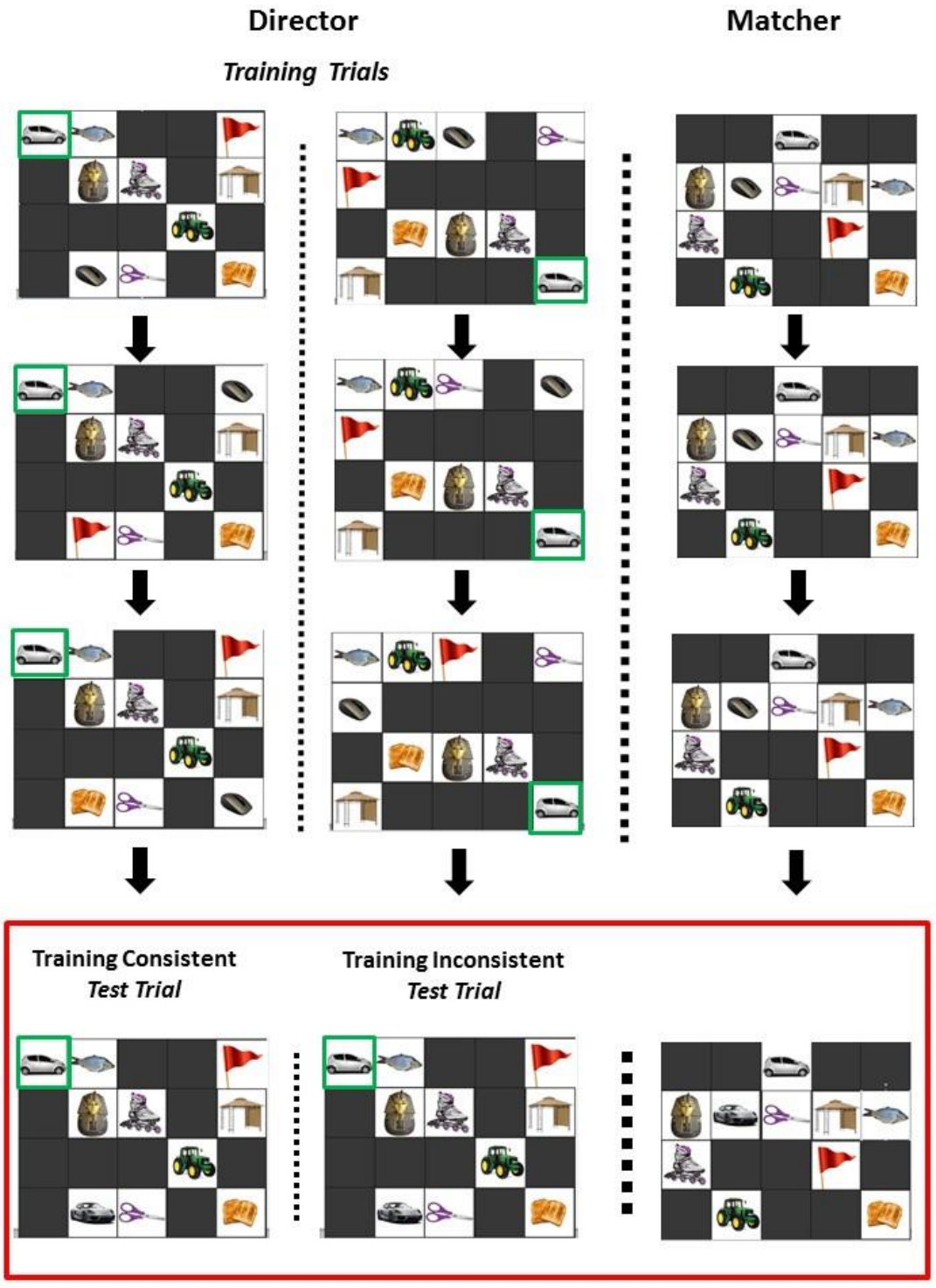

Figure 1 - Example of stimuli in Singleton -> Contrast condition:
Example of 3 training trials followed by a test trial. The column on the left shows the Director's view of the stimuli where the test trial is consistent with the arrangement in the training phase - *Training Consistent* condition. The middle column shows the alternative *Training Inconsistent* condition. The training trials highlight the target object in a green rectangle – in this case the "the car". The test trial presents participants with the target object "the car" again, but unlike the training trials it also introduces a new "sport scar" object. This may prompt the Director to underspecify their description of the target object to the Matcher ("select the car"). The training trials also present the "computer mouse" which acts as a foil for the "sports car" shown in the test phase. Note that the Matcher's view remains fixed throughout with the "sports car" replacing the "computer mouse" in the test trial.

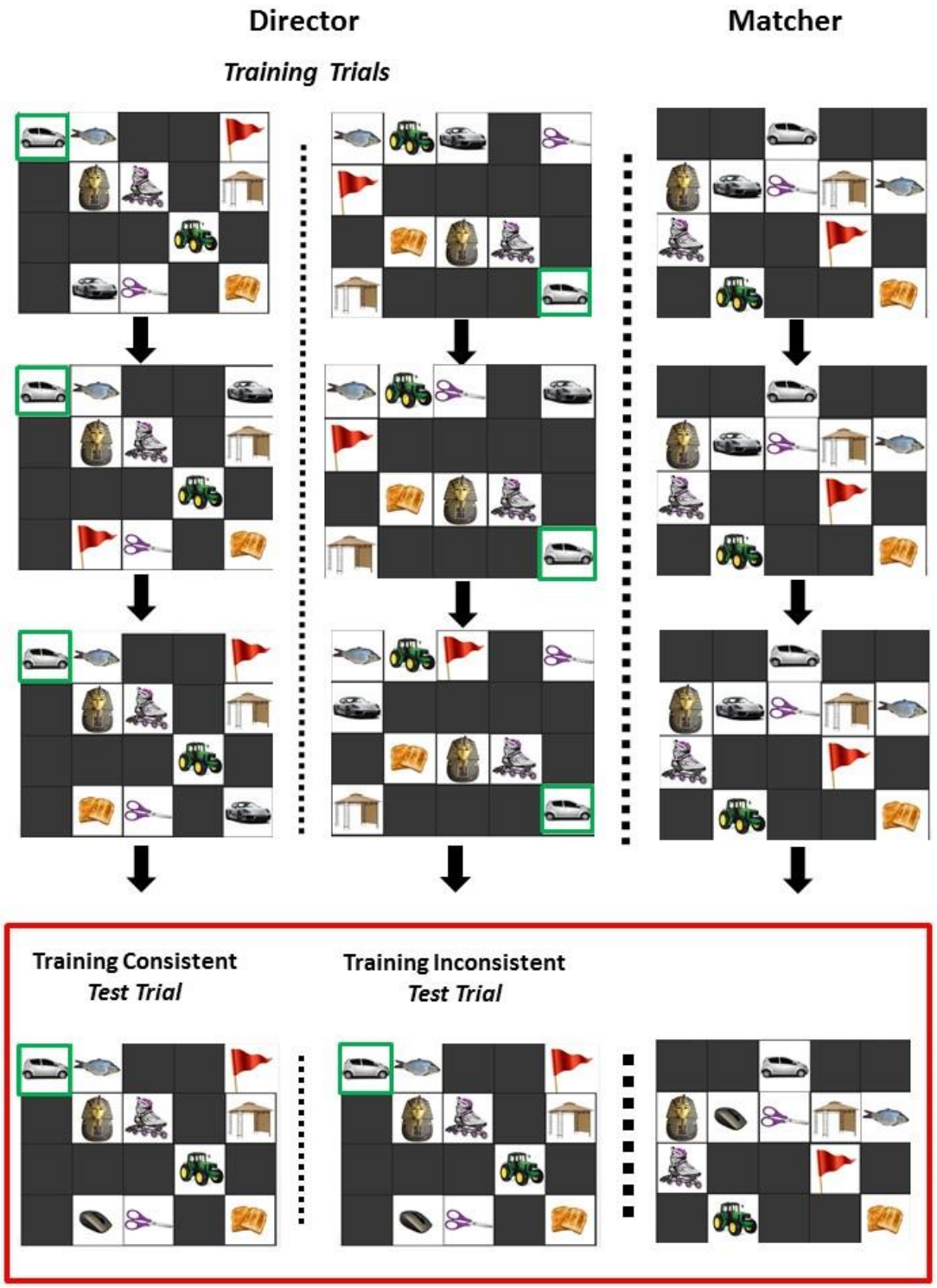

Figure 2 - Example of stimuli in Contrast -> Singleton condition:
Example of 3 training trials followed by a test trial. The column on the left shows the Director's view of the stimuli where the test trial is consistent with the arrangement in the training phase - *Training Consistent* condition. The middle column shows the alternative *Training Inconsistent* condition. The training trials highlight the target object in a green rectangle – in this case the "the car". Note that the competitor object – "the sports car" is also present in the grid. The test trial presents participants with the target object "the car" again, but unlike the training trials the foil object - "the computer mouse" has replaced the competitor. This may prompt the Director to overspecify their description of the target object to the Matcher ("select the *family* car"). Note that the Matcher's view remains fixed throughout with the "computer mouse" replacing the "sports car" in the test trial.

Materials
The parameters governing each display in the experiment are defined in tables within the sqlite3 database EESP3.db in the github repository at https://github.com/dalejbarr/EESP3.

Each display consists of a five-by-four grid containing objects of different size and colour (Figure 1). The experiment will contain 48 "sequences" of trials, each consisting of a number of training trials followed by a single test trial. (Throughout this document, we use the term "sequence" to refer to the collection of training and test trials all associated with a single target/competitor/foil triplet.) Each sequence appeared an equal number of times in all four conditions of the 2x2 designs, counterbalanced across participants.

For each sequence, the number of training trials was randomly selected, with a range from six to nine. The motivation for varying training sequence length is to make the occurrence of the test trial unpredictable. Given these parameters, each experimental session may contain between 336 (7 x 48) and 480 (10 x 48) trials.

For each sequence, seven to ten filler items were randomly chosen from a database of stimulus images. The displays were then checked manually to ensure that the filler items were sufficiently dissimilar to the target so as not to influence descriptions of the target.

Apparatus
The experimental stimuli will be presented on a 19" LCD Dell desktop computer monitor (4:3 aspect ratio, resolution 1024 x 768 pixels). A microphone will be placed above the participant's computer monitor to record their descriptions of the target letter for each trial. The audio will be tagged using Audacity 2.0.6 software. Eye movements will be recorded during each trial using an Eyelink 1000 (SR Research) eye tracker (sampling rate set to 250Hz).

Sequencing of Trials
Each of the two blocks of trials (in which 24 sequences were presented) was further divided up into six sub-blocks, each of which contained the training and test trials for four sequences. The motivation for this was to have all of the training/test trials for a given block in relative proximity within the sequence, but to also make the position of the test trial for each sequence unpredictable. Trials for the first five of the six sub-blocks were sequenced as follows. First, the last fifteen trials of the sub-block were created, consisting of (a) the four test trials from the four sequences, at serial positions three, seven, eleven, and fifteen within the fifteen trial sequence; (b) the last training trial for three of the four sequences, with one at position four or five (randomly chosen), another at position eight or nine (randomly chosen); and the third at position twelve or thirteen (randomly chosen); (c) the third and fourth training trials for each of the four sequences in the next sub-block, which filled up the remaining empty slots of the final fifteen. After the final fifteen trials were determined in this way, the remaining training trials from the current four sequences, as well as the first two training trials from the next four sequences, were randomly shuffled to form the first part of the sub-block.

The sixth sub-block within each block was determined similarly, with the exception that there were no new training trials from the next sub-block to be intermingled. For this block, the last nine trials were constructed first, with test trials for each of the four sequences appearing at serial positions one, five, eight, and nine. Positions six and seven had the last two training trials for the sequence tested at eight and nine; position two had the last training trial for the sequence tested at position five; and positions three and four had the second to last training trials for the series tested at eight and nine.

Procedure

Upon arrival each participant will be given an 'instruction' sheet detailing the task and their role during the experiment (see Appendix 1). Participants will sit opposite the eye tracker and computer screen. The experimenter (confederate) will sit behind the participant facing a separate computer monitor. The layout of the room is designed so as to ensure that neither the participant nor the experimenter will be able to see the each other's monitor. The participant will play the role of Director and the experimenter will play the role of the Matcher.

In each trial the Director will be asked to verbally name the target object so that the Matcher can identify the item on their monitor and select it using a mouse. In order to discriminate the target object from the filler objects, the target will be highlighted within a green square in the Director's display (see Figure 1 and Figure 2). As the arrangement of images within the Matcher's will differ in a unpredictable way from that of the Director, the speaker will have to describe the features of the highlighted item, rather than use the target's grid location as a description.

Unlike Experiment 1, which had a preview of the target location before the full set of images appeared, the location of the target object will appear at the same time as the rest of the images within the grid. Audio recording of the Director's response begins simultaneously with the presentation of the main display. The trial will end when the Matcher selects the object designated by the Director. The Director cannot see the Matcher's screen or mouse pointer, and will receive no feedback regarding whether the trial was completed correctly. If the Director fails to provide sufficient information to identify the target, the Matcher will ask the director for clarification (e.g. "which one do you mean?"). Any such clarification exchanges will appear in the audio recording for the trial and will be noted during later transcription.

PREDICTIONS AND STATISTICAL ANALYSIS

Main measurements

Our analysis will focus on three categories of measurements: (1) speech content and performance; in particular, use of a descriptive modifier and speech fluency; (2) speech onset latency, defined as the time taken to produce the first content word as measured from the onset of the display; and (3) eye gaze behaviour.

Transcription and coding of the audio files

For each of the 48 sequences for each director, we will transcribe and code the audio recordings for two trials: (1) the last trial of the training sequence; and (2) the test trial. The last training trial is needed in order to provide baseline data for the speech onset latency in the test trial. Each trial will be transcribed and coded for fluency and adjective use. Fluency will be coded into one of five categories, as shown in the table below. We have included a new category of fluency (LE for lengthened speech) in addition to the four categories we used previously in Experiment 1.

Code	**Description**	**Example(s)**
FL	Fluent speech	"the family car", "the car", "car"
UP	Unfilled pause (occurring after speech onset)	"the... silver car"
FP	Filled pause (um/uh)	"um... the car"
RE	Repaired utterance	"car... yeah the family car", "car... uh…family car"
LE	Lengthened speech	"the s(*ssss*…)ilver car"

We will also code whether or not a descriptive modifier is used, defined by the following categories:

Code	Description	Example(s)
NO	No modifier	“car”, “the car”, “the silver car”
PR	Pre-nominal modifier	“family car” , “normal car”
PO	Post-nominal modifier	“car, the family car”, “car, family one”
DE	Deleted adjective	“fa—uh… just the car”
AS	Addition due to self-repair	“car... family car ”
AO	Addition due to other-repair	“car...” [Matcher: “Which one?”] “Oh, the family one”

Similarly to Experiment 1, onset times of utterances will be measured in milliseconds. The following criteria will be applied when identifying utterance onsets:

- Trials will be discarded if the speech is unidentifiable
- Any filled pauses or articles will be ignored (um, uh, the); speech onset is identified as the first content word (e.g., adjective or noun), even if the adjective referred to colour rather than size (e.g., for “uh, the silver car” onset would be taken to be at the onset of the word ‘blue’).
- If directors correct themselves after an error (e.g. 'white car...eh sorry silver car’) onset of the correction (i.e. silver) will be recorded; however, such repaired utterances will not be used in the analysis of speech onset.

Data preparation and statistical analysis

One concern is that some directors might opt for a strategy of “hyper-describing” target objects; providing long, rich descriptions that would differentiate targets from nearly any possible competitor objects; moreover, doing so even when there is no competitor in the display. The problem with this behaviour is that on test trials in the singleton-contrast condition, directors could simply continue using the modified description, which would then spuriously appear to be appropriately specified. We plan to identify these participants by coding whether or not in the final training trial for each sequence in the Singleton-to-Contrast condition, they inappropriately the modifier that would have differentiated the target from the (absent) competitor. We will delete all data from speakers who did this on more than half of these trials. For all remaining participants, we will also exclude on a trial-by-trial basis any test trials in the singleton-contrast condition where on the last training trial speakers used a modifier that would distinguish the target from the competitor. (In the contrast-singleton condition, this is less of an issue because speakers must use size modifiers during training or the addressee will be unable to resolve the reference; however if we find that a speaker repaired an utterance—e.g., “the car, uh the family car” —in the last training trial for this condition, we will discard the following test trial from the analysis.)

We will also check the quality of the materials to determine whether there were certain stimulus items that should be excluded. In particular, we will consider the last training trial of each series for each item in which the critical object was a foil, and remove from the analysis any target item for which more than 50% of speakers used a description that would have distinguished it from the corresponding competitor.

The statistical analysis for the production data (modifier use and speech onset) will be performed using linear mixed-effects models with directors (subjects) and sequence (item) as crossed random

factors (Baayen, Davidson, & Bates, 2008). All analyses will use the maximal random effects structure justified by the design (Barr, Levy, Scheepers, & Tily, 2013), which implies by-subject and by-item random intercepts and by-subject and by-item random slopes for both main effects (Training-Test Consistency and Shift Direction) and their interaction. We will derive *p*-values using the t-to-z heuristic (i.e., deriving p-values from the standard normal distribution for the t statistic), as this enables us to perform one-tailed tests (see below). Models will be estimated using the lme4 package in R (version 1.1-7 or higher). All independent variables will be deviation coded. The analysis of modifier use will assume a logit link and binomial variance function, whereas the analysis of onset times will use an identity link with a gaussian variance function. Should the maximal models fail to converge, we will simplify the model in the following sequence of steps. First, we will constrain the covariances from both the items and subjects matrices to zero (Barr et al., 2013). If the model still fails to converge, we will then seek to simplify by interatively dropping the component with the smallest estimated variance.

Our analyses of the eye-tracking data will use synchronized permutation tests (Pesarin, 2001) because of the unknown dependency structure at the level of individual frames of eye data.

Analyses and predictions

The basis for our estimate of a sample size of 36 participants (power = .85) was derived from a pilot study conducted prior to Experiment 1. Our main prediction is that speakers will misspecify referents at a higher rate in the Consistent Training-Test condition than in the Inconsistent Training-Test condition. To maximize power (especially important given that the DV for this analysis is binary), we have opted to test for the main effect of Training-Test Consistency using a one-tailed test. Please see the documentation for the first study for further information about the power calculation (https://osf.io/4akir/). Although the previous study was unsuccessful, we believe that the design was not ideal, because the re-use of letter stimuli as targets could have led to crosstalk in memory across sequences that masked any effects of retrieval fluency. With the numerous changes to the procedure to improve sensitivity, we see no reason to boost our sample size.

Our second main prediction concerns the speech onset latency for appropriately specified descriptions. Our prediction is that speakers will experience more difficulty shifting from the entrained description to a more contextual appropriate description in the Consistent Training-Test condition than in the Inconsistent Training-Test condition, due to a more fluent retrieval of the entrained response. This analysis will include *only* trials where the target was appropriately specified both at test as well as in the last training trial before test. The dependent variable is the speech latency for the test trial minus the speech latency for the final training trial for that sequence; in other words, the change in speech latency incurred by abandoning the entrained description. Our power analysis suggested .93 power for a two-tailed test with N = 36.

For the eyetracking data, we predict a lower proportion of gazes to non-target images in the grid prior to the onset of speech in the Consistent Training-Test condition than in the Inconsistent Training-Test condition; this would reflect less consideration of context due to a strong memory signal.

In sum, we have two key predictions:

(1) A main effect of Training-Test consistency on misspecification, with more frequent misspecification in the Consistent condition, alpha=.05, one tailed;

(2) For appropriately specified descriptions, a main effect of Training-Test consistency on speech onset latency (relative to the last training trial), with longer relative delays in the Consistent condition, alpha=.05, two-tailed;

We also make two additional (less critical) predictions:

(3) Greater rate of underspecification than overspecification (based on the previous experiment); in other words, a higher rate of misspecification in the Singleton-to-Contrast condition than in the Contrast-to-Singleton condition, alpha=.05, two-tailed;

(4) Fewer non-target fixations prior to speech onset in the Consistent than in the Inconsistent condition, alpha=.05, two-tailed.

APPENDIX 1: Participant Instructions

Social Description Task – Information Sheet

In this experiment you will play the role of the ***Director*** and the experimenter will play the role of the ***Matcher***. You will be seated at a computer monitor and presented with a series of 5x4 grids containing different objects. In each trial a single object will be highlighted by a green outline. Your task is to verbally name this item so that the experimenter is able to select it on a separate computer monitor (please see ***Fig. 1*** below).

Although, the experimenter's monitor will contain the same objects as those that appear on your screen, they will be arranged in a completely random order. Therefore, it is unlikely that the objects will appear in the same locations as those shown on your screen. In order to provide an accurate instruction to the experimenter, you must avoid using the *spatial location* of the target item in your description. You may, however, describe the item in any other way that you think may help the experimenter to locate the target object.

Throughout the experiment your responses will be recorded and your eye movements will be tracked. There will be an opportunity to take a break during the experiment.

Please ask the experimenter *now* if you have any questions about your role in the study. There will be a full debrief after the experiment is finished.

An example of the type of grid that you will see →

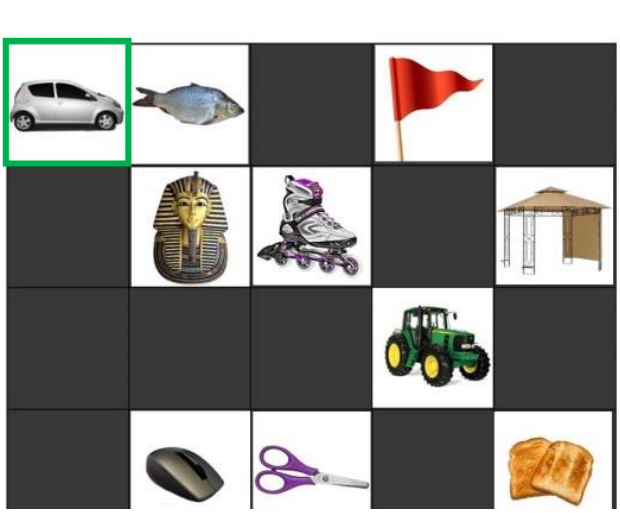

Figure 1: Example of the display on the *Participant's screen*. The Participant will identify the highlighted target object ("car" in this example) to the Experimenter. Once the Experimenter has selected the letter, a new trial will begin.

APPENDIX 2: List of all Target and Critical objects used

Target	**Competitor**	**Foil**
Egg in shell	Egg yolk	White flower petal
Family car	Sports car	Grey computer mouse
Wall clock	Digital clock	'Dr. Beats' speakers
Office phone	Mobile phone	Remote control
Reading glasses	Drinking glasses	Test beakers
Kitchen knife	Swiss army knife	USB stick
Mountain bike	Motor bike	'Go' Kart
Leather glove	Boxing glove	Bean bag
Gold key	Car key	Ping pong bay
Riding saddle	Bicycle saddle	Putter
Camcorder	CCTV camera	Hairdryer
Computer mouse	Mouse	Squirrel
Orange	Orange slice	Sunset picture
Sun hat	Cowboy hat	Wooden bowl
Gun	Toy gun	Hook
AA battery	Car battery	Box
Bedroom lamp	Lava lamp	Rocket
Money (notes)	Money (coins)	Bolts and screws
Boot	Car boot	Breadbin
Red apple	Green apple	Pear
Bicycle helmet	Builders helmet	Mellon

Acoustic guitar	Electric guitar	Frying pan
Garden spade	Beach spade	Spatula
Horse	Rocking horse	Cradle
Mirror	Hand mirror	Wreath
Bumblebee	B letter	D letter
Smoking pipe	Kitchen pipe	Flute
Chair	Baby highchair	Ironing board
Candle	Melted candle	Vase
Teapot	Teapot with cosy	Woolly hat
Fan	Electric fan	Drain cover
Yellow t-shirt men's	Yellow t-shirt women's	Yellow tea towel
Padlock unlocked	Padlock locked	Handbag
Cheese	Blue cheese	Sponge
Wine glass	Glass of red wine	Decanter
Coffee cup	Coffee cup and saucer	Plant pot
Saw	Electric saw	Blender
Bat	Baseball bat	Chopsticks
Human eye	I letter	L letter
Headphones	Headphones(ear buds)	Ear plugs
Ballpoint pen	Pen without lid	Pencil
Spoon	Wooden spoon	Wooden spatula
Bin	Pedal bin	Black jug
School bell	Bicycle bell	Bauble
Open umbrella	Closed umbrella	Nail file
Potatoes	Peeled potatoes	Lemons
Lighter with flame	Lighter	Flask

Door long handle	Door knob	globe

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390-412. doi:10.1016/j.jml.2007.12.005.

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

Pesarin, F. (2001). *Multivariate permutation tests: with applications in biostatistics*. Chichester: Wiley.

Tulving, E., & Thomson, D.M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*, 352-373.