WORKING MEMORY AND REFERENTIAL COMMUNICATION

WORKING MEMORY AND REFERENTIAL COMMUNICATION: AN INVESTIGATION OF THE COGNITIVE FACTORS AFFECTING THE PRODUCTION OF OVERSPECIFIED REFERRING EXPRESSIONS

BY JULIE BANNON, B. Sc. (Hons)

A Thesis Submitted to the School of Graduate Studies for Partial Fulfillment of the Requirements for the Degree Master of Science

McMaster University

McMaster University MASTER OF SCIENCE (2018) Hamilton, Ontario (Psychology, Neuroscience, and Behaviour)

TITLE: Working memory and referential communication: An investigation of the cognitive factors affecting the production of overspecified referring expressions

AUTHOR: Julie Bannon, B.Sc. (Hons) (University of Toronto)

SUPERVISOR: Dr. Karin Humphreys

NUMBER OF PAGES: x, 59

Lay Abstract

Interactive communication often involves speakers relaying information to a conversational partner about objects in the environment, a phenomenon typically referred to as referential communication. A significant focus of previous research in this area is on how speakers chose to identify objects for a conversational partner. The focus of the current research is to examine the underlying cognitive mechanisms that support this behaviour. Participants were asked to communicate with a partner about objects on a computer screen while completing a secondary memory task. The findings of this research suggest that individual working memory capacity influences the amount of information speakers choose to include in their referential descriptions. Further, we show that including unnecessary information in object descriptions (i.e., referring to object attributes in the absence of contrastive objects) leads to a reduced speech rate, likely because of increased demands on speech planning.

Abstract

Language production often requires speakers to convey information to a conversational partner about objects in their environment. According to Grice's Maxim of Quantity (1975), speakers should provide only the precise amount of information needed to identify an object. However, it is frequently observed that speakers will include redundant adjectives in their referring expressions, rendering their descriptions overspecified. The majority of the research investigating overspecification has focused on how scene characteristics influence the likelihood of this behaviour. To date, less is known about the internal characteristics of the speaker that may play a role in the production of overspecified descriptions, and in referential communication more generally. The current experiment investigates the role of working memory in the generation of referential descriptions and examines how this interacts with manipulations of scene characteristics and cognitive load. Participants were asked to provide instructions to a confederate about which object to select from an array of either three or six unrelated objects while they simultaneously remembered a series of either zero, three, or five numbers. Participants also completed an operation span task to measure their individual working memory capacity (WMC). Results showed a main effect of array size for speech onset times, confirming that speakers are faster to initiate their speech when there are fewer objects in the display. Further, there was a significant three-way interaction between array size, cognitive load, and operation span scores, indicating that speakers with lower WMC are more likely to use redundant adjectives for three object arrays under low levels of load. Finally, there was a significant, negative correlation between speech rate and adjective use, indicating that speakers adjust their rate of speech depending on their choice of referring expression. The results of this research suggest a potential role for individual WMC in the production of overspecified descriptions.

Key words: language production, referential communication, overspecification, working memory, cognitive load, speech planning

Acknowledgments

I would like to thank my supervisor, Karin Humphreys, for her continued support and enthusiasm towards the research. I would also like to thank the members of committee, Scott Watter and Lili Service for their valuable input into this thesis. Finally, thank you to my wonderful friends and family, and especially to my parents, for all their support and encouragement.

Table of Contents

List of Tables and Figures

Declaration of Academic Achievement

This thesis presents original research investigating the role of working memory in the production of referential descriptions. I am the sole author of this thesis. The research project outlined in this document was developed by myself in collaboration with my supervisor, Dr. Karin Humphreys, with input from my committee members Dr. Scott Watter and Dr. Elisabet Service. For this reason, I have chosen to use the collective pronoun "we" throughout this thesis. I programmed the experiment for this project, and data was collected by myself with the assistance of undergraduate research practicum students. I conducted all statistical analysis for this research based on the guidance of my supervisor Dr. Karin Humphreys.

Introduction

Effective communication requires speakers to construct messages that are contextually appropriate and easily understood by the listener. This is especially evident in the context of referential communication, which requires speakers to produce accurate descriptions of objects for a conversational partner to identify. One important factor for speakers to consider is how much information is needed to convey a particular message. According to Grice's (1975) Maxim of Quantity, speakers should include only the precise amount of information needed, and nothing more or less. However, it has been demonstrated that speakers often include reference to object features that are not necessary for identification, resulting in overspecified descriptions (e.g., Koolen, Gatt, Goudbeek, & Krahmer, 2011; Rubio-Fernandez, 2016; Tarenskeen, Broersma, & Geurts, 2015). Early research on referential overspecification done by Pechmann (1989) suggests that overspecification is a result of the incrementality of speech production. Pechmann argued that speakers initiate their speech before they have fully scanned the visual display, and therefore refer to object features before determining whether they are distinguishing. Since Pechmann's (1989) early work on the relationship between speech planning and referential overspecification, research has largely focused on how the visual context affects the likelihood of overspecification, and has often ignored the possible role of individual differences in executive function. The purpose of the current research is to investigate the impact of individual working memory capacity (WMC) – a system frequently implicated in language processing and production – on the formulation of referring expressions, and particularly on the production of overspecified descriptions.

Further, we seek to better understand the relationship between overspecification and speech planning. This thesis will begin with a review of the literature on referential communication, and the relationship between working memory language production. It will then present data from a new experiment that investigates how internal characteristics of the speaker impact the production of referential descriptions.

Referential Communication

Spoken language generally involves two or more individuals exchanging information. In referential communication, speakers are frequently required to communicate with a partner about an object present in the here and now. This means that speakers must formulate referring expressions that enable the listener to make an inference about which object is being referenced. This process has often been examined in situations where conversational partners are presented with hard-to-describe objects such as tangrams. In such studies it is often observed that speakers will shorten their descriptions over time as the two partners form conceptual pacts for how to refer to a particular object (Brennan & Clark, 1996). This means that speakers must work together to formulate accessible labels for objects. However, in everyday communication, it is much more likely that speakers will produce one-shot references for typical objects in their environment. Although speakers are generally able to do this with ease, research has shown that the way in which people construct referring expressions is not always efficient. This is evident in simple referential communication tasks that require a speaker to identify everyday objects such as clothing or furniture for a partner. In such a situation, the simplicity of the task often demands only that speakers provide a label for the object as no additional information is necessary for identification. However, as noted above, it is frequently observed that speakers will include superfluous reference to features of an object, rendering their referring expressions overspecified (e.g., referring to a shirt as *'the yellow shirt'* when there is only one shirt present). The use of these overspecified referring expressions has been shown to be influenced by the context of the visual display. For example, people are more likely to overspecify when presented with polychrome versus monochrome displays (Koolen, Goudbeek & Krahmer, 2013; Rubio-Fernandez, 2016), when an object is low in colour diagnosticity (Sedivy, 2003; Westerbeek, Koolen & Maes, 2015), or when an object is presented in an atypical colour (e.g., a purple banana; Westerbeek, et al., 2015). The influence of context and object features in the production of overspecified references invites the question of whether such behaviour serves a communicative purpose, or whether it is due to other internal characteristics influencing the speaker.

Rubio-Fernandez (2016) suggests that including reference to colour during object identification may be a communicative tool that helps draw the listener's attention towards the objects (see also Arts, Maes, Noordman, & Jansen, 2011). When participants were informed that listeners in a pilot study had difficulty identifying objects described by the participants, they were more likely to include redundant colour adjectives in their descriptions. This suggests that speakers may intentionally overspecify their referring expressions to avoid ambiguity and make it easier for a listener to quickly locate a target object. However, there is conflicting evidence as to whether overspecification is beneficial for the listener. Some studies of language comprehension have found that overspecified references can facilitate faster object identification when descriptions include reference to location in addition to object properties (Arts et al., 2011), or when searching for objects in a virtual environment (Paraboni & van Deemter, 2014). When real-time measures of comprehension are used (e.g., eye-tracking, ERPs), the evidence suggests that unnecessary prenominal adjectives may cause initial confusion for the listener (Engelhardt, Baily, & Ferreira, 2006; Engelhardt, Demiral, & Ferreira, 2011; but see Tourtouri, Delogu, & Crocker, 2017). Engelhardt and colleagues (2006) found that in a rating task, participants did not judge overspecified descriptions to be infelicitous, but when the same descriptions were heard in a comprehension task, eye-movements indicated initial confusion when an unnecessary prenominal adjective was included. Further, Engelhardt et al. (2011) found that response times were slower when the target object was identified by a label and an additional modifier, and ERPs revealed evidence of an N400 shortly after the onset of unnecessary adjectives in object descriptions. Although it is possible that overspecification is a communicative strategy, this evidence from online language comprehension suggests that overspecification may not be an effective tool to make object identification easier for the listener. It is therefore important to consider the internal influences on communication that are present during referential tasks that may impact speakers' lexical choices when producing object descriptions.

Koolen, Gatt, Krahmer, van Gompel, and van Deemter (2017) found that when speakers were shown a visual display for a short, limited amount of time, they were more likely to include redundant adjectives than when they were given unlimited time to view the display. Based on this finding, they suggest that overspecification may result from a reliance on heuristics during speech production: when there is insufficient time to view the display, speakers may include prenominal adjectives that increase the chance that their descriptions will differentiate the object from the rest of the scene. In essence, this means that speakers will include more redundant adjectives when they do not thoroughly inspect the visual display. This explanation is consistent with Pechmann's (1989) theory that overspecification results from the incrementality of speech production, namely that redundant adjectives are included in descriptions before the speaker has determined whether they are distinguishing. The results of Koolen et al. (2017) further suggest that the inclusion of redundant adjectives may be influenced by the visual search that is required to identify both the target object and the object properties that are required to differentiate the target from distractors in the scene. Indeed, it has been shown that certain characteristics of the visual scene can affect colour overspecification. For example, speakers are more likely to include colour adjectives in their descriptions when there is more clutter in the visual scene (Koolen, Krahmer, & Swerts, 2016). Such results are in line with the idea that when there are more objects present in a display, speakers will adopt a communicative strategy that allows them to choose a description for an object without comparing the target against each individual item. Beyond simple colour adjectives, the complexity of the visual scene has also been shown to affect the content of referential descriptions. Elsner, Clarke, and Rohde (2018) found that increasing grid size and object heterogeneity increases the use of coordinates in descriptions, whereas object descriptions that include reference to colour, shape, or size tend to decrease in homogeneous displays. Similarly, Clarke, Elsner, and Rohde (2013) demonstrate that the length of referential expressions increases as visual scenes become more complex. It would seem that the lexical choices that speakers make when formulating descriptions vary depending on the content and complexity of the visual scene, with the use colour adjectives being more likely when the target's colour is salient and scenes are less complex, and alternative descriptions (e.g., reference to coordinates or landmarks) becoming more likely as scene complexity increases.

In addition to influencing the content of referential expressions, variation in the visual scene has also been shown to impact speech planning. Increasing the number of distractor images present in a visual display can affect the time it takes for speakers to plan their referential descriptions, as evidenced by increased onset times for targets presented among a larger number of distractors. This increase in speech onset appears to function linearly as the number of objects in the scene increases (Gatt, Krahmer, van Deemter, & van Gompel, 2017), and is present both in highly complex scenes containing grids with multiple geometric shapes (Elsner, et al., 2018) and in relatively simple scenes containing multiple objects of the same type (Gatt, et al., 2017). However, it is important to note that Gatt and colleagues (2017) found that this effect was dependent on the type of adjectives used to distinguish the object. In fact, the effect was not present when only colour was needed to distinguish the target object from competitors. Gatt et al. (2017) argue that this is because the colour of an objects "pops-out", making it easy to quickly encode, whereas the use of size adjectives requires one to compare the target object against each competitor. This suggests that the inclusion of redundant adjectives in object descriptions may place different demands on speech planning depending on the type of modifier use in the description. Specifically, it is possible that the encoding of colour adjectives is more automatic in comparison to the use of more complex modifying terms such as coordinates or landmarks.

The current body of research leaves several questions unanswered. First, it is apparent that the number of objects in the visual scene affects both object descriptions and initiation times, but it is not clear how this precisely affects speech planning and overspecification. Specifically, while it has been shown that speakers take longer to initiate their speech when using certain combinations of adjectives when there are more objects present (Elsner et al., 2018; Gatt et al., 2017), there is little evidence regarding whether including redundant adjectives in the absence of a competitor objects affects speech planning as measured by delayed onset times. Second, much of the literature on referential communication, and specifically overspecification, focuses on the effects of the visual scene on object descriptions. However, there is little research that examines how internal characteristics of the speaker affect referential communication. It is important to consider the speaker in theories of overspecification because producing referring expressions can be a cognitively demanding task. For example, the speaker must consider the knowledge state of the addressee, the amount of information required to identify the object, and the planning of the final utterance. Although there is growing evidence that certain executive functions such as individual WMC play a role in language production, it is not well understood how WMC affects speech planning times or lexical decisions during referential communication.

Working memory and language production

Working memory is the cognitive system used for the temporary storage and maintenance of information relevant to the execution of complex tasks (Baddeley, 1992, 2003). This is important for theories of language production because speaking requires that a person is able to simultaneously plan and execute speech that is grounded in both discourse and visual contexts. Individual WMC has been shown to play a role in both language comprehension (e.g., Daneman & Merikle, 1996; Huettig & Janse, 2016; Otten & van Berkum, 2009) and production (e.g., Daneman, 1991; Fehringer & Fry, 2007; Hartsuiker & Barkyuysen, 2006; Slevc, 2011), although the relationship between WMC and language production is much less studied.

One common method of examining the role of WM in spoken language is to tax the system by requiring speakers to complete a secondary task, thereby increasing the amount of information that must be held in WM. To date, the majority of research investigating the relationship between WM and speech production has focused on the role of WM in sentence planning. For example, Power (1985) found that participants had shorter speech onset latencies when they were asked to remember either 3 or 6 digits while producing a sentence. Other studies using verbal working memory tasks have found that speakers are less likely to mention accessible items early in a sentence when verbal WM demands are increased (Slevc, 2011). However, WM load does not appear to affect the phrasal scope of planning. In a study by Martin, Yan, and Schnur (2014), participants were slower to initiate sentences that began with complex rather than simple noun phrases, but this effect was not exaggerated when participants were asked to remember either two words or a dot pattern while producing sentences. In addition to speech initiation times, WM load has been shown to affect other speech characteristics such as speech rate. For example, Carro, Goudbeek, and Krahmer (2014) found that individuals articulate words faster when performing a cognitively demanding task. Cognitive load has also been shown to cause older adults to reduce their rate of speech (Kemper, Herman, & Lian, 2003), suggesting at least some influence of cognitive capacity on an individual's ability to manage additional load. Finally, individual WMC has also been implicated in syntactic planning. In an experiment by Hartsuiker and Barkhuysen (2006), it was shown that speakers with lower reading spans were more likely to make subject-verb agreement errors when formulating sentences under load, whereas speakers with high reading spans were not affected by the secondary task.

This body of evidence suggests WM plays a role in language production, but there is currently little research investigating the role of WM in referential communication. The few studies that do examine WM load and reference production have found that speakers adjust their speaking strategies while referring under load. For example, it has been shown that speakers are less likely to align their descriptions with a previously primed reference structure when referring under load (Goudbeek & Krahmer, 2011). This could be a result of speakers' decreased ability to pay attention to information previously presented to them when their working memory system is overloaded. Speakers also demonstrate increased pronoun use when referring to discourse-salient characters in conversation (Vogels, Krahmer, & Maes, 2015), suggesting that speakers do adopt different referential strategies when they are forced to manage competing task demands. It has been argued that individual variation in the use of pronouns in reference can be explained by individual differences in WMC. Hendriks' (2016) ACT-R model of reference argues that individuals with lower WMC are more likely to use pronouns in cases where a full noun would be preferred, and less likely to use pronouns for discourse-salient targets. It should be noted that the predictions of this model stand in contrast to the results of Vogels and colleagues (2014), making it difficult to determine the precise role of WM in lexical choices during reference production. Further, the ACT-R model focuses on attempting to explain individual differences both over- and underspecification in relation to pronoun use during discourse, and does not consider cases of overspecification that involve reference to object properties.

One possibility is that individuals with lower WMC may overspecify less because overspecification requires encoding additional information beyond the object name, and therefore holding more information in memory during language production. Alternatively, speakers with lower WMC may overspecify more because selection of the optimal description for a target requires consideration of entire visual display in order to formulate a referring expression that can be most easily understood by the listener. The latter view suggests that the inclusion of redundant adjectives may be a simpler strategy because it increases the chance that a referring expression will sufficiently identify the target object without considering the entire display or the needs of the listener. Further, there is currently little evidence to suggest that overspecification causes difficulties in speech planning, and there is reason to believe that overspecification may be a useful tool for those with lower WMC. According to both Pechmann (1989) and Koolen et al (2017), overspecification results from speakers choosing a prenomial adjective to describe a target before deciding on the optimal description to identify the target. This suggests that speakers do not scan the full display before beginning to produce overspecified descriptions. There is further evidence demonstrating differences in visual scanning between high and low WMC individuals. In a study by Swets, Jacovina, and Gerrig (2014), speakers provided a sentence to a partner about how to move objects on a screen. On critical trials, the target was either unique in the display, or needed to be differentiated from a competitor object of the same type. Analysis of scanning behaviour revealed that participants who scored low on a reading span task spent less time gazing at the competitor object prior to sentence onset. These results suggest a relationship between scanning behaviour and WMC such that those with lower reading span scores are less likely to fixate on the entire display, and therefore reduce their scope of advance planning. If, as Pechmann (1989) argues, overspecification results from incremental planning such that speakers produce redundant adjectives before scanning the entire visual scene, then it is likely that individuals with lower WMC will rely more heavily on the use of prenomial adjectives in referential communication because of a reduced tendency fixate on more objects in the display prior to speech onset. In the present study, we manipulate characteristics of the visual display, as well as the amount of additional cognitive load, in order to determine the role of scene characteristics and WMC in the production of overspecified descriptions.

The current study

The purpose of the current study is to investigate the effects of cognitive load on referential overspecification, and to determine how this speaking strategy is mediated by individual WMC. We also seek to better understand how WMC influences individual speech characteristics such as planning time and rate of speech. Participants in this experiment were asked to take part in a referential communication task in which they had to verbally identify objects for a partner while keeping a series of numbers in memory. The level of cognitive load (i.e., the number of items participants were asked to remember) and number of images on the screen were manipulated.

Based on the previous literature, several predictions can be made about the results of this study. First, based on the finding that speakers are more likely to overspecify when there is more clutter in the visual scene, we expect that rates of overspecification will be higher for six rather than three object arrays. Second, we expect that rates of overspecification will increase as load increases. This prediction is based on theories of referential communication which suggest that overspecification is used by speakers to alleviate search effort by choosing object descriptions before scanning the entire scene. If this is what speakers are doing, then this strategy may be particularly useful under increased load because it would alleviate the need to focus on anything other than the target object. Although it is possible that overspecification will decrease under load because of the additional conceptual planning involved in producing colour adjectives, there is not currently evidence to suggest that the production of overspecified descriptions is taxing on the speaker. We also expect that speakers with lower WMC will produce more overspecified descriptions. This third prediction is based on evidence that speakers with low WMC are less efficient in examining the visual scene, and are more likely to rely on simplistic strategies when completing a cognitively demanding task.

Finally, it is expected that this experiment will extend previous findings that speakers are faster to initiate their speech when there are fewer objects present in a visual display. Gatt and colleagues (2017) previously demonstrated that onset times increase when more objects are included in the display (see also Elsner et al., 2018), but only when speakers needed to include reference to size or size and colour, and not for colour alone. However, they do not present data indicating that this effect is present when speakers are never required to use adjectives to describe objects in the visual display. In this experiment, we test whether this effect is present when speakers are not required to use adjectives to describe an object, and so any use of adjectives will result in overspecified descriptions. We expect that the number of objects in the display will affect speech onset times, with speakers taking less time to initiate their speech when there are fewer objects present.

In addition to measures of speech onset and overspecification, we also measure rate of speech in relation to the independent variables. The evidence concerning speech rate and cognitive load is somewhat mixed, with Power (1985) finding no differences in speech rate as load increased, while other work has demonstrated decreased speech rate under higher amounts of load, although this was found in older adults (Kemper, et al, 2003). However, Fehringer and Fry (2007) found that individuals with lower WMC had greater pause durations in their L2, suggesting that some aspects of speech rate may be related to WM. Although it is not clear how the manipulations of this particular study will affect individual rate of speech, it is included as a measure in order to more accurately determine the relationship between WMC and speech rate. Further, measuring speech rate allows us to examine whether the inclusion of redundant adjectives causes speakers to reduce their speech rate as it may require them to make lexical decisions as they simultaneously produce speech, therefore taxing their planning system.

Method

Participants

Participants were 36 (*mean age* = 18.75, *range* = 18-22) students from the McMaster University undergraduate community who received partial course credit for their participation. All participants reported native or native-like fluency in English and were not colour blind.

Materials and Design

The experiment employed a 2 (Array size: three or six images) X 3 (Load level: zero (no load), three (low load), or five (high load) numbers) within-subjects design. Critical trials consisted of images that were clipart depictions of everyday objects. Images were taken from normative databases (Moreno-Martinez & Montoro, 2012; Saryazdi, Bannon, Rodrigues, Klammer, & Chambers, 2018) or from online image repositories. They were chosen to be similar in style and brightness of colour. Filler targets were either line drawings or essentially colourless objects (e.g., a whisk), and were chosen because it would be difficult to describe them based on colour attributes. The purpose of the fillers objects was to prevent participants from forming a habit over the course of the experiment such that they describe all objects using colour adjectives.

There were 36 experimental trials, half of which presented the target along with two other images (three object arrays), and half contained five other images (six object arrays). All images in the array were differently coloured and semantically unrelated to the target object so any reference to object properties in this task would result in overspecified descriptions. Filler trials consisted of either two or four objects. The experimental trials were broken down into sets of five, each of which contained three critical trials and two filler trials. The experiment contained a total of twelve sets of five trials, resulting in a total of 60 trials overall (36 critical and 24 fillers). At the beginning of each set, a series of randomly generated numbers were displayed to the participant. Four sets required participants to remember three numbers (low load condition), four required them to remember five numbers (high load condition), and four did not display any numbers (no load condition). This resulted in a total of six different lists, such that each object appeared equally in both three and six image arrays and at each level of load. The experiment was programmed and administered using PsychoPy version 1.85.4 (Pierce, 2007), and voice recordings were captured using version 2.2.2 of Audacity© recording software.

Verbal working memory capacity was measured using the Automated Operation Span (AOPSAN) assessment and was administered online via Millisecond Test Library (https://www.millisecond.com/download/library/ospan/). In the AOSPAN task, participants are presented with a simple math equation and must subsequently verify the answer. After confirming the answer to the equation, they are shown a screen with a letter on it and asked to it, after which they are presented with a new equation. The presentation of equations and letters alternates until the end of the set, after which the participant must report back the letters in the correct order in which they were presented. This task was chosen because it has been shown to be a good measure of individual WMC (Unsworth, Heitz, Schrock, & Engle, 2005), and because evidence suggests that the ability to successfully manage dual task demands such as in this experiment (i.e., remembering a string of numbers while speaking) can be predicted by complex span tasks (Redick et al, 2016).

Procedure

The testing room contained two desks placed next two each other, each with an individual monitor. The participant's computer tower was positioned on the desk so that it blocked the view of a confederate's screen. When participants were brought into the room, a confederate was already seated at one of the desks. The confederate was an undergraduate research assistant who was posing as another student who had signed up to participate in the experiment for credit. Participants were told that, because they arrived second, they would be the speaker in the experiment. They were informed that they would be completing a task in which they would have to use a full sentence to give the other participant an instruction about which object to click on from an array presented on their screen while completing a secondary memory task. Participants were told that the confederate would be presented with the same images, but that they would not necessarily be in the same positions on the screen. The confederate and participant computers were not connected, and so the participant was instructed to wait for verbal confirmation to move onto the next trial, which they could proceed to by pressing the space key. After the participant gave their instruction, the confederate responded with "okay" and never asked for clarification about which object was to be selected.

The experiment began with one practice set in which the participant was shown only two numbers to remember. After the practice set, they were given the opportunity to ask any clarification questions before proceeding to the experimental task. This also provided the experimenter the opportunity to correct them if they were completing the task incorrectly, such as not using complete sentences to give their instruction. At the beginning of each experimental set, a series of numbers appeared on the screen for 1500ms and then disappeared. Each trial began with a fixation cross in the middle of the screen for 500ms, followed by presentation of the images accompanied by a tone. One of the images on their screen was cued by a frame that appeared 1500ms after the images. This continued for five trials, and then a screen appeared indicating to the participant that the set had ended and asked them to report back the numbers in the correct order. The experiment consisted of twelve sets of five and took approximately fifteen minutes to complete. After the task was finished, the participant was informed about the identity of the confederate and was asked whether they believed the confederate was another naïve participant.

Following the experimental task, participants completed the AOPSAN. They were told that the AOSPAN required them to remember a series of letters while they verified the correctness of simple math equations. Detailed instructions and practice trials were provided to the participants in the AOSPAN program. Upon completion of the AOSPAN, participants were thanked for their participation and dismissed.

Data Preparation

A total of 1296 instructions were produced during the experiment. Of these, 38 (3% of the data) were removed because the participant was unable to correctly identify the object or because the participant did not use a complete sentence to give their instruction, resulting in a final data set of 1258 instructions. Instructions for each critical trial were transcribed and coded by research assistants who were blind to the conditions of the experiment. All of the speech timing measures (i.e., speech rate and speech onset time) were extracted using Praat version 6.0.16 (Boersma & Weenink, 2016).

Measures

Speech Rate Speech rate was calculated as syllables per second and refers to the rate of speech for the entire instruction up until the offset the target name. In cases in which the speaker included a postnominal object description or offered a second name for the target (e.g. *"Click on the bucket…or the pail"*), the length of the utterances was measured from the onset until the end of the first head noun and divided by the number of syllables uttered up to that point to obtain an accurate measure of speech rate for initial instructions. Filled pauses were not included in the syllable count.

Speech Onset Speech onset times give a measure of the time from the initial presentation of images to the time at which the participant initiated their instruction. If a speaker included a filled pause at the beginning of their instruction (e.g., *"uhh Click on the…"*), onset time was considered to be the beginning of the word *'click'*.

Noun Phrase (NP) Onset Because this task allowed speakers to use the same structure of utterance for each instruction (i.e., *"Click on the..."*), speech initiation times may be a less reliable indicator of planning because the same initial structure is likely easier to retrieve. For this reason, we include an additional measure that examines the amount of time taken to produce the NP. NP onset times give a measure of the time from the initial presentation of the images to the onset of the NP. The onset of the NP is considered to be the time at which the speaker begins uttering either the head noun or a prenomial adjective followed by the head noun. Determiners such as *'the'* were not included in the calculation of NP onset. This was because participants were told to provide instructions in the form of *"Click on the...X"*, and so it is likely that determiners may be planned as part of the initial structure of the utterance and not as part of the NP.

Overspecification Instructions were considered overspecified if they contained a redundant prenominal colour adjective. Only the speaker's first attempt at identifying the target was considered for analysis, and instances in which speaker included adjectives postnominally (e.g., *"Click on the bucket…the pink one"*) were not included in this measure.

The decision to focus solely on prenominal colour adjectives in this analysis was made because we can be more certain that the use of redundant colour adjectives likely arises from a similar communicative strategy. Colour is also the most common property to be used in overspecified references (Tarenskeen et al., 2015), and represented 97% of all overspecified descriptions produced in this experiment.

Analysis

Analysis of the results was conducted based on two subsets of the data. The first subset, which will be referred to as the full data set, consists of all utterances produced in the experiment, excluding ones that were removed on the basis of inability to name an object or failure to use a complete sentence. The second subset, which we refer to as 'correct trials', is a subset of the data which includes only the trials from sets in which participants were able to accurately recall all the numbers in the correct order, as well as all trials from the no load condition. The decision to subset the data this way for analysis was motivated by the fact that the nature of the secondary memory task makes it possible for participants to trade off which task they are giving priority to. It is possible that an incorrect report of the numbers reflects a greater focus on the primary speaking task, and therefore less load. However, in the subset based on correct trials, we can be sure that participants were also focusing on remembering the numbers and were experiencing the desired amount of load.

Statistical analyses were conducted using R open-source software, Version 1.1.453 (R Core Team, 2017). Adjective use, speech onset, NP onset, and speech rate were analysed using mixed effects models from the lme4 package, Version 1.1-17 (Bates, Maechler, Bolker, & Walker, 2015) and lmerTest package, Version 3.0-1 (Kuznetsova, Brockhoff, & Christensen, 2017) in R. Summary statistics can be found in Table 1. Continuous measures were modelled using linear mixed effects models (Baayen, Davidson, & Bates, 2008), and dichotomous measures were modelled using generalized logit mixed effects models (Jaeger, 2008). Load condition (effect coded with the no load condition as the reference group) and array size were modelled as within subjects fixed factors, and operation span scores were centered and included as a continuous predictor. Participant and item were modelled as random factors, with random intercept terms included for both participant and item. Load condition and array size were included as by-participant slopes, and operation span scores, load condition, and array size were included as by-item slopes. See Table 2 for a summary of the analysis based on the full data set, and Table 3 for a summary based on the analysis of the correct trials. Correlations were conducted to measure the relationship between adjective use and the speech timing measures (i.e., onset time and speech rate). Summary of the correlations can be found in Table 4.

It should be noted that operation span scores were based on the absolute score for each participant. Although it has been suggested that partial credit unit scoring is preferred (Conway et al., 2015), absolute scores were chosen in this case because they reflect the same scoring procedure as was used to determine whether participants correctly recalled the numbers in the experimental task (i.e., all numbers correctly recalled in the correct order).

Results

Task Performance

In order to ensure that the memory task was sufficiently difficult, analysis was conducted to determine the effect of array size, load condition, and operation span scores on task performance (i.e., whether participants were able to accurately recall the numbers). This analysis was based on a subset of the data which excluded trials in the no load condition. Results revealed significant main effects of operation span scores (β = 1.08, SE = 0.28, z = 3.83, $p < 0.001$) and load condition (β = -0.60, SE = 0.11, z = -5.23, $p < 0.001$), and a significant interaction between operation span and load condition (β = 0.56, SE = 0.13, z = 4.48, $p < 0.001$). Follow up analysis revealed that the effect of operation span scores was significant in both the low load ($\beta = 0.45$, $SE = 0.24$, $z = 1.96$, $p = 0.049$) and high load ($\beta = 1.68$, $SE = 0.39$, $z = 4.34$, $p < 0.001$) conditions, although the difference in the low load condition is much closer to the threshold for significance. These results confirm that individuals with poorer WMC have more difficulty with the memory task, and that the high load condition was more difficult than the load low condition.

Speech Rate

The results did not reveal any significant effects of array size, load condition, or operation span score on speech rate. This was true for analysis based on the full data set, as well as the subset of correct trials. Results for speech rate by array size and load condition can be found in Figure 1.

Speech Onset

Results based on the full data set revealed a significant main effect of array size on speech initiation times ($\beta = -0.03$, $SE = 0.01$, $t(23.76) = -4.06$, $p < 0.001$), indicating that participants were faster to begin talking for three object arrays than six object arrays. This effect was also significant for the analysis based on the correct trials ($\beta = -0.04$, $SE = 0.01$, $t(26.79) = -4.39$, $p < 0.001$). No other effects reached significance for this measure. Results for speech onset time by array size and load condition can be found in Figure 2.

Noun Phrase (NP) Onset

Analysis of the onset of the NP revealed similar results to analysis of overall initiation times. Results based on the full data set revealed a significant difference in NP onset for three and six object arrays ($\beta = -0.04$, $SE = 0.01$, $t(26) = -3.17$, $p = 0.004$), with speakers initiating the NP earlier for three object arrays. Analysis of the subset of correct trials revealed the same effect of array size on NP onset times ($\beta = -0.05$, $SE = 0.01$, $t(41.94) = -3.38$, $p = 0.002$), and a marginal interaction between load condition and array size ($\beta = 0.03$, $SE = 0.02$, $t(541.7) = 1.73$, $p = 0.085$). A follow up analysis revealed that speakers were faster to produce the NP for three object arrays in the no load and high load condition, but the difference was not significant in the low load condition (although the average NP onset for three object arrays was still numerically lower than average onset for six object arrays ($M = 2.20$ v. $M = 2.25$, respectively) in the low load condition). Results for NP onset time by array size and load condition can be found in Figure 3.

Overspecification

Analysis based on the full data set did not reveal any significant main effects of array size, load condition, or operation span on the use of redundant adjectives. However, there was a marginal interaction between load condition and array size (no load vs. high load: $\beta = -0.32$, $SE = 0.18$, $z = -1.80$, $p = 0.071$), and a marginal three-way interaction between array size, load condition, and operation span scores (no load vs. high load: $\beta = -0.32$, $SE = 0.18$, $z = 1.76$, $p = 0.079$). Analysis based on the subset of correct trials again did not reveal any significant main effects, however there was a significant three-way interaction between array size, load condition, and operation span scores (no load vs. low load: $\beta = -0.80$, $SE = 0.31$, $z = -2.60$, $p = 0.009$; no load vs. high load: $\beta = 1.27$, $SE = 0.48$, $z = 2.63$, $p = 0.009$). Further analysis revealed that the interaction between operation span and array size was significant in the low load ($\beta = -0.59$, $SE = 0.26$, $z = -2.25$, $p = 0.025$) and high load conditions ($\beta = 1.37$, $SE = 0.67$, $z = 2.04$, $p = 0.041$), but not in the no load condition (β = -0.31, *SE* = 0.26, *z* = -1.18, *p* = 0.239). Although the simple effects did not reach significance, there was a marginal effect of operation span scores on adjective use in the three object arrays in the low load (β = -0.54, *SE* = 0.29, *z* = -1.86, *p* = 0.063) and high load (β = 2.97, *SE* = 1.66, *z* = 1.79, *p* = 0.074) conditions, but this effect was not present for the six object arrays. The results indicate a tendency for individuals with lower WMC to produce more overspecified descriptions for three object arrays under low levels of load, whereas individuals with higher WMC produced more overspecified descriptions under high load for the three object arrays. These results suggest a complicated relationship between WMC and the production of redundant adjectives. However, the results should be interpreted with extreme caution because the simple main effects do not quite reach significance. Further, the overall percentage of overspecified descriptions produced under high cognitive load in the subset of correct trials was only 5%, which may comprise of too few data points to draw any firm conclusions about individual variation in this condition. Results of overspecification by array size and load condition can be found in Figure 4. Results of overspecification and operation span scores by load condition for three object arrays in the subset of correct trials can be found in Figure 5.

Relationship between dependent variables

The purpose of calculating correlations between the dependent variables is primarily to determine the relationship between overspecification and measures of speech timing (in this case, speech onset and speech rate). For this reason, correlational analysis was only conducted on the full data set. Because array size has been shown to affect speech onset times (Gatt et al., 2017), we first calculated the relationship between overspecification, speech onset, and speech rate separately for three and six object arrays collapsed across load conditions. Significant negative correlations were observed between adjective use and speech onset (*r*(628) = -0.13, *p* = 0.001) and adjective use and speech rate (*r*(628) = -0.12, *p* = 0.004) for three object arrays. For six object arrays there was a significant negative correlation between adjective use and speech rate (*r*(626) = -0.08, *p* = 0.034), and a marginal positive correlation between adjective use and NP onset (*r*(626) = 0.07, *p* = 0.075). These results indicate that speakers slow down their rate of speech when producing overspecified descriptions regardless of the number of objects in the display. However, speakers are faster to initiate their speech when producing overspecified descriptions for three object arrays, but slower to initiate the NP for overspecified descriptions for six object arrays. Boxplots for speech rate by overspecification can be found in Figure 6, and boxplots for speech onset by overspecification can be found in Figure 7. Finally, boxplots for NP onset by overspecification can be found in Figure 8.

Further analysis was conducted to determine how this relationship was affected by load conditions. Interestingly, for three object arrays, adjective use and speech onset were only significantly correlated in the no load condition (*r*(209) = -0.14, *p* = 0.037) and marginally correlated in the high load condition (*r*(209) = -0.13, *p* = 0.068), while adjective use and speech rate were only significantly correlated in the low load condition (*r*(206) = -0.14, *p* = 0.040). For six object arrays, the correlation between adjective use and speech rate was marginal for the no load (*r*(209) = -0.12, *p* = 0.088) and low load (*r*(208) = -0.13, *p* = 0.053) conditions, but none of the individual correlations between adjective use and other speech planning measures reached significance when broken down by load condition.

Discussion

The current experiment investigated whether individual differences in WMC affect speech characteristics and lexical choices in the production of referring expressions. Participants were asked to provide instructions to a confederate about which object to select from an array of three or six, while remembering a series of numbers. Speech rate, speech onset time, and the use of redundant adjectives were measured in relation to array size, level of cognitive load, and individual operation span scores. The results showed a main effect of array size on both speech onset and NP onset time for both the full data set and the reduced data set of trials in which participants correctly recalled the numbers. Further, a significant interaction between array size, load condition, and operation span score on the use of redundant adjectives was observed just in the subset of correct trials. No significant main effects were observed for the speech rate measure, but speech rate was negatively correlated with the use of adjectives for both the three and six object arrays, suggesting that speakers reduce their speech rate when producing overspecified descriptions. The results of this experiment provide preliminary evidence for a role of WMC in overspecified referring expressions, as well as expand on previous findings regarding the effect of scene characteristics on speech initiation times. Further, the results begin to establish a relationship between speech rate, speech onset, and the production of overspecified descriptions.

Of particular interest in this experiment was the relationship between cognitive load, WM, and overspecification. It was predicted that speakers would overspecify more for six object than three object arrays, that adjective use would increase as load increased, and finally that individuals with lower WMC would produce more overspecified descriptions than those with higher WMC. The lack of difference in rates of overspecification between three and six object arrays was surprising given the results of Koolen and colleagues (2016), who showed that adding visual clutter to a scene can increase the number of overspecified expressions. The results also appear to stand in contrast to Pechmann's (1989) theory that overspecification results from the incrementality of speech production. If, as he argues, speakers include prenominal adjectives in their descriptions before they completely scan the display to determine if the features are distinguishing, then one might expect that when there are fewer objects, and so less scanning is required, that speakers might be faster to determine whether a prenominal adjective is required, and therefore overspecify less. Further, according to Swets and colleagues' (2014) experiment, speakers with lower WMC spend less time gazing at competitor objects in the display. Therefore, if overspecification is related to decreased visual scanning, as observed by Pechmann (1989), then one might expect individuals with lower WMC to rely more heavily on this communicative strategy, but this effect was not present in the data. However, because we do not have information regarding the scanning behaviour of participants in this experiment, it is not possible to determine how thoroughly participants inspected the visual displays, and therefore difficult to determine the precise reason why the effects were not evident. One possible reason that the results of this experiment don't directly support the hypotheses is because the visual displays in the current experiment never included contrastive objects that would require participants to use adjectives at least sometimes to distinguish them. This may have led to participants using fewer adjectives overall and therefore masked possible differences in adjective use between high and low WMC individuals, as well as possible differences between the array sizes. Indeed, the overall proportion of overspecified referring expressions produced in this experiment was lower than what has been commonly found in past research (14% of all expressions in the current experiment included redundant colour adjectives, whereas other experiments have seen rates closer to 30%; e.g., Koolen et al., 2013; Tarenskeen et al., 2015).

Although the results of the current experiment do no support the above hypotheses in a straightforward manner, they provide preliminary evidence to suggest WMC plays a role in the use of redundant adjectives. There was a significant three-way interaction between array size, load condition, and operation span scores in cases where participants correctly recalled the numbers, indicating that individual differences in WMC may play a role in how speakers respond to manipulations of scene characteristics and cognitive load during language production. Specifically, for three object arrays, speakers with lower operation span scores produced more overspecified descriptions under low levels of cognitive load, but speakers with higher operations span scores did so more under high levels of load. This pattern of results indicates that individuals with lower WMC may be continuing to rely on overspecification for scenes with fewer objects under low levels of load, but as load increases they adjust their strategies to produce minimally specified descriptions. This suggests that when there are more objects in the scene, WMC does not appear to make a difference in rates of overspecification, but when there are fewer objects present, speakers with high and low WMC diverge in their choice of referring expressions depending on the level of load. This result is especially interesting considering the results of Swets and colleagues (2014) which suggested that individuals with lower WMC spend less time gazing at competitor objects in the display. It is possible that individuals with lower WMC overspecified more under low load only for the three object arrays because, even in cases when minimal inspection of the visual scene is required, they continue to focus primarily on the target object and therefore do not take the time to determine whether adjectives are required to distinguish the object. Further, the fact that individuals with higher WMC were more likely to overspecify for three object arrays under high load suggests that WMC may enable speakers to maintain a single referring strategy as load increases, whereas those with lower WMC opt to produce minimally specified descriptions that do not require the encoding of any additional information. Further, this suggests that the decision to omit adjectives in referring expressions is not based on one's perception of the optimal way to refer to an object, but rather results from increased processing demands placed on the speaker.

Although these results provide preliminary evidence for a role of WMC in referential communication, it is important to interpret them with caution. The interaction between array size, load level, and operation span scores was only significant in the analysis based on the subset of correct trials, and the significance of the simple effects in each load condition was only marginal. Because participants with lower WMC were less likely to correctly recall the numbers, they are somewhat underrepresented in this subset. This further reduced the overall number of overspecified referring expressions, making this analysis underpowered. Further research conducted on a larger sample size is required to determine whether the relationship between WMC and overspecification is robust.

In addition to examining individual differences in the production of overspecified descriptions, this experiment investigated speech planning, as measured by speech initiation times. The results showed a significant main effect of array size on speech onset times for both the initial onset of the utterance, as well as initiation of the noun phrase, but no effect of cognitive load or operation span. The results failed to replicate findings that cognitive load can impact speech onset times. Power (1985) gave participants the same memory task as the one used in this study (except, under the high load condition, participants were asked to remember six numbers instead of five) and found that speakers were faster to initiate speech under increased load. It is important to note that in Power's experiment, participants were given two words and asked to form a sentence from them. His task therefore required greater conceptual planning which could contribute to variation in speech initiation times, whereas in the current task, the sentences that participants were asked to produce followed a relatively strict structure.

Despite not finding significant differences in onset times for the load conditions, the difference in speech onset depending on the number of objects in the display is particularly interesting because it extends previous knowledge regarding the effect of distractor objects on speech planning. Previous work by Gatt and colleagues (2017) showed that speakers initiate their speech later when there are more objects in the display, but their effect was only significant in cases where participants were required to distinguish the target using a size or size and colour adjective. When speakers needed only to include colour adjectives, increasing the number of distractors did not affect speech onset times. In the current experiment, we show that the number of distractor objects does in fact have an overall effect on speech planning. Notably, the stimuli used by Gatt and colleagues (2017) were objects of all the same type and so required distinguishing adjectives for identification, whereas in the current experiment, all objects in the display were semantically unrelated, meaning that any use of adjectives to identify the target was unnecessary. Here we show a possible role for visual search in formulating referring expressions, even in cases where the target object does not require modification to be distinguished from distractors. The results presented here support the idea that visual search plays a role in message formulation, with a greater number of distractor objects inhibiting faster encoding.

Considering the effect of array size on speech onset times, the lack of differences in speech onset in relation to WMC is surprising. Swets and colleagues (2014) noted that participants with higher reading spans were faster to initiate speech for displays with no contrastive objects, similar to the displays used in this experiment which never contained semantically related or contrastive objects (although it should be noted that the effect found by Swets et al. (2014) was only marginal). Based on their findings, one might expect that, in the current experiment, participants with higher WMC would be faster to initiate their speech because of a more flexible planning scope. However, this did not turn out to be the case. This discrepancy in findings is unlikely to be a result of differences in the task because Swets and colleagues used a similar experimental task in which participants were asked to provide an instruction to another participant about how to move objects on a screen. Swets et al. further identified differences in scanning behaviour of high and low WMC speakers, noting that individuals with lower WMC spent less time examining the objects in the display. Again, given the results that participants are slower to initiate speech for six object arrays, when more visual scanning would be required, it is surprising that individuals with high WMC are not faster to encode their message. It has also been argued that working memory is what helps ground language in the visual scene (Huettig, Olivers, & Hartsuiker, 2011). Although this argument focuses primarily on visual search using the visual world paradigm in language comprehension, one might still expect WM to guide the visual search in language production, such that participants must link objects in their visual field to stored knowledge about how to identify the object, making it easier for individuals with higher WMC to determine the optimal description.

This experiment also sought to examine the relationship between overspecification and aspects of speech timing, in this case speech onset and speech rate. As noted above, there were no significant main effects of speech rate, but there were significant negative correlations between speech rate and adjective use, as well as speech onset times and adjective use. There was a significant negative correlation between speech onset and adjective use for three object arrays. This result supports the idea that adjective use results from incremental language production. When people begin to speak sooner, they presumably are spending less time examining the visual scene and cut down the amount of time available to decide on an appropriate object description. Although the relationship between speech onset and adjective use supports Pechmann's (1989) theory of overspecification, it is important to note that the correlation between NP onset and adjective use did not reach significance for the three object arrays. This suggests that, while speakers were faster to initiate overspecified instructions for three object arrays, they produced the NP at roughly the same time regardless of whether and adjective was used. However, for six object arrays, the opposite pattern was seen. The relationship between speech onset and overspecification was not significant for displays with six objects, but there was a marginal positive correlation between NP onset and overspecification indicating that speakers are somewhat slower to initiate the NP when the include adjectives in their descriptions. One possible explanation for this is that the inclusion of modifying terms in an object description requires additional conceptual planning, leading to pauses before the NP is produced. This may also explain why we found a significant negative correlation between speech rate and adjective use for both three and six object arrays (although this did not reach significance for any of the individual load conditions). This relationship suggests that speakers slow down their speech when they incorporate adjectives in their object descriptions, but it is possible that this slowed speech results from slightly longer pauses before adjectives are inserted. Overall, the relationships between the overspecification and speech onset, and overspecification and speech rate suggest that the inclusion of adjectives in object descriptions influences individual elements of speech planning, likely due to increased demands placed on conceptualization or word retrieval when including adjectives in the noun phrase.

Conclusion

The majority of previous research in referential communication has focused on communicative and contextual factors involved in the production of referring expressions. The current experiment expands on this body of research by demonstrating a role for WM in referential communication. Although the main effects of operation span, load condition, and array size did not reach significance, the significant interaction between these variables indicates differences in the way individuals with high and low WMC respond to manipulations of scene characteristics and task demands when formulating their referring expressions. This result, to our knowledge, is the first to implicate WMC in the production of overspecified descriptions, but requires further investigation in a larger sample size.

In addition, this research extended previous findings about the impact of the visual scene on speech onset times. Here we showed that, even in the absence of contrasting objects in the display, speakers take longer to plan their utterances when there are more distractor objects present. Finally, we show a negative relationship between the use of redundant adjectives and speech rate that is indicative of increased conceptual planning required to produce overspecified descriptions.

References

Arts, A. Maes, A. Noordman, L., & Jansen, C. (2011). Overspecification facilitates object identification. *Journal of Pragmatics, 43*(1), 361-374. doi:10.1016/j.pragma.2010.07.013

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*(4), 390–412. doi: 10.1016/j.jml.2007.12.005

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556-559. doi:10.1126/science.1736359

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews: Neuroscience, 4*(10), 829-839. doi:10.1038/nrn1201

Bates, D., Machler, M., Bolker, B., Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. doi:10.18637/jss.v067.i01

Brennan, S. E. & Clark, H. H. (1996). Conceptual pacts and lexical choices in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482-1493. doi: 10.1037/0278-7393.22.6.1482

Boersma, P. & Weenink, D. (2016). Praat: Doing Phonetics by Computer. Version 6.0.16, http://www.praat.org

Clarke, A. D. F., Elsner, M., & Rohde, H. (2013). Where's Wally: The influence of visual salience on referring expression generation. *Frontiers in Psychology*, 4, 329. doi:10.3389/fpsyg.2013.00329

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Willhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin and Review*, *12*(5), 769-786. doi:10.3758/BF03196772

Daneman, M. (1991). Working memory as a predictor of verbal fluency. *Journal of Psycholinguistic Research*, *20*(6), 445-464. doi:10.1007/BF01067637

Daneman, M. & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin and Review*, *3*(4), 422-433. doi:10.3758/BF03214546

Elsner, M., Clarke, A., & Rohde, H. (2018). Visual complexity and its effects on referring expression generation. *Cognitive Science*, *42*, 940-973. doi:10.1111/cogs.12507

Engelhardt, P. E., Bailey, K. G. D., & Ferreira, F. (2006). Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language*, *54*(4), 554-573. doi:10.1016/j.jml.2005.12.009

Engelhardt, P. E., Demiral, S. B., & Ferreira, F. (2011). Over-specified referring expressions impair comprehension: An ERP study. *Brain and Cognition*, *77*(2), 304-314. doi:10.1016/j.bandc.2011.07.004

Fehringer, C. & Fry, C. (2007). Hesitation phenomena in language production of bilingual speakers: The role of working memory. *Folia Linguistica*, *411*(1-2), 37-72. doi:10.1515/flin.41.1-2.37

Gatt, A., Krahmer, E., van Deemter, K., & van Gompel, R. P. G. (2017). Reference production as search: The impact of domain size on the production of distinguishing descriptions. *Cognitive Science*, *41*, 1457-1492. doi: 10.1111/cogs.12375

Goudbeek, M., & Krahmer, E. (2011). Referring under load: Disentangling preference-based and alignment based content selection processes in referring expression generation. In K. van Deemter, A. Gatt, R. van Gompel, & E. Krahmer (Eds.), *Proceedings of PRE-Cogsci: Bridging the gap between computational, empirical and theoretical approaches to reference*. Boston, 20 July 2011.

Grice, H. P. (1975). Logic and conversation. *1975*, 41-58.

Hartsuiker, R. J. & Barkhuysen, P. N. (2006). Language production and working memory: The case of subject-verb agreement. *Language and Cognitive Processes*, *21*(1-3), 181-204. doi:10.1080/01690960400002117

Hendriks, P. (2016). Cognitive modelling of individual variation in reference production and comprehension. *Frontiers in Psychology*, *7*, 506. doi:10.3389/fpsyg.2016.00506

Huettig, F. & Janse, E. (2016). Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. *Lanuage, Cognition, and Neuroscience*, *31*(1), 80-93. doi:10.1080/23273798.2015.1047459

Huettig, F., Olivers, C. N. L., & Hartsuiker, R. J. (2011). Looking, language, and memory: Bridging research from the visual world and visual search paradigms. *Acta Psychologica*, *137*(2), 138-150. doi:10.1016/j.actpsy.2010.07.013

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434-446. doi:10.1016/j.jml.2007.11.007

Kemper, S., Herman, R. E., & Lian, C. H. T. (2003). The costs of doing two things at once for younger and older adults: Talking while walking, finger tapping, and ignoring speech in noise. *Psychology and Aging*, *18*(2), 181-192. doi:10.1037/0882-7974.18.2.181

Koolen, R., Gatt, A., Goudbeek, M., & Krahmer, E. (2011). Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, *43*(13), 3231-3250. doi:10.1016/j.pragma.2011.06.008

Koolen, R., Gatt, A., Krahmer, E., van Gompel, R. P. G., & van Deemter, K. (2016). Viewing time affects overspecification: Evidence for two strategies of attribute selection during reference production. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 159-164)

Koolen, R., Goudbeek, M., & Krahmer, E. (2013). The effect of scene variation on the redundant use of color in definite reference. *Cognitive Science*, *37*(2), 395-411. doi:10.1111/cogs.12019

Koolen, R., Krahmer, E., & Swerts, M. (2016). How distractor objects trigger referential overspecification: Testing the effects of visual clutter and distractor distance. *Cognitive Science*, *40*(7), 1617-1647. doi:10.1111/cogs.12297

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2017). lmerTest Package: Tests in

Linear Mixed Effects Models. *Journal of Statistical Software, 82*(13), 1–26. doi:10.18637/jss.v082.i13

Martin, R. C., Yan, H., & Schnur, T. T. (2014). Working memory and planning during sentence production. *Acta Psychologica, 152*, 120-132. doi:10.1016/j.actpsy.2014.08.006

Masson Carro, I., Goudbeek, M., & Krahmer, E. (2014). On the automaticity of reduction in dialogue: Cognitive load and repeated multimodal references. n P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 976-981). Austin, TX: Cognitive

Moreno-Martinez, F. J. & Montoro, P. R. (2012). An ecological alternative to Snodgrass & Vanderwart: 360 high quality colour images with norms for seven psycholinguistic variables. *PLoS One, 7*(5), e37527. doi:10.1371/journal.pone.0037527

Otten, M. & van Berkum, J. J. A. (2009). Does working memory capacity affect the ability to predict upcoming words in discourse? *Brain Research, 1291*, 92-101. doi:10.1016/j.brainres.2009.07.042

Paraboni, I. & van Deemter, K. (2014). Reference and the facilitation of search in spatial domains. *Language, Cognition, and Neuroscience, 29*(8), 1002-1017. doi:10.1080/01690965.2013.805796

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics, 27*(1), 89-110. doi:10.1515/ling.1989.27.1.89

Peirce, J. W. (2007). PsychoPy – Psychophysics software in Python. *Journal of*

Neuroscience Methods, *162*(1-2), 8-13. doi:10.1016/j.jneumeth.2006.11.017

Power, M. J. (1985). Sentence Production and Working Memory. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *37*(3), 367-385. doi:10.1080/14640748508400940

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Redick, T. S., Shipstead, Z., Meier, M. E., Montroy, J. J., Hicks, K. L., Unsworth, N., Kane, M. J., Hambrick, D. Z., & Engle, R. W. (2016). Cognitive predictors of a common multitasking ability: Contributions from working memory, attention control, and fluid intelligence. *Journal of Experimental Psychology: General*, *145*(11), 1473-1492. doi:10.1037/xge0000219

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

Saryazdi, R., Bannon, J., Rodrigues, A., Klammer, C., & Chambers, C. (2018). Picture perfect: A stimulus set of 225 pairs of matched clipart and photographic images normed by Mechanical Turk and laboratory participants. *Behavior Research Methods*, 1-13, doi:10.3758/s13428-018-1028-5

Sedivy, J. (2003). Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, *32*(1), 3-23. doi:10.1023/A:1021928914454

Slevc, L. R. (2011). Saying what's on your mind: Working memory effects on sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(6), 1503-1514. doi:10.1037/a0024350

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 174-215. doi:10.1037/0278-7393.6.2.174

Swets, B., Jacovina, M. E., & Gerrig, R. J. (2014). Individual differences in the scope of speech planning: Evidence from eye- movements. *Language and Cognition*, *6*(1), 12-44. doi:10.1017/langcog.2013.5

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

Tourtouri, E., Delogu, F., & Crocker, M. (2017). Specificity and entropy reduction in situated referential processing. In G. Gunzelmann, A. Howes, T. Tenbrink & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 3356–3361). London: Cognitive Science Society

Unsworth, N., Heitz, R. P., Schrock, J. C ., & Engle R. W. (2005). An automated version of the operation span task. *Behavior Research Methods, 37*(3), 498-505. doi:10.3758/BF03192720

Vogels, J., Krahmer, E., & Maes, A. (2015). How cognitive load influences speakers' choice of referential expressions. *Cogntive Science, 39*(6), 1396-1418. doi:10.1111/cogs.12205

Westerbeek, H., Koolen, R., & Maes, A. (2015). Stored object knowledge and the production of referring expressions: The case of color typicality. *Frontiers in Psychology*, *6*, 935. doi:10.3389/fpsyg.2015.00935

Table 1

Summary Statistics for Three and Six Object Arrays in Each Load Condition

	Three Obj Array			Six Obj Array		
	No Load *M(SD)*	Low Load *M(SD)*	High Load *M(SD)*	No Load *M(SD)*	Low Load *M (SD)*	High Load *M (SD)*
Speech Rate						
All Data	5.14 (1.43)	5.21 (1.51)	5.27 (1.31)	5.03 (1.39)	5.19 (1.48)	5.16 (1.36)
Correct Only	5.14 (1.43)	5.27 (1.59)	5.27 (1.30)	5.03 (1.40)	5.17 (1.43)	5.03 (1.33)
Speech Onset						
All Data	1.57 (0.24)	1.60 (0.23)	1.62 (0.36)	1.66 (0.29)	1.67 (0.26)	1.65 (0.24)
Correct Only	1.57 (0.24)	1.59 (0.22)	1.59 (0.33)	1.66 (0.29)	1.67 (0.24)	1.64 (0.24)
NP Onset						
All Data	2.14 (0.39)	2.21 (0.46)	2.14 (0.45)	2.25 (0.48)	2.25 (0.43)	2.22 (0.45)
Correct Only	2.14 (0.39)	2.20 (0.51)	2.08 (0.41)	2.25 (0.48)	2.25 (0.42)	2.18 (0.46)
Adjective Use						
All Data	14%	13%	15%	16%	16%	13%
Correct Only	14%	10%	6%	16%	13%	4%

Table 2

Summary of Results for Linear Mixed Effects Analysis for Full Data Set

Effect	*Estimate*	*SE*	*df*	*t/Z*	*p*
Speech Rate					
(Intercept)	5.15	0.17	61.62	30.84	<0.001
O-span	-0.06	0.10	34.42	-0.63	0.530
Array Size	0.05	0.04	34.18	1.36	0.183
Load (N-L)	0.003	0.05	31.76	0.05	0.959
Load (N-H)	0.05	0.04	29.49	1.22	0.231
Array Size x Load (N-L)	-0.03	0.04	908	-0.75	0.453
Array Size x Load (N-H)	0.03	0.04	1045	0.61	0.540
O-Span x Array Size	0.001	0.03	1058	0.31	0.757
O-span x Load (N-L)	0.02	0.05	47.37	0.44	0.660
O-span x Load (N-H)	0.001	0.04	49.24	0.18	0.859
O-span x Array Size x Load N-L	-0.01	0.04	1090	-0.19	0.852
O-span x Array Size x Load N-H	0.01	0.04	1096	0.30	0.763
Speech Onset					
(Intercept)	1.63	0.03	43.56	63.06	<0.001
O-span	<0.01	0.02	34.43	0.02	0.981
Array Size	-0.03	0.01	23.76	-4.06	<0.001
Load (N-L)	<0.01	0.01	51.17	0.15	0.880
Load (N-H)	0.01	0.01	36.95	0.90	0.372
Array Size x Load (N-L)	<0.01	0.01	713.3	-0.19	0.848
Array Size x Load (N-H)	0.01	0.01	966.90	1.54	0.124

O-Span x Array Size	<0.01	0.01	33.57	-0.29	0.774
O-span x Load (N-L)	<0.01	0.01	42.91	-0.29	0.777
O-span x Load (N-H)	<0.01	0.01	43.38	0.29	0.775
O-span x Array Size x Load N-L	-0.01	0.01	1053	-1.06	0.290
O-span x Array Size x Load N-H	<-0.01	0.01	1071	-0.26	0.794
Noun Phrase Onset					
(Intercept)	2.20	0.05	62.07	48.90	<0.001
O-span	<-0.01	0.03	34.26	-0.14	0.886
Array Size	-0.04	0.01	26.00	-3.17	0.004
Load (N-L)	0.01	0.02	1121	0.68	0.497
Load (N-H)	-0.01	0.02	58.02	-0.58	0.563
Array Size x Load (N-L)	0.02	0.01	774	1.33	0.185
Array Size x Load (N-H)	<-0.01	0.01	1038	-0.31	0.756
O-Span x Array Size	<-0.00	0.01	33.33	-0.12	0.903
O-span x Load (N-L)	0.01	0.01	696.4	0.48	0.634
O-span x Load (N-H)	-0.01	0.02	53.74	-0.62	0.535
O-span x Array Size x Load N-L	-0.02	0.01	1115	0.26	0.296
O-span x Array Size x Load N-H	<0.01	0.01	1112	-1.05	0.792
Adjective Use					
(Intercept)	-3.61	0.63		-5.72	<0.001
O-span	-0.22	0.58		-0.39	0.697
Array Size	-0.11	0.15		-0.71	0.475
Load (N-L)	-0.13	0.19		-0.68	0.505
Load (N-H)	0.01	0.18		0.04	0.971

Array Size x Load (N-L)	-0.25	0.19	-1.35	0.177
Array Size x Load (N-H)	0.32	0.18	1.80	0.071
O-Span x Array Size	-0.15	0.13	-1.19	0.235
O-span x Load (N-L)	-0.13	0.18	-0.72	0.473
O-span x Load (N-H)	-0.28	0.18	-1.52	0.128
O-span x Array Size x Load N-L	-0.22	0.18	-1.20	0.229
O-span x Array Size x Load N-H	0.32	0.18	1.76	0.079

Note. Significance is tested with lmerTest using Satterthwaite approximations for degrees of freedom

Table 3

Summary of Results for Linear Mixed Effects Analysis for Correct Trials

Effect	*Estimate*	*SE*	*df*	*t/Z*	*p*
Speech Rate					
(Intercept)	5.12	0.17	59.14	30.04	<0.001
O-span	-0.03	0.10	42.21	-0.31	0.761
Array Size	0.07	0.04	42.30	1.60	0.116
Load (N-L)	0.07	0.06	34.22	1.20	0.238
Load (N-H)	-0.02	0.07	36.92	-0.29	0.772
Array Size x Load (N-L)	-0.05	0.05	587.50	-1.06	0.288
Array Size x Load (N-H)	0.06	0.06	780.99	1.14	0.255
O-Span x Array Size	0.01	0.04	779.23	0.29	0.771
O-span x Load (N-L)	0.04	0.06	74.42	0.74	0.461
O-span x Load (N-H)	0.01	0.07	57.49	0.18	0.857
O-span x Array Size x Load N-L	0.05	0.05	819.63	0.97	0.322
O-span x Array Size x Load N-H	-0.05	0.06	799.61	-0.84	0.401
Speech Onset					
(Intercept)	1.62	0.03	45.23	61.26	<0.001
O-span	0.01	0.02	37.20	0.31	0.757
Array Size	-0.04	0.01	26.79	-4.39	<0.001
Load (N-L)	0.01	0.01	63.52	0.81	0.419
Load (N-H)	-0.01	0.01	35.36	-0.58	0.564
Array Size x Load (N-L)	<-0.01	0.01	531	0.12	0.908
Array Size x Load (N-H)	0.01	0.01	771.30	0.58	0.565

O-Span x Array Size	<0.01	0.01	53.50	0.09	0.930
O-span x Load (N-L)	-0.01	0.01	64.71	-0.58	0.564
O-span x Load (N-H)	0.01	0.01	47.18	0.94	0.353
O-span x Array Size x Load N-L	-0.02	0.01	789.20	-1.64	0.102
O-span x Array Size x Load N-H	0.01	0.01	796	0.65	0.514
Noun Phrase Onset					
(Intercept)	2.19	0.05	61.83	45.27	<0.001
O-span	0.01	0.03	38.43	0.27	0.792
Array Size	-0.05	0.01	41.94	-3.38	0.002
Load (N-L)	0.02	0.02	86.19	0.95	0.343
Load (N-H)	-0.03	0.02	69.89	-1.36	0.179
Array Size x Load (N-L)	0.03	0.02	541.70	1.73	0.085
Array Size x Load (N-H)	-0.02	0.02	779	-1.16	0.265
O-Span x Array Size	<0.01	0.01	797.5	0.07	0.943
O-span x Load (N-L)	-0.01	0.02	79.46	-0.42	0.678
O-span x Load (N-H)	0.02	0.02	71.79	0.66	0.512
O-span x Array Size x Load N-L	-0.03	0.02	819.6	-1.60	0.110
O-span x Array Size x Load N-H	0.02	0.02	825.1	0.95	0.342
Adjective Use					
(Intercept)	-3.99	0.69		-5.77	<0.001
O-span	0.10	0.62		0.16	0.874
Array Size	-0.32	0.24		-1.32	0.188
Load (N-L)	0.12	0.32		0.39	0.700
Load (N-H)	-0.70	0.44		-1.59	0.111

Array Size x Load (N-L)	-0.10	0.30	-0.31	0.755
Array Size x Load (N-H)	-0.01	0.44	-0.02	0.983
O-Span x Array Size	0.15	0.25	0.60	0.548
O-span x Load (N-L)	-0.38	0.31	-1.23	0.220
O-span x Load (N-H)	0.35	0.48	0.74	0.461
O-span x Array Size x Load N-L	-0.80	0.31	-2.60	0.009
O-span x Array Size x Load N-H	1.27	0.48	2.63	0.009

Note. Significance is tested with lmerTest using Satterthwaite approximations for degrees of freedom

Table 4

Correlations Between Dependent Variables for the Full Data set

	Speech Onset	NP Onset	Adjective Use
Three Object Arrays			
All Data			
Speech Rate	-0.01	-0.48***	-0.12**
Speech Onset		0.64***	-0.13**
NP Onset			0.01
No Load			
Speech Rate	-0.11	-0.56***	-0.11
Speech Onset		0.60***	-0.14*
NP Onset			0.06
Low Load			
Speech Rate	-0.11	-0.61***	-0.14*
Speech Onset		0.48***	-0.13
NP Onset			-0.01
High Load			
Speech Rate	0.06	-0.30***	-0.10
Speech Onset		0.81***	-0.13
NP Onset			0.01
Six Object Arrays			
All Data			
Speech Rate	-0.07	-0.51***	-0.08*
Speech Onset		0.61***	-0.01
NP Onset			0.07
No Load			
Speech Rate	-0.10	-0.59***	-0.12
Speech Onset		0.60***	0.05
NP Onset			0.10
Low Load			
Speech Rate	-0.04	-0.47***	-0.13
Speech Onset		0.68***	-0.02
NP Onset			0.07
High Load			
Speech Rate	-0.06	-0.46***	-0.01

Speech Onset	0.56***	-0.07
NP Onset		0.03

Figure 1: Speech rate by array size and load condition for the full data set (a) and the subset of correct trials (b).

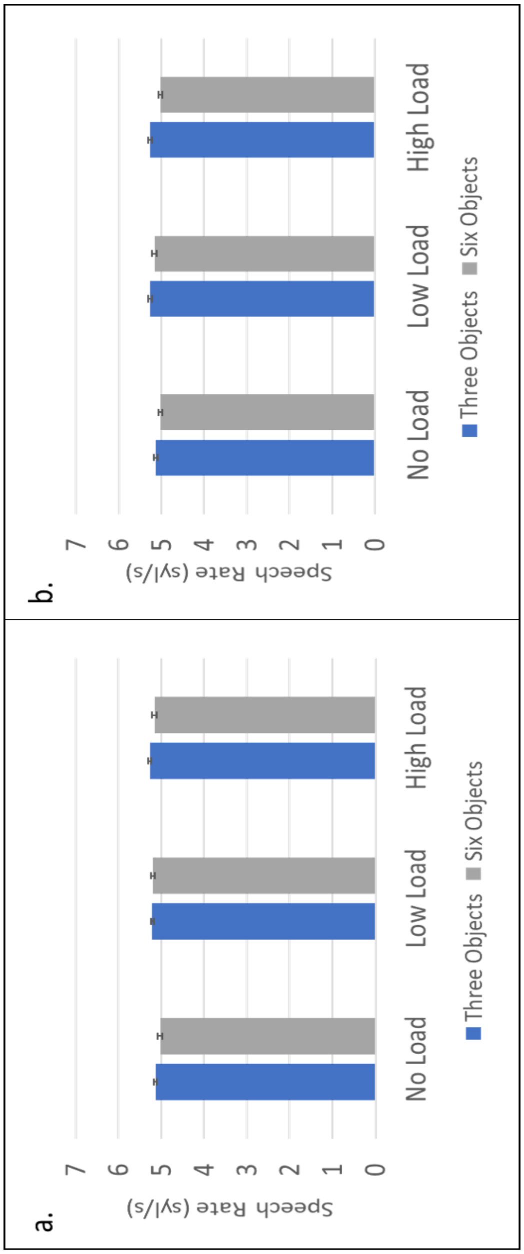

Figure 2: Speech onset times by array size and load condition for the full data set (a) and the subset of correct trials (b).

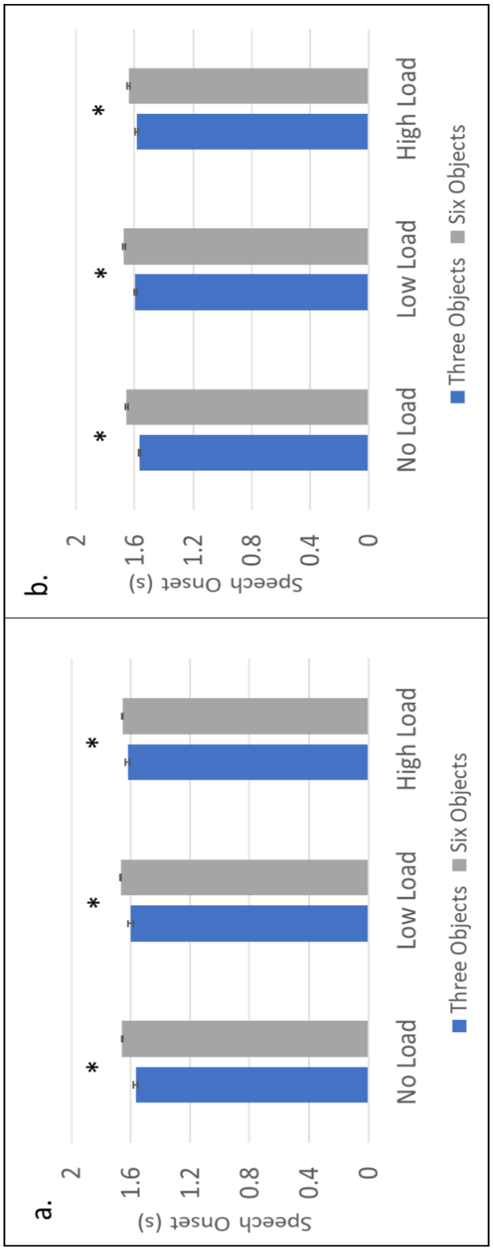

Figure 3: Noun phrase onset times by array size and load condition for the full data set (a) and the subset of correct trials (b).

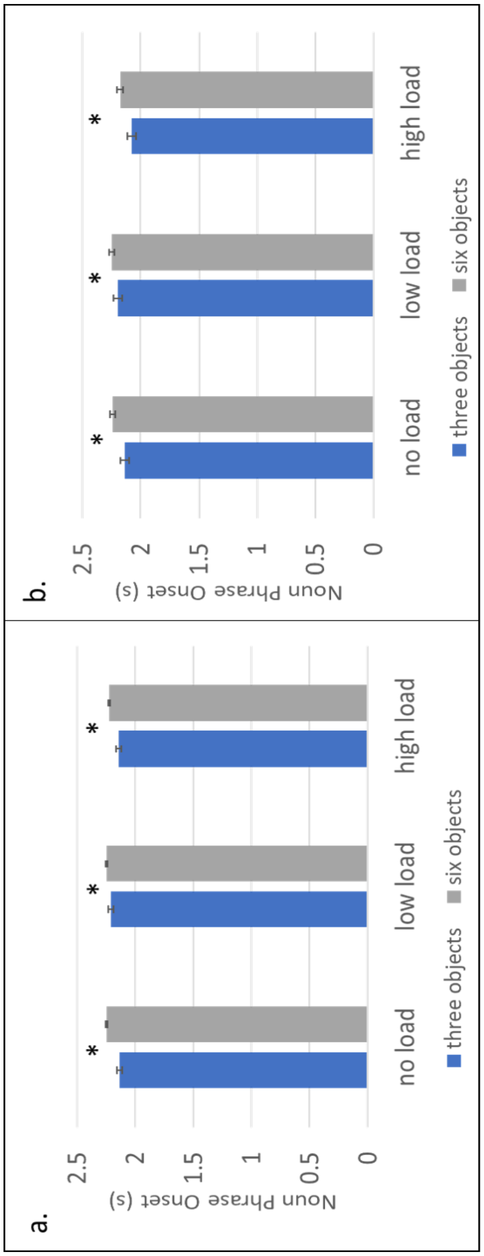

Figure 4: Overspecification by array size and load conditions for the full data set (a) and the subset of correct trials (b).

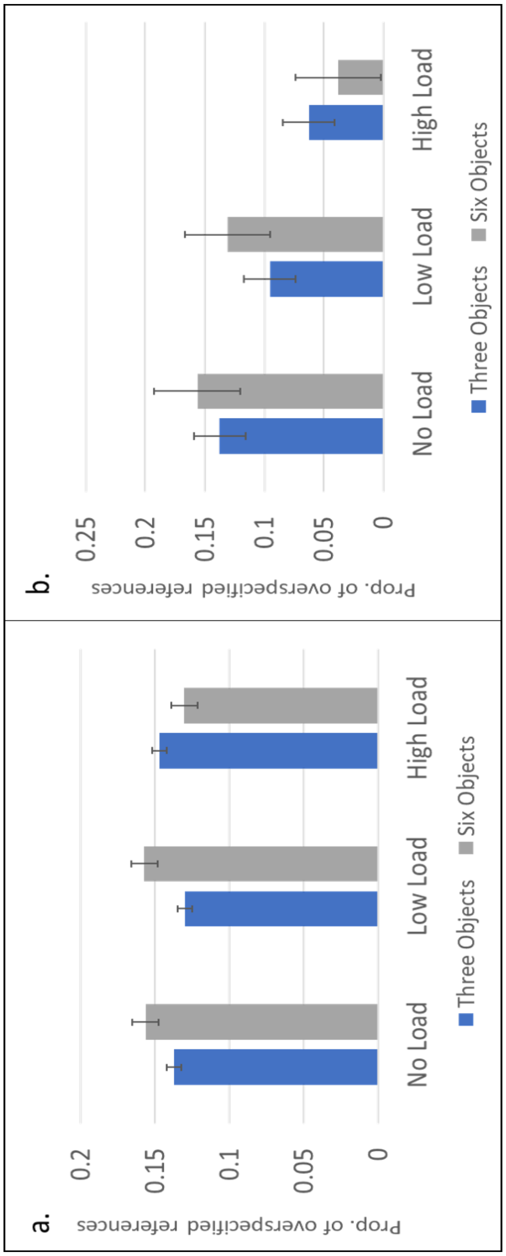

Figure 5: Overspecification by operation span scores for three object arrays in the no load (a), low load (b), and high load (c) conditions.

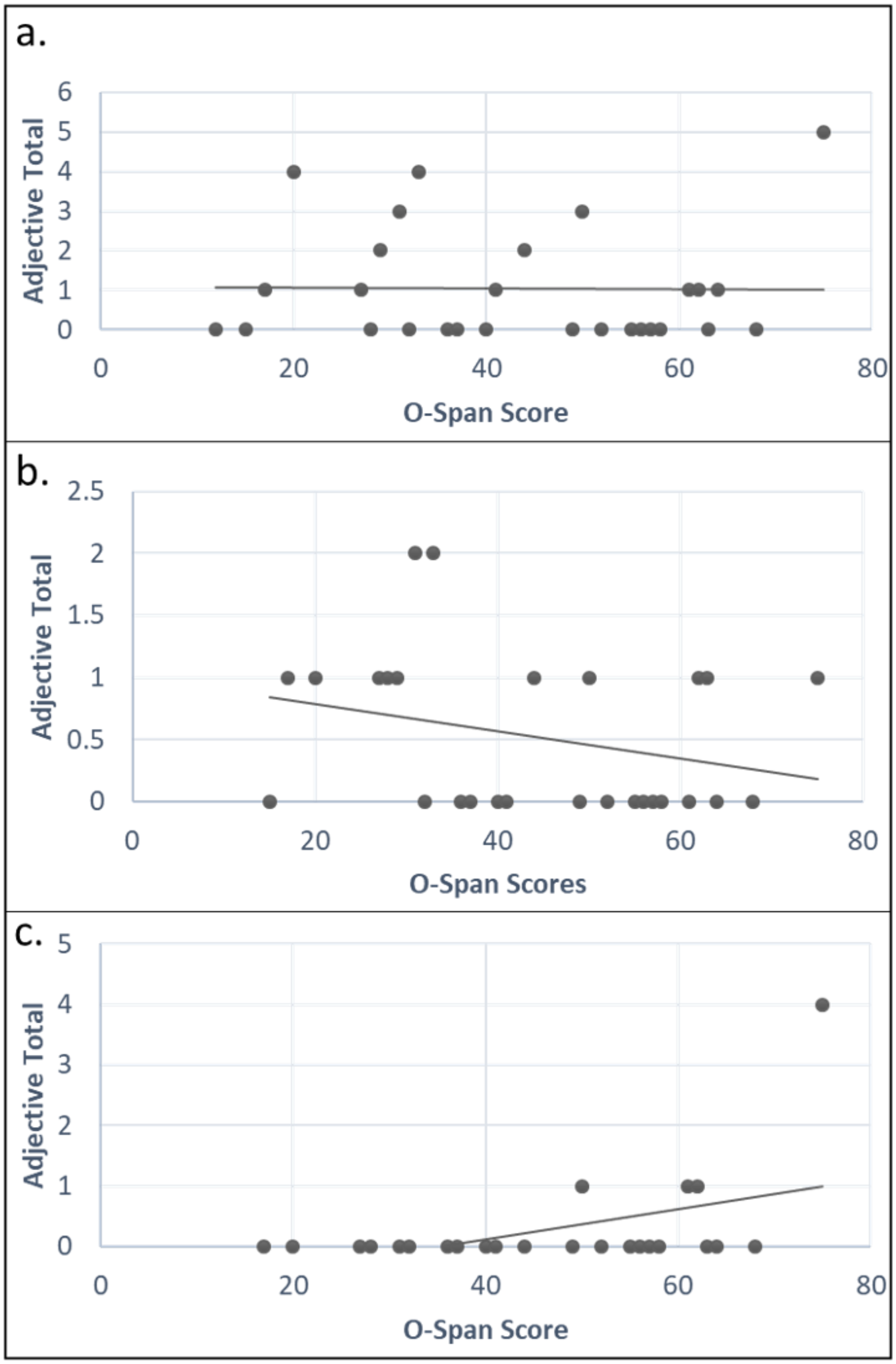

Figure 6: Boxplots for speech rate for three (a) and six (b) object arrays.

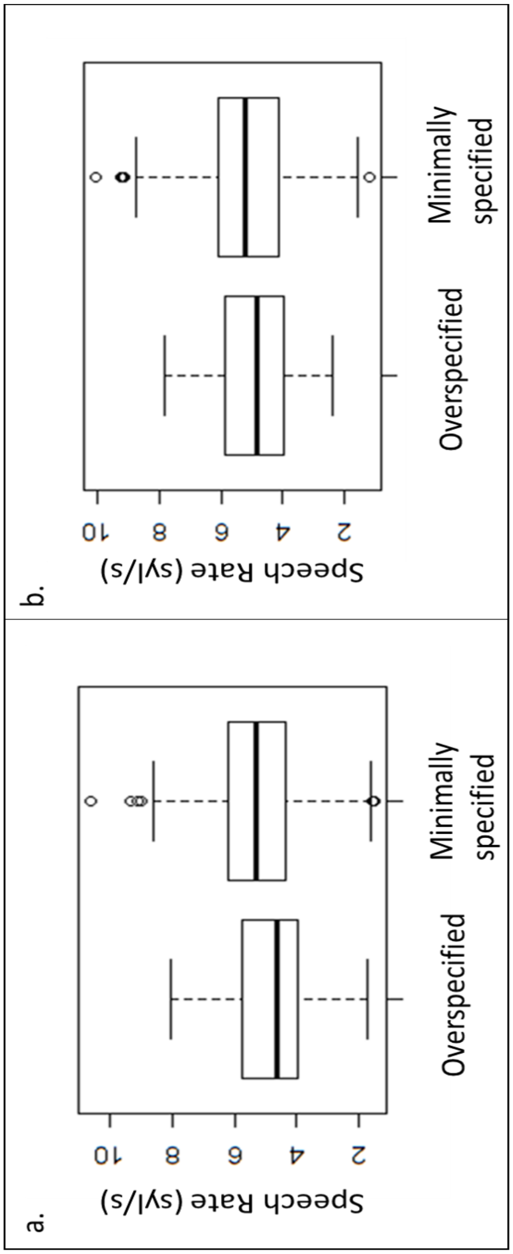

Figure 7: Boxplots for speech onset times for three (a) and six (b) object arrays.

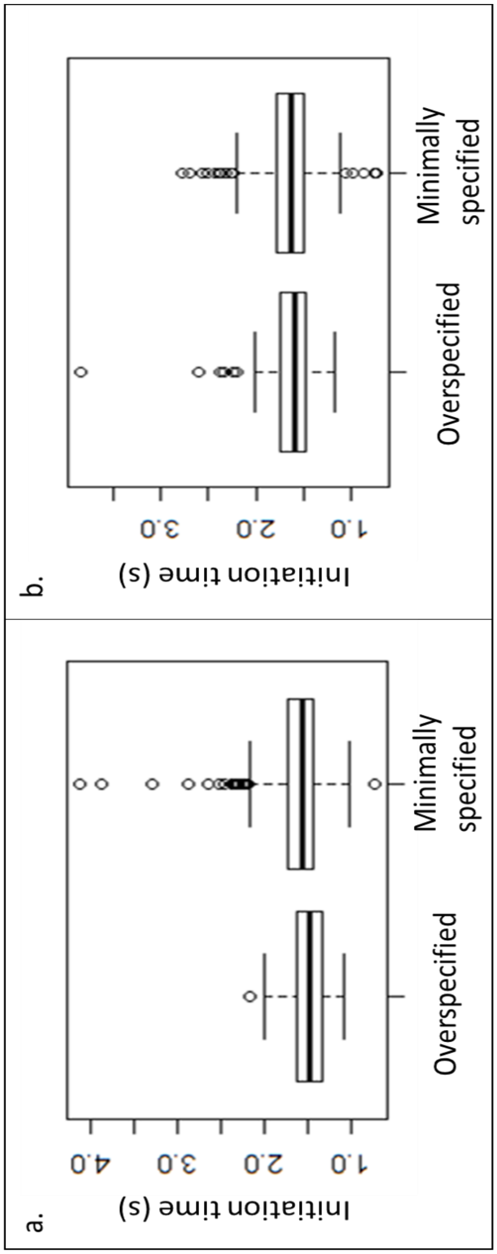

Figure 8: Boxplots for noun phrase onset for three (a) and six (b) object arrays