Is Retrieval Fluency a Heuristic in Audience Design?

Kieran John O'Shea, MA, MSc.

Submitted for the degree of Doctor of Philosophy, May 2017

Kieran J. O'Shea
Institute of Neuroscience & Psychology,
University of Glasgow,
58 Hillhead Street,
Glasgow,
G12 8QB.

Abstract

Across three experiments, we sought to test the key assumption of Horton and Gerrig's (2005a) memory-based model of common ground and audience design. Horton and Gerrig (2005a) argue that ordinary memory processes can serve as a proxy for more complex computations about common ground. Their key claim is that conversational partners act as memory cues for the retrieval of potentially relevant information through a process of resonance in episodic memory. Although studies have demonstrated effects in reference generation that are consistent with ordinary memory processes (Horton & Gerrig, 2005a), there has been no direct test to date of the key claim, which would require experimentally dissociating the effects of episodic memory from effects of common ground.

Similarly to Horton and Gerrig (2005a), we hypothesised that memory plays a crucial role in audience design. Influenced by the work of Gann and Barr (2014), we formed an alternative *retrieval fluency hypothesis* for audience design. We predicted that the fluency with which a speaker's expressions are retrieved would be dependent upon the degree that the referent and the retrieval context match the original encoding context (Tulving & Thomson, 1973). Our hypothesis proposed that expressions that were more fluent and had *stronger memory signals* would more likely be deemed contextually appropriate by the speaker – resulting in less consideration of context relative to expressions yielding weaker memory signals.

To test this we used a referential communication game, with participants playing as 'Director' providing descriptions to the 'Matcher' experimenter. In our first two experiments, we manipulated the *visual context* that target items appeared in. This was a communicatively irrelevant feature of the stimuli display that was presented to participants. Whilst these manipulations were salient to the Director they were not relevant to the actual description of the target objects. This enabled us to test whether visual features in the environment (that were irrespective of common ground) cued memory during audience design performance. In our third experiment, we manipulated the Director's perceptual experience – a communicatively relevant cue that is normally strongly correlated with common ground. In this study, we de-confounded the visual appearance of a potential addressee from the speaker's pragmatic knowledge of whom they were interacting with. Crucially, this enabled us to directly test the assumption that episodic effects are a key source of partner specificity in reference production (Horton, 2007; Horton & Gerrig, 2005a).

In Experiment 1, participants were shown a grid containing letters of various sizes and colours. We altered the appearance of the "competitor" and "foil" items, which alternated between training and test trials, so that participants had to adapt their descriptions at the test phase in order to avoid misspecifying descriptions. We expected speakers to experience greater retrieval fluency when the visual context in test trials was highly similar to the training trial configuration. It was predicted that this would result in them continuing to use the same description as before - making more descriptive errors than when presented with configurations that were dissimilar between the training and test phase. However, we found a lack of support for our hypothesis, as there was no main effect of visual context on reference production.

In Experiment 2 minor adjustments were made to the configuration and sequencing of objects and the stimuli presented to participants. In this experiment, Directors described pictures of everyday objects to the Matcher. Experiment 2 provided weak statistical support in favour of the retrieval fluency hypothesis for audience design and suggested that visual context impacted upon reference generation. More specifically, participants appeared to rely on the strength of the memory signal present when designing descriptions for the listener.

In Experiment 3, participants described target items to one of two Matchers using an interactive webcam design. At the test phase, the visual experience of the Director (participant) was controlled independently of the pragmatic situation, meaning that who the Director saw and whom they were speaking to did not always coincide. To the extent speakers use memory as a proxy for common ground, we expected misspecifications to be higher when participants viewed the same Matcher as they saw when they originally entrained on descriptions during training (effect of visual consistency). Furthermore, to the extent they use common ground, we expected misspecification to depend on their knowledge of who hears the description (effect of pragmatic consistency). Contrary to the memory-based model, there was no evidence that speakers misspecified more when viewing the same Matcher than when viewing a different Matcher. We also found no significant difference in misspecification rate when speakers believed that they had addressed the same Matcher versus a different Matcher.

In all three experiments we found a high misspecification rate in referential descriptions, indicating clear evidence of reliance on episodic memory. However, we did not find evidence in support of the retrieval fluency hypothesis for audience design. Our results also failed to support the key claims outlined in Horton and Gerrig's (2005a) memory-based

model. In particular, the results of Experiment 3 cast doubt on the assumption of partner specificity in audience design.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to take this opportunity to thank a few people. Firstly, I wish to thank my supervisor Dr. Dale Barr. Thank you for all your patience and support throughout my PhD and of course for introducing me to R programming! You have genuinely changed the way I think about science and data collection.

I wish to thank the ESRC (Economic and Social Research Council) for funding my doctoral work. I also wish to thank Dr. Sara Sereno for her continued support throughout my postgraduate studies here at Glasgow. I have always valued your advice and friendship. I would also like to give a special mention to the late Prof. Paddy O'Donnell – who always gave up his time to offer support to myself and many other junior academics at the University.

I wish to thank Dr. Dominique Knutsen for her insight and encouragement throughout the project. Thanks to Caitlyn Martin, David Ralston and Cameran Khalid for their help with data collection and their perseverance in assisting me to code thousands of speech production files!

My thanks also go to Dr. Ben Dunn, Dr. Christoph Scheepers, Dr. Phil McAleer, Dr. Maria Gardani, Marc Becirspahic, Stephanie Boyle and Dr. Gemma Learmonth for their support and good humour during my PhD work – it wouldn't have been the same without you! Thanks also to Dr. Lisa DeBruine and Prof. Ben Jones for welcoming me into the Face Research Lab and for their patience and understanding while I was finishing the write up.

My thanks goes to my parents, Sylvia and Eddie and my sister Caitlin. I'm very lucky to come from such a loving family who have always been supportive of my attempt to make a career out of academia – thank you for all you have done for me over the years!

Finally, I wish to thank my loving partner Ailsa. You've always been there for me through thick and thin, through happy days and grumpy days (and I know I've been very grumpy of late!). I'm extremely grateful to have your unconditional love and support.

Declaration

I declare that this thesis is my own work and was completed under the normal terms of supervision.

_____

Kieran J. O'Shea

Data reported throughout this thesis has been presented at the following conferences:

Conference Talks:

O'Shea, K.J., Ralston, D.F., & Barr, D.J. (2015). Does Retrieval Fluency impact upon Audience Design in Referential Communication? *Talk presented at the 30th Annual Conference of the British Psychological Society Postgraduate Affairs Group.* Glasgow, UK.

Conference Posters:

O'Shea, K.J., Martin, C.M., & Barr, D.J. (2017). Dissociating Effects of Common Ground and Episodic Memory on Partner Specificity in Production. *Poster presented at Annual Conference for Architectures and Mechanisms for Language Processing Conference (AMLaP)*. Lancaster, UK.

O'Shea, K.J., Ralston, D.F., & Barr, D.J. (2015). Is Retrieval Fluency a Heuristic in Audience Design? *Poster presented at Annual Conference for Architectures and Mechanisms for Language Processing Conference (AMLaP)*. Valletta, Malta.

O'Shea, K.J., Ralston, D.F., & Barr, D.J. (2015). Tailoring Descriptions to Suit the Listener's Needs: Does Retrieval Fluency Processing Impact upon Audience Design in Joint Communication? *Poster presented at 6th Joint Action Meeting.* Budapest, Hungary.

Chapter 1 –

Evidence for the Memory-Based Model of Referential Communication

1.1 – Audience Design and the Cooperative Principle

Reference is essential to successful communication – where a speaker attempts to enable the addressee to identify a particular referent in a given context (Horton & Keysar, 1996). Generally, we can assume that when someone speaks their main goal is to be successfully understood (Ferreira, 2008). However, successful communication between interlocutors often depends on the ability of the speaker to adapt their referential description to meet the addressees' informational needs. For example, this could involve a simple alteration whereby the speaker talks louder in a busy environment to ensure that their description is audible. Alternatively, this process may involve a more complex alteration where the speaker adapts their terminology to benefit a more inexperienced addressee (Ferreira, Slevc, & Rogers, 2005). This process of "tailoring" information for the conversational partner is known as *audience design* (Clark & Murphy, 1982). When engaging in audience design the speaker will take into account the listener's perspective in order to produce an utterance that the addressee is able to understand (Knutsen & Le Bigot, 2012).

The traditional view of audience design in language production argues that speakers are beholden to Grice's (1975) "Cooperative Principle". When an interlocutor successfully engages in audience design we can say that the speaker has fulfilled Grice's *Maxim of Quantity*:

1) Make your contribution as informative as is required (for the current purposes of the exchange).
2) Do not make your contribution more informative than is required.

According to these maxims speakers should consider the listener's current knowledge when deciding how to design their utterance (Horton & Gerrig, 2002). Consequently, speakers should strive to provide "optimal" descriptions to the addressee – providing the minimally sufficient information required for the listener to identify the referent within a shared context.

However, there are always numerous ways to describe the same referent (for example *"the blue denim jeans"* could refer to the same item as *"the darker pair of jeans"* or *"the pair*

on the left hanger") so how do speakers decide what level of information is sufficient within a given context? Or to put it more succinctly – how does the speaker decide whether a description is optimal or not?

1.2 – Establishing Common Ground

Tailoring a description to suit the listener's informational needs requires the speaker to account for audience-related factors such as previously established communicative conventions as well as the addressee's own expertise (Fussell & Krauss, 1989a). Most interactions occur between interlocutors who have varying levels of prior acquaintanceship. Accordingly, communicators will establish their own framework of mutual knowledge, which will be determined by the extent of the interactions they have shared in the past. The private knowledge they share, coupled with more general knowledge, can be used to formulate message construction and understanding (Fussell & Krauss, 1989b). The information both interlocutors share can be termed as their *common ground* – the beliefs, assumptions and mutual knowledge shared by both the speaker and the addressee (Clark & Carlson, 1981; Clark & Marshall, 1981). Thus in order for the speaker to tailor a description for the addressee, and therefore satisfy Grice's Maxims, they must be able to anticipate what information the listener already knows – they must establish the extent of their common ground with the listener (Fussell & Krauss, 1992).

During audience design speakers have been known to alter their speech to adapt their utterances for particular audiences (experts vs. novices: Isaacs & Clark, 1987, adults vs. children: Glucksberg, Krauss, & Weisberg, 1966, native vs. non-native speakers: Bortfeld & Brennan, 1997) and will often use the information that lies within their common ground to do so. The extent to which interlocutors share common ground will impact upon the ease of communication – particularly if the information they wish to discuss involves unusual topics or very specific details (Fussell & Krauss, 1989a). If there is a broad consensus of common ground it therefore follows that the communicative process should be simpler – the speaker should have considerably less difficulty finding the right phrase or terminology to express a subtle meaning (Fussell & Krauss, 1989a).

One aspect of dialogue that simplifies the process of audience design (and consequently can help to ascertain common ground between interlocutors) is the establishment of *linguistic/referential precedents* (Barr & Keysar, 2002; Kronmüller & Barr, 2007, 2015) or *lexical entrainment* (Brennan & Clark, 1996; Garrod & Anderson, 1987; Garrod & Doherty, 1994). In dialogue it is common to make reference to the same entities multiple times within a given discourse. Thus interlocutors come to associate particular referential

expressions with specific referents, such as calling a particular item of clothing "*the blue denim jeans"*.

Consider the following two excerpts from a conversation between Mark and Jane in the clothes department store:

(1)

M: Maybe you should look around more?

J: No…pass me the blue denim jeans.

M: Which ones?

J: Uh…the pair on the left hanger.

(2)

M: Maybe you should look around more?

J: No…pass me the blue denim jeans.

M: Okay but I prefer the black pair…

In Scenario (1) there appears to be a breakdown in communication between Mark and Jane. Jane asks for Mark to pass her *"the blue denim jeans"*. She appears to believe that Mark will understand her utterance. In this instance, however, it is clear that Jane and Mark have not yet established a common ground for referring to this particular item of clothing. As such, they have not yet established a referential precedent for "*the blue denim jeans"* which helps to explain why Mark seeks clarification from Jane *("Which ones?"*).

Scenario (2) contrasts with (1), as it appears that the two interlocutors have established a referential precedent for the phrase *"the blue denim jeans"*. Whether this is actually the case or not, Jane's reference succeeds anyway as Mark clearly understands which pair of jeans she is referring to. He responds by letting her know that he prefers a different pair ("*Okay but I prefer the black pair*"). Referential precedents simplify the speaker's task

because they need not decide how to conceptualise and describe a referent each time they encounter it. The process becomes easier as the speaker can just retrieve from memory the expression they used for that referent on a previous occasion - whilst doing some minimal checking to make sure that the precedent is still contextually adequate.

One hallmark of the existence of referential precedents is that once speakers have entrained upon a particular description, they will continue using that description even when the context has changed in a manner which makes the expression over-informative (Brennan & Clark, 1996; Deutsch & Pechmann, 1982). This overspecification can cause the speaker to violate Grice's (1975) *Maxim of Quantity* by providing more information than is required (Van Der Wege, 2009). Consider the following excerpt from Scenario (3) – Jane has purchased her *blue denim jeans* and is now looking to complete her outfit:

(3)

J: Do you think this t-shirt would match?

M: Maybe…how about this one?

J: I'm not sure it would go with the blue denim jeans.

M: Sure it would…it would definitely match your jeans!

Here we can see that whilst being unimpressed with Mark's choice of clothing, Jane continues to use the previously entrained description for her new jeans – *"the blue denim jeans"*. Considering that she has purchased her jeans and is no longer looking at similar items in the department store, Jane has no need to continue to refer to her purchase as *"the blue denim jeans"*. In this instance Mark notices the overspecification and begins to develop a new referential precedent of his own by shortening the description to "jeans" – *"It would definitely match your jeans!"*

Jane's reference to *"the blue denim jeans"* highlights a common trait among interlocutors – speakers are more likely to overspecify as the result of an existing precedent when speaking to an addressee who shares the precedent than when speaking to a new addressee (Brennan & Clark, 1996). This finding has been taken as evidence of partner-specificity of precedents, according to which speakers choose their expressions based on the information they believe is mutually held with the addressee (Brennan & Clark, 1996).

Most of the time conversations play out relatively smoothly. However, it is when addressees have difficulty following the speaker (such as the example in shown Scenario (1) above) that we can begin to see some of the problems that forming inadequate referring expressions can cause. If a speaker underspecifies their description by being too vague they can confuse the listener. On the other hand, overspecifying an utterance can prove to be unhelpful or even insulting for the addressee (Horton, 2008).

Generally, it is likely that speakers will try to abide by the *Maxim of Quantity* (Grice, 1975) and as a consequence of this, the listener will have particular expectations of the speaker (Clark, 1992). Listeners will expect the speaker to provide a suitable amount of information to enable referent identification and will therefore be perturbed by underspecifications (Engelhardt, Bailey, & Ferreira, 2006). As such, listeners will assume that speaker's descriptions have been optimally designed for their specific needs (Clark, Schreuder, & Buttrick, 1983). Thus if a modifier is used it should be relevant to the contextual setting (Engelhardt et al., 2006; Levison, 2000).

When misspecification does occur however, speakers appear more likely to overspecify their utterance rather than leave addressees with an underspecified description (Deutsch & Pechmann, 1982; Ferreira et al., 2005; Gann & Barr, 2014). Various studies have shown that overspecification is a common feature of referential descriptions and occurs when contextual support is available to speaker but not the addressee (Horton & Keysar, 1996; Nadig & Sedivy, 2002; Wardlow Lane & Ferreira, 2008; Wardlow Lane, Groisman, & Ferreira, 2006) and also when contextual support is completely unavailable to the speaker (Deutsch & Pechmann, 1982; Engelhardt et al., 2006; Pechmann, 1989).

For example, Deutsch and Pechmann (1982) found that overspecifications were commonplace and were produced frequently on over one quarter of trials in their study. The authors argued that these misspecifications were actually beneficial for the listener and that rather than hinder listeners' understanding, overspecification led to a more effective performance. Conversely, Engelhardt et al. (2006) found a similar rate of overspecification in their study (speakers overspecified descriptions on nearly one third of trials) but argued that participants' eye movements revealed confusion with overly-specific descriptions. Engelhardt (et al., 2006) and Sedivy, Tanenhaus, Chambers and Carlson, (1999), have argued that overspecifications may lead to a lack of comprehension and an impairment in communication.

Whilst there has been debate over the merit of overspecifying descriptions, research indicates that speakers will frequently adapt unsuitable or misspecified descriptions based

on feedback received from the addressee. For example, in a referential communication game, Horton and Gerrig (2002) had participants describe items to two separate matchers who had different subsets of knowledge. In test trials participants were tasked with describing referents to the alternative matcher from the one that they had established a precedent with. The authors found greater audience design after the second partner switch compared to the first switch. This indicated that the feedback speakers received from the first switch motivated them to consider the listener's needs more carefully in subsequent interactions (Barr & Keysar, 2006; Horton & Gerrig, 2002). In a similar referential task Gann and Barr (2014) found that participants relied on feedback (when available) to moderate their referential descriptions to addressees. However, when feedback was unavailable, speakers depended on their own self-assessments of referential adequacy. Gann & Barr (2014) suggest that in these instances, speakers will often rely on a process monitoring and adjustment to incrementally adapt utterances in order to suit the listener's referential needs.

1.3 – Clark's "Optimal Design" vs. The Monitoring and Adjustment Model

Herbert H. Clark and colleagues (Clark & Carlson, 1981; Clark & Marshall, 1981; Clark & Murphy, 1982; Clark et al., 1983; Clark & Wilkes-Gibbs, 1986) have provided the most influential account of common ground in communication. Similarly to Grice (1975), Clark et al. (1983) highlight the conversational goal of "*The Principle of Optimal Design*" (p. 246) - the speaker must design his utterance in a way which he believes is *optimal* for the listener. Accordingly, as a consequence of this principle, the listener must be able to understand the meaning of the utterance in coordination with the rest of the common ground they share with the speaker. It is argued that interlocutors use a series of *co-presence heuristics* to decipher what information lies within their common ground (Clark & Marshall, 1981; Clark & Murphy, 1982). These co-presence heuristics are used to short-circuit a potentially infinite recursive process and enable interlocutors to solve the *mutual knowledge paradox* (see Clark & Marshall, 1981 for a background summary).

The heuristics relied upon can be split into three main categories:

(a) *Community membership*: this depends upon information that is part of the socio-cultural background that two interlocutors share. Each shared community/sub-community (for example Mark and Jane are both Glaswegians) will have a common body of knowledge, assumptions and beliefs that those in that particular community will assume to be universally known. In the example above we can conclude that Mark and Jane both have a shared knowledge of the city of Glasgow.

(b) *Physical co-presence*: information that is shared or experienced in the physical environment. For example, Mark and Jane both visit the clothes department store together. The department store therefore forms part of their common ground. When Jane refers to something she sees on display, she can assume that Mark has a common understanding of the scene she is referring to.

(c) *Linguistic co-presence*: information that is shared as part of a conversation. Once Mark has understood which item Jane is referring to with the phrase "*the blue denim jeans*" both interlocutors can assume that this term (and the item associated with it) is now part of their common knowledge.

Clark and Marshall (1978, 1981) state that the complex process of definite reference requires a particular type of memory representation which helps the individual to encode whether information in a particular scenario meets the *triple co-presence heuristics*. Thus Clark and colleagues suggest that communicators use a *reference diary* to keep track of this process. In order for interlocutors to design and understand references they must consult their reference diary to do so (Clark & Marshall, 1978; Clark & Murphy, 1982). Accordingly, this "diary" helps an individual to keep note of the events in which they have taken part with others. Consider once again the interaction between Mark and Jane. In order for Jane to now refer to "*the cashier*" she must be sure that Mark had been present when she interacted with the cashier at the till of the clothes department store. If Jane does not have this event stored within her reference diary (or if she does not have another basis for common ground readily available) she cannot be certain that Mark will understand that "*the cashier*" is part of their common ground (Clark & Murphy, 1982). Using her reference diary Jane will tailor utterances towards her common knowledge with Mark. Furthermore, in accordance with this model, it is likely that when interacting with Jane, Mark will also confine the information he considers to mutual knowledge (Clark & Carlson, 1981).

The concept of a reference diary is appealing as it identifies memory encoding and retrieval as having a crucial role in the formation of descriptions in conversational common ground (Horton, 2008). That is, knowledge of one's own experiences in combination with an understanding of the knowledge and beliefs that others hold must be stored and retrieved in some manner. However, although idea of a reference diary is a useful construct, it does not fully explain how memory and common ground interact to help the speaker to produce optimal descriptions for the listener (Horton, 2008; Horton & Gerrig, 2016).

The main critique of Clarks et al.'s (1983) Principle of Optimal Design is that it assumes that communicators are capable of maintaining very detailed records of individuals, which are readily available in memory to help the speaker to design their utterances. Moreover, Clark's Optimal Design theory does not explain how an individual decides what the correct level of detail to encode would be, as evidence of triple co-presence is likely to be available in most instances (Horton, 2008). If individuals encoded triple co-presence in every possible occasion then the information stored in one's reference diary would quickly become representationally unbounded. Furthermore, if information was encoded in a more selective manner then it would be unclear what the selection criteria would be (Horton, 2008).

Horton and Keysar (1996) attempted to build on the insights offered by Clark et al.'s (1983) by proposing an alternative model that attempts to outline the role of common ground in language production. The authors argue that the Optimal Design Model is flawed as it focuses on the final product of the production system without considering the role of common ground in the production process. Horton and Keysar (1996) compare and contrast the *Initial Design Model* (incorporating the principle of optimal design proposed by Clark et al. 1983) to their alternative *Monitoring and Adjustment Model*. Whilst the Initial Design Model takes the addressees' perspective into account (the speaker uses only information which is incorporated in the common ground) the Monitoring and Adjustment Model does not consider common ground in the initial planning of utterances. Horton and Keysar (1996) argue that knowledge of what the conversational partner does or does not know may be too costly to use routinely when planning descriptions. Additionally, in some cases the information that is available to the speaker may already form part of the speaker's common ground with the listener.

Thus the Monitoring and Adjustment Model argues that speakers plan descriptions by using information that is readily available to themselves irrespective of whether the listener shares this information in their common ground with the speaker (Horton & Keysar, 1996). If a speaker adopts this model then it is likely that they will occasionally include information in their description that is not comprehensible to the listener. Horton and Keysar (1996) therefore assume that the speaker will monitor their speech and adjust any descriptions that contain content which lies outwith the mutual knowledge between the speaker and the addressee. The Monitoring and Adjustment model argues that common ground functions as a *correction mechanism* during referential communication.

Notably, even if speakers follow the alternative Initial Design Model (in line with Clark et al. 1983), there will be occasions where the speaker may make an error and produce an utterance that falls outwith the common ground. Therefore the role of monitoring in the Initial Design Model is simply to detect any errors made. Accordingly, in the Initial Design Model speakers rely on common ground as utterances are planned using mutual knowledge from the offset (Horton & Keysar, 1996).

Horton and Keysar (1996) directly tested both models in an experiment which required the participant to play the role of the speaker in a communication game with a confederate who played the role of the listener. Participants had to describe a series of objects for the confederate to identify. In order to tailor descriptions towards the listener's referential needs the speaker was required to occasionally add an adjective into their description when the stimuli appeared in the "shared context" condition (for example "it's the *small* circle"). However, on other occasions when the stimuli appeared in the "privileged context" it was not necessary for the speaker to provide an adjective in their utterance.

Horton and Keysar's (1996) study provided evidence showing that interlocutors followed the Monitoring and Adjustment Model. Results indicated that when participants were not under any time constraints they seemed to incorporate common ground in their descriptions. However, when time constraints were added, participants appeared to discard their consideration of common ground. These results suggest that under pressure speakers lack the sufficient resources and time to monitor their utterances for correction. As a result of this, they tend to fall back on their initial egocentric descriptions (Keysar, Barr, & Horton, 1998). Thus utterances which initially looked like they were specifically tailored for the listener only happened to appear like they were designed in such a way. Horton and Keysar (1996) argue that this is evidence that speakers were not engaging in audience design by accounting for common ground in the initial planning of descriptions – they were following the Monitoring and Adjustment model and adapting descriptions for the addressee when necessary.

1.4 – Common Ground in Comprehension

Although the initial proposal of common ground in language use (Clark & Carlson, 1981, 1982) was heavily challenged (Johnson-Laird, 1982; Sperber, 1982; Sperber & Wilson, 1982) most researchers now agree that it is a concept which plays an important role in comprehension (Keysar, Barr, Balin, & Paek, 1998). Since Clark and Carlson's early work there have been a number of influential studies that have developed the original theory and enhanced our understanding of common ground in referential communication. In

particular, studies that reveal how common ground impacts upon comprehension have provided important insights into how common ground affects language production in audience design.

For example, Keysar et al. (1998) conducted two experiments in which participants played the role of addressee and interpreted instructions from a confederate speaker. The authors introduced two alternative hypotheses that outline the role of common ground in audience design. *The Restricted Search Hypothesis* proposes that the search for referents in conversation is limited to items which are in common ground. Keysar et al. (1998) note that it would be logical for listeners to limit their search to referents within the common ground as speakers are expected to follow the principle of optimal design (Clark et al., 1983). Thus under this hypothesis, pragmatic knowledge of common ground will lead the search for conversational referents from the very beginning of the interaction (Keysar et al., 1998).

The Unrestricted Search Hypothesis offers an alternative view of the role of common ground. This hypothesis suggests that when addressees understand definite reference their search for referents is not guided by mutual knowledge. For example, under this hypothesis when Jane refers to "*the cashier*" when talking to Mark, Mark's unrestricted search will select an available "cashier" regardless of whether or not he/she is in common ground with Jane. This hypothesis is supported by previous findings (Horton & Keysar, 1996; Keysar, 1994, 1998; Keysar, Barr, & Horton, 1998) and suggests that under certain conditions comprehenders do not assume speakers follow the Principle of Optimal Design. Consequently, it is proposed that communicators do not rely on common ground unless they make an error (Keysar et al., 1998). Similarly to Horton and Keysar's (1996) Monitoring and Adjustment Model, the authors propose the *Perspective Adjustment Model.* This model argues that speakers monitor their descriptions and if a violation of common ground is detected their utterance plans are revised.

The results obtained supported the Perspective Adjustment Model. Reaction time and error rate data provided evidence for the Unrestricted Search Hypothesis – when participant's own privileged knowledge provided them with a potential referent which was inaccessible to the speaker, their unrestricted search caused greater response times and more errors when responding to the questions put forward by the speaker (Keysar et al., 1998). Furthermore, in a second experiment, results indicated that when a potential competitor referent was visible to the listener (but not to the speaker) saccade launch towards the

target item was delayed for an average of 180ms – further indicating that participants followed the Unrestricted Search Hypothesis (Keysar, et al., 1998).

These results suggest that when the interlocutors had differing perspectives the addressee's unrestricted search selected the wrong referent and required the listener to correct their initial search. Keysar et al. (1998) argue that the slow response times to correct mistakes reflected the interference of non-mutual referents. Accordingly, the Perspective Adjustment Model explains this pattern of results - common ground acts as a correction mechanism for interpretation errors (Keysar et al., 1998). Similarly to the participants in Horton and Keysar's (1996) study, addressees were shown to interpret descriptions from an egocentric perspective (Keysar, Barr, & Horton, 1998).

Keysar, Barr, Balin, and Brauner (2000) also found evidence indicating that interlocutors follow an "unrestricted search" when interpreting language in conversation. Keysar et al. (2000) proposed that addressees will occasionally use an egocentric approach which will lead them to consider potential referents which are not in common ground with the speaker. In this study participants played the role of the addressee in a communication game with a confederate director. The director received a photograph of the grid (showing where the objects were supposed to be placed) and instructed the addressee in moving the objects around the grid to match the photograph.

During the trial the director provided the addressee with an ambiguous instruction – for example "*move the small candle*". Importantly, the addressee had a shared perspective with the director that enabled them to view one potential referent. However, the addressee also had their own privileged perspective that provided an additional potential referent which was occluded from the director's view. It was hypothesised that if the participant initially considered the candle which was occluded from the director, this would suggest that the addressee was adopting an egocentric interpretation in their search for referents (see Figure 1 for example of stimuli).

The results of the eye tracking study revealed that participants fixated on the object (which was occluded from the confederate speaker) nearly twice as often when it contained a competitor referent (e.g. another candle) compared to the control condition when the location contained a non-referent. Furthermore, participants spent 242ms longer fixating on the occluded item in the competitor condition compared to the control condition (Keysar et al., 2000). The egocentric approach appeared to be so compelling for participants that it was able to override their knowledge that the speaker could not possibly see the occluded

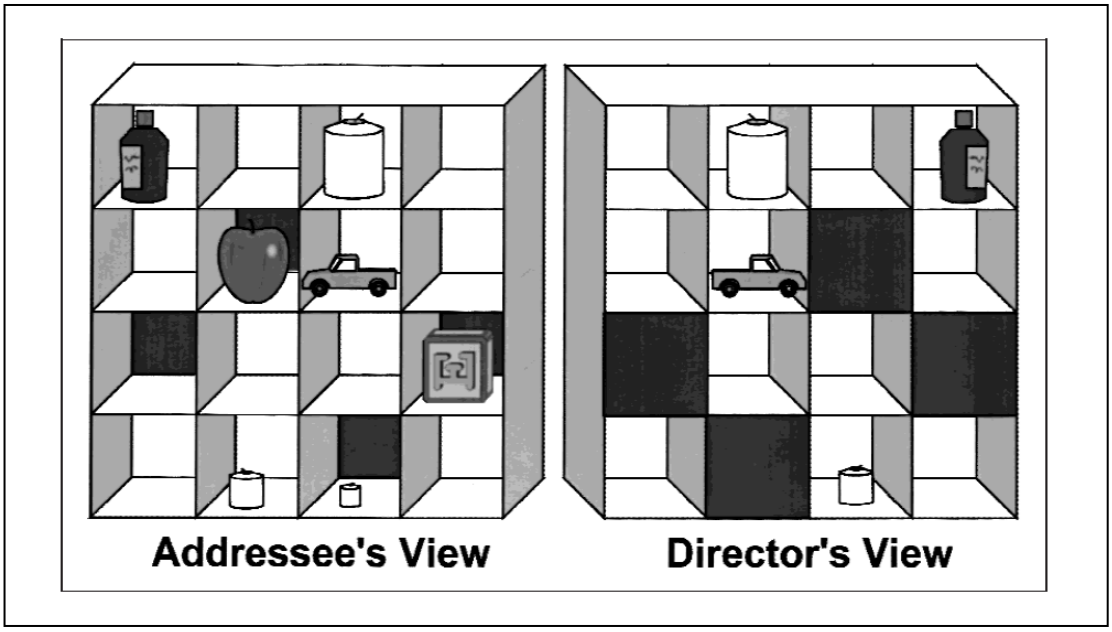


Figure 1: Stimuli from the Keysar et al. (2000) study. The occluded slots in the grid ensure that the addressee and director have distinct views of the grid. The addressee has privileged information as they can see behind the occluded slots which block the director's view. In this example the addressee hears a key instruction (referring to "the small candle"). Based on this description, the addressee may potentially pick out a different candle (the occluded candle) from the one the director is referring to (the shared candle). *Taken from Keysar, Barr, Balin and Brauner (2000)*.

item. Thus the results of Keysar et al. (2000) further demonstrate that listeners are prone to using an egocentric perspective when interpreting referential descriptions and do not always take into account their common ground with the speaker.

Whilst Keysar and colleagues provide substantial evidence which supports the Perspective Adjustment Model, both Hanna, Tanenhaus, and Trueswell, (2003) and Nadig and Sedivy, (2002) argue that these findings also support the *Partial Constraint Hypothesis*. This hypothesis assumes that common ground is one of a number of cues influencing the comprehension system. According to this model, the effects of common ground are immediate but only partial, as other cues may be available to the individual that provide additional information which is not in the common ground of the two interlocutors (Hanna et al., 2003; Nadig & Sedivy, 2002).

Nadig and Sedivy (2002) supported this hypothesis by recording the eye movements of five year old children whilst they played in a referential communication game with an adult confederate speaker. Similarly to Keysar et al. (2000), the authors found interference from private knowledge but also found strong evidence indicating that children consulted common ground in both comprehension and language production. Nadig and Sedivy (2002) argue that their findings indicate that children use rapid common ground constraints in comprehension and therefore refute Horton and Keysar, (1996) and Keysar et al.'s

(1998) suggestion that common ground is delayed to a later "monitoring" stage in processing. Hanna et al. (2003) found similar results when testing adult participants in an eyetracking study. During the early stages of comprehension listeners were more likely to look at the target shape which was in common ground compared to a matching shape which was only visible to the participant. Together these findings appear to support the Partial Constraint Hypothesis and suggest that interlocutors do not adopt a completely egocentric approach to referential communication (Barr & Keysar, 2006).

Notably, Pickering and Garrod's (2004) *Interactive Alignment Model* also provides an alternative account which differs from the traditional view of common ground posited by Clark and colleagues (Clark & Carlson, 1981; Clark & Marshall, 1981; Clark & Murphy, 1982; Clark et al., 1983; Clark & Wilkes-Gibbs, 1986). The Interactive Alignment Model proposes that conversational representations between interlocutors become aligned at different linguistic levels at the same time. Communicators do this by utilising each other's choice of sounds, words, meanings and grammatical forms (Garrod & Pickering, 2004). During referential communication the overlap between communicators' representations is such that a particular contribution by the speaker will result in the appropriate changes being made in the listener's own representation or will initiate the process of interactive repair (Pickering & Garrod, 2004). Interlocutors therefore build up a series of aligned representations which form the *implicit common ground* (information shared between interlocutors). The formation of implicit common ground means that communicators do not have to develop separate representations for themselves and their communicative partner (Garrod & Pickering, 2004).

Pickering and Garrod (2004) argue that speakers adapt their utterances only when information is accessible from their own situational model. This accessibility is from aligned representations which reflect the implicit common ground and can therefore be incidentally helpful to the listener. This idea is similar to previous research which has suggested that speakers can produce utterances that may appear to be helpful for the listener without the speaker actually designing their description with the listener in mind (e.g. Brown & Dell, 1987; Horton & Keysar, 1996).

Pickering and Garrod (2004) believe that implicit common ground is built up through an automatic process and is utilised in straightforward processes of repair. Communicators only rely on *full common ground* when it is absolutely necessary. Thus full common ground acts to repair misalignment. This interpretation is in line with the view of Horton and Keysar (1996) and Keysar et al. (1998) who argue that common ground acts as a

correction mechanism. Full common ground is predominantly only used in times of difficulty when interlocutors have become radically misaligned (Pickering & Garrod, 2004).

1.5 – Challenging Clark's Principle of Optimal Design

The research reviewed thus far indicates that although there are varying accounts detailing the role of common ground in referential communication most appear to differ with Clarks et al's. (1983) original view of "Optimal Design". As previously noted, the "Optimal Design" model assumes that speakers adhere to the Principle of Optimal Design which specifies that speakers will only include information in their description which is included in the common ground of the speaker and addressee (Clark et al., 1983). The studies outlined by Horton and Keysar (1996), Keysar et al. (1998); Keysar et al. (2000) and Pickering and Garrod (2004) diverge from this view and tend to support the idea that "full" common ground (Clark & Carlson, 1981; Clark & Marshall, 1981; Clark & Murphy, 1982; Clark et al., 1983, Clark & Wilkes-Gibbs, 1986) may be unnecessary for routine referential communication. Both Horton and Keysar (1996) and Pickering and Garrod (2004) agree that the process of assessing common ground is "too costly" to incorporate regularly into every single interaction with another interlocutor. Instead consideration of common ground is viewed as an optional process which may be undertaken by the speaker when resources are not too taxing.

Keysar et al. (1998) go further by challenging Clark and Carlson's (1981) assumption about optimality in common ground. Keysar et al. (1998) argue that their *Perspective Adjustment Model* may in fact be considered "optimal" if one accounts for the additional cost associated with consulting common ground throughout an interaction in Clark and Carlson's (1981) "Optimal Design" approach. The additional demand that common ground places on an individual's cognitive resources may make following the Perspective Adjustment Model worthwhile – even if it results in the occasional referential error (Keysar et al. 1998).

The idea that speakers choose their utterances based on information which is more readily accessible to themselves, rather than their addressee, is supported by a large variety of evidence suggesting egocentric tendencies in language production (Engelhardt et al., 2006; Ferreira & Dell, 2000; Gann & Barr, 2014; Wardlow Lane & Ferreira, 2008; Wardlow Lane, Groisman, & Ferreira, 2006; Wardlow Lane & Liersch, 2012). These findings will be reviewed in more detail in Chapter 4. Importantly, they imply that speakers will frequently include information in their descriptions that is unhelpful or misleading for the

listener (Gann & Barr, 2014). For instance, there is no evidence that speakers who have entrained on calling a very typical candle as "*the unmelted candle*" are any more likely to revert spontaneously and autonomously to the basic-level description "*the candle*" when the precedent is not in common ground with the listener as compared to when it is (Gann & Barr, 2014, see also Brennan & Clark, 1996). Thus consideration of the addressee's informational needs is only one factor which governs whether or not a speaker continues to follow an established precedent or whether the speaker tailors their description to suit the current context of the interaction.

1.6 – A Memory-Based Approach to Common Ground and Audience Design

Following this initial debate, Horton and Gerrig (2005a) introduced an alternative model which reconceptualised the role of common ground in referential communication. In their influential paper "*Conversational Common Ground and Memory Processes in Language Production*" Horton and Gerrig argue that the characteristics frequently attributed to conversational common ground are actually properties of *ordinary memory processes*. Crucially, the memory-based model emphasises the role that ordinary encoding and retrieval processes play in communication (Horton, 2008; Horton & Gerrig, 2005a, 2005b).

The authors outline two separate processes - *Commonality Assessment* and *Message Formation*, which they argue represent the different aspects involved in tailoring descriptions for addressees. Horton and Gerrig (2005a) identify both commonality assessment and message formation as playing a key role in audience design. When a speaker considers commonality assessment they take into account the likelihood that a specific piece of information is shared with the addressee. For example, when Jane turns to Mark and says *"I'm going to Naomi's flat later"* she assumes that Mark knows who *"Naomi"* is. According to the memory-based model, commonality assessment frequently develops from the speaker's automatic recognition that particular information can be considered familiar or not with a specific context. This apparent familiarity can also influence message formation – with speakers more likely to use certain forms of reference if the appropriate linguistic representations are accessible at that particular time (Horton & Gerrig, 2016). Importantly, when the speaker engages in message formation they consider how best to construct their description in relation to their commonality belief. Thus when Jane refers to "*Naomi"* she has to consider whether this utterance is the most effective way of referring to her friend. Without providing any surname or additional detail Jane assumes that Mark can uniquely identify "*Naomi"* by her first name alone.

Horton and Gerrig (2005a) note that although both commonality assessment and message formation are related they involve separate and unique aspects of audience design. Jane's belief that she shares knowledge with Mark differs from her consideration of how to design utterances which account for this belief (Horton & Gerrig, 2005a). Importantly, the success of the speaker's *memory retrieval* will determine whether commonality assessment functions effectively or not. Commonality assessment is dependent upon the normal episodic memory traces that are encoded during everyday interactions. Conversely, message formation is influenced by the speaker's estimation of the information which is accessible in the addressee's own memory (Horton & Gerrig, 2005a).

In addition to this, Horton and Gerrig (2005a) argue that the establishment of both commonality assessment and message formation as separate concepts helps to identify two possible ways in which audience design could fail. Firstly, audience design may fail if the speaker incorrectly assumes commonality between themselves and the listener. For example, if Mark replies to Jane by saying -*"Naomi…who?"* it becomes clear that Jane has incorrectly assumed that Mark shares commonality with her. Alternatively, Jane may provide too much detail and assume that she does not share commonality with Mark - *"I'm going to Naomi Mawson's flat later"*. This may even cause Mark to correct Jane – *"Yes I know who Naomi is!"*

Secondly, audience design can fail due to the speaker's inability to successfully adjust their message formation. In such an instance the speaker will provide an utterance which is unsuccessful in specifying who the intended referent is, despite the referent being mutually known to both interlocutors. In this case, Mark would have to seek clarification from Jane – *"Which Naomi are you talking about?"* Arguably, both of these possible failures in audience design highlight ways in which the speaker may adopt a more egocentric approach to language production by producing utterances which are comprehensible to themselves without fully accounting for the addressee's referential needs.

1.7 – Partner Specificity in Audience Design

Central to Horton and Gerrig's (2005a) theory is the idea that conversational partners can act as *memory cues* for the retrieval of information. This retrieval takes place via a process known as *resonance* – a quick, passive and effortless mechanism that enables cues in working memory to interact in parallel with information stored in long term memory (Horton, 2008; Ratcliff, 1978). Previously, Brennan and Clark (1996) proposed a similar idea to this by underlining the role of *partner specific* conceptual pacts between interlocutors. Accordingly, Brennan and Clark (1996) argue that when communicators

entrain on a description, the mapping between the referent and the entrained expression is linked with the interlocutors involved in the entrainment, thus making it part of their common ground (Brennan & Clark, 1996; Brown-Schmidt, 2009; Clark, 1992, 1996, Clark & Marshall, 1978, 1981). Similarly, Horton and Gerrig (2005a) argue that individuals function as highly salient cues and can enable the automatic retrieval of associated information. Crucially, according to this model, memories that are most frequently and consistently associated with a particular cue will be most likely to be available for reference production (Horton, 2007; Horton & Gerrig, 2005a).

Following Horton and Gerrig's (2005a) initial paper, Horton (2007) argued that conversational partners can act as contextual cues in the same manner that different rooms or physical contexts can cue automatic retrieval of information. In a picture-naming task, Horton (2007) found that naming latencies were shortest for responses which were associated with the original partner the description had been entrained with compared to descriptions associated with a new conversational partner. In this study, Horton (2007) suggests that the salience of conversational partners as memory cues influences the accessibility of lexical and conceptual information associated with that individual even in the absence of an intent to communicate with that person. Thus the key idea behind the memory-based model is similar to that of Brennan and Clark (1996): if an interlocutor develops a strong enough association between their conversational partner and relevant information there is a high likelihood that the information will be regarded as shared knowledge between both communicators (Horton, 2008). However, whilst this is an appealing idea, recent work by Brown-Schmidt and Horton (2014) failed to replicate Horton's original findings - raising some doubt over the proposal that conversational partners can act as memory cues in referential communication.

Brennan and Hanna (2009) highlight that the memory-based model gains support from studies which show that common ground established with a specific partner can be considered in the earliest moments of language processing (Hanna & Tanenhaus, 2004; Hanna, Tanenhaus & Truswell, 2003; Metzing & Brennan, 2003; Nadig & Sedivy, 2002). However, the authors also note that the memory-based model's assertion of partner specificity is incompatible with Pickering and Garrod's (2004) alignment theory which argues that precedent, not the speaker's identity, is important. Partner specificity also lacks support from two-stage models, which argue that early language processing is egocentric in nature and that partner specific adjustments materialize later as more effortful

amendments or repairs (e.g. Brown & Dell, 1987; Ferreira & Dell, 2000; Horton & Keysar, 1996; Keysar et al., 2000; Keysar, Barr, Balin, et al., 1998; Kronmüller & Barr, 2007).

In line with the two-stage model approach to reference production, researchers have found an overall lack of empirical support for the memory-based model. For example, Barr and Keysar (2002) failed to find evidence supporting the role of partner specificity in entrainment. The authors argued that if entrainment is partner specific then a precedent established with a speaker should be constrained when an entirely new speaker uses a previous expression. Barr and Keysar (2002) predicted that a new speaker would cause addressees to be slower to look at and reach out for target objects in their experiment. The results of the study showed that addressees were equally as fast to look at and reach out for objects irrespective of whom the speaker was. The authors concluded that this was because addressees relied on referring precedents because they were available in memory and not because they were partner specific (Barr & Keysar, 2002). Additionally, further evidence indicates that entrainment effects reflect general expectations about language use which are not linked to a listener's partner specific beliefs (Kronmüller & Barr, 2007).

However, other researchers have challenged these findings. For example, Metzing and Brennan (2003) questioned the methodological validity of Barr and Keysar's (2002) results and found evidence for partner specificity in memory using a similar paradigm. Furthermore, Brown-Schmidt (2009) suggests that the lack of live interaction between the participant and confederate in Barr and Keysar's (2002) study may have impacted upon performance in their experiment. In Barr and Keysar's (2002) design participants moved objects around a grid according to the instructions provided by a confederate. Brown-Schmidt (2009) argues that this prevented participants from collaboratively establishing entrained descriptions. Additionally, Brown-Schmidt, Yoon, and Ryskin, (2015) provide a similar argument noting that the conversational partner is more likely to be encoded with information when they are communicatively relevant to the conversation. The authors suggest that this enables the partner to become more strongly bound in memory. Brown-Schmidt et al. (2015) note that partner specific effects are absent or reduced in experiments that incorporate limited partner interaction in their design (e.g. Barr, 2008; Barr & Keysar, 2002; Brown-Schmidt, 2009; Brown & Dell, 1987; Kronmüller & Barr, 2007) in comparison to studies which involve extensive interactions between participants and show greater partner specific effects (Brown-Schmidt, 2009; Hanna et al., 2003; Heller, Grodner, & Tanenhaus, 2008; Lockridge & Brennan, 2002).

Taking into account the disparity in these findings we felt it was necessary to further test the concept of partner specificity in audience design. In the remainder of this thesis, I set out to investigate an alternative *retrieval fluency hypothesis* which seeks to further our understanding of how memory influences audience design and tests some of the key assumptions of Horton and Gerrig's (2005a) memory-based model.

1.8 – Thesis Motivation and Hypothesis

Our decision to develop an alternative hypothesis is motivated by a lack of conclusive evidence in favour of Horton and Gerrig's (2005a) memory-based model. Several studies have failed to support the assumption of partner specificity in common ground (e.g. Barr & Keysar, 2002; Kronmüller & Barr, 2007, 2015). Other research has shown that partner specificity only occurs in interactive dialogue settings and suggests that stimulus characteristics and the number of critical trials in the study may also effect the outcome (Brown-Schmidt, 2009). Furthermore, we note that support for the memory-based model has frequently been based on Horton's (2007) study (e.g. Brown-Schmidt, 2009, 2012; Brown-Schmidt et al., 2015; Gorman et al., 2013; Horton, 2008; Horton & Slaten, 2012). Notably, the findings from Horton's study are characterised by a low effect size and have recenlty failed to replicate (Brown-Schmidt and Horton, 2014).

As mentioned previously, resonance plays an important role in Horton and Gerrig's (2005a) theory and helps to facilitate the concept of partner specificity in audience design. Since resonance involves a parallel search of memory, this makes it possible for a range of associated information to become available on the basis of relatively local cues (Horton, 2008). Horton (2008) notes that the memory-based model draws on previous evidence from the memory literature. For example the "Search of Associative Memory" (SAM, Gillund & Shiffrin, 1984) and the "Retrieving Effectively from Memory" (REM, Shiffrin & Steyvers, 1997) models both identify memory retrieval as being a cue dependent search of long-term memory. In particular, the REM states that contextual information available when encoding is very likely to be be incorporated as part of relevant memory traces (Horton, 2008; Shiffrin & Steyvers, 1997).

Accordingly, along with partner specificity, the memory-based model therefore suggests that interlocutors will store additional episodic representations of the contextual information of a conversation in their memory (e.g. context of surroundings, lighting in the room, colour of objects) and depending on the strengh of these memories, these factors should all influence how the speaker produces a description for the listener. Crucially, although some authors (e.g. Gorman, Gregg-Harrison, Marsh, & Tanenhaus, 2013; Hanna

et al., 2003; Metzing & Brennan, 2003) have found evidence supporting partner specificity in common ground, we note that research thus far has failed to account for the effect that these additional episodic representations may have on audience design performance. In order to determine whether partner specificity plays a significant role in audience design, it is important to de-confound these additional contextual effects available in memory, from common ground. Thus to provide more conclusive evidence in favour of Horton and Gerrig's (2005a) memory-based model, and in particular their supposition of partner specificity in audience design, experiments testing this theory must be able to distinguish between effects of memory and effects of common ground. If additional episodic representations are not controlled for, merely showing that memory can impact upon communication does not provide support for Horton and Gerrig's (2005a) memory-based model.

In this thesis, I set out an alternative hypothesis which proposes that during audience design, rather than repeatedly consulting their common ground with a conversational partner, speakers make snap judgements regarding the contextual appropriateness of a referring expression using *heuristic assessments*. We suggest that speakers will often avoid generating new descriptions by using a form of attribute substitution (Kahneman & Frederick, 2002) – using a previous description that is more readily available in their memory. Thus speakers will often provide descriptions which appears to be shaped with the addressee's informational needs in mind, when in fact they are actually basing their utterances on the heuristic attribute of "ease of recall" (Barr, 2014). In particular, we test whether speakers judge the appropriateness of a given expression as a function of *retrieval fluency* - the relative ease or difficulty with which they are able to process information (Oppenheimer, 2008). A key factor which may influence the speaker's likelihood to use the retrieval fluency heuristic is the impact that episodic representations (contextual cues available in the environment e.g. colour of objects, visual similaity between past and present contexts) may have on memory. Our hypothesis accounts for the effect these representations may have during audience design and therefore serves as a further test of Horton and Gerrig's (2005a) memory-based model. In the following chapter, I will outline our retrieval fluency hypothesis in further detail and provide an overview of the logic and design of the experiments that will follow.

Chapter 2 –

The Retrieval Fluency Hypothesis

2.1 – Retrieval Fluency as a Theoretical Concept

As outlined in Chapter 1, our alternative hypothesis enables us to further test the key assumptions of Horton and Gerrig's (2005a) memory-based model. Our hypothesis proposes that rather than continually consulting their common ground with an addressee during audience design, speakers make snap judgments regarding the contextual appropriateness of a referring expression using heuristic assessments. Following recent work by Gann & Barr (2014), we investigate the hypothesis that speakers judge the appropriateness of a given referring expression as a function of *retrieval fluency* - of how easily that expression comes to mind (Oppenheimer, 2008) when attempting to linguistically encode the referent. However, before we outline our retrieval fluency hypothesis in full, it is important to highlight the research that has influenced the development of our theory.

The notion of fluency as a cue in decision making has a long history (Alter & Oppenheimer, 2009; Oppenheimer, 2008), but it has received little attention in the context of audience design and language production. Processing fluency is defined as an individual's subjective experience of the ease or difficulty with which they are able to process information (Oppenheimer, 2008). According to (Alter & Oppenheimer, 2009, p. 220) all cognitive tasks can be labelled along a continuum from "effortless" to "highly effortful" which creates a parallel metacognitive experience ranging from *"fluent"* to *"disfluent"*. Furthermore, Alter & Oppenheimer (2009) identify five "tribes of fluency" which can impact upon an individual's experience: perceptual fluency, embodied cognitive fluency, linguistic fluency, higher order cognitive fluency and memory-based fluency (see Alter & Oppenheimer, 2009 for a comprehensive overview). With respect to our *retrieval fluency hypothesis* it is the latter of these "tribes" – memory-based fluency (i.e. retrieval fluency) that we are primarily interested in.

As a sub-category of processing fluency, retrieval fluency can be understood as the relative ease with which an individual is able to bring to mind expressions or examples which conform to a specific rule (Alter & Oppenheimer, 2009). Therefore we can surmise that expressions which have *stronger* levels of fluency (more fluent) are more easily retrievable in memory in comparison with expressions that have *weaker* levels of fluency (disfluent).

Perhaps the most notable example of retrieval fluency is provided by Tversky & Kahneman (1973) in their seminal paper detailing the role that the availability heuristic plays on an individual's judgements. Although Tversky & Kahneman don't use the specific term "fluency" in their paper, their work clearly demonstrates the role that retrieval fluency plays on memory (Alter & Oppenheimer, 2009). For example, participants were asked to retrieve words from memory that either began with the letter "K" or had "K" as the third letter in the word (Tversky & Kahneman, 1973). Participants were significantly better at retrieving words beginning with the letter "K" due to the ease of retrieval (*greater retrieval fluency*) experienced in their memory. This led to participants judging words beginning with the letter "K" to be more frequent in comparison to those which had "K" as the third letter. In line with this, research has indicated that fluency can have an effect upon judgements across a wide range of domains (Oppenheimer, 2008). These include judgements on intelligence (Oppenheimer, 2006), truthfulness (McGlone & Tofighbakhsh, 2000; Reber & Schwarz, 1999), likability (Bornstein & Dagostino, 1992; Reber, Winkielman, & Schwarz, 1998; Zajonc, 1968) and famousness (Jacoby, Woloshyn, & Kelley, 1989).

2.2 – Episodic Memory and the Encoding Specificity Principle

Since the early 1970's researchers have made the distinction between episodic and semantic memory (Tulving, 1972, 2002). Unlike semantic memory, which enables us to store our general knowledge (Ashcraft & Radvansky, 2010) episodic memory refers to the ability to learn, store and retrieve information about our own personal experiences (Dickerson & Eichenbaum, 2010). Since its theoretical conception, researchers have focussed on understanding how episodic experiences are stored and processed in memory. For example, early work by (Craik & Lockhart, 1972) focussed on the idea that storage in episodic memory is influenced by *depth of processing*. This theory suggests that information that is processed at a shallow level (receiving only incidental attention) is stored less effectively than information processed at a deeper level. The authors proposed that deeper processing (which involves the elaboration of the representation of information stored in memory) is associated with more detailed, stronger and longer lasting memory traces (Craik & Lockhart, 1972; Ashcraft & Radvansky, 2010).

Other research has focussed on how episodic memory stores specific types of information. For example, Palmeri, Goldinger, and Pisoni (1993) studied the role of episodic memory in voice and speech encoding. Their results suggested that voice information is encoded in memory automatically without conscious or strategic processes. Through episodic

memory, voice information can be stored in robust multidimensional representations that are retained in long-term memory for prolonged periods of time (Palmeri et al., 1993). In later research Goldinger (1996), extended this finding by showing that episodic traces of spoken words can impact upon recognition memory for a day and perceptual identification for up to a week after initial encoding.

Logan, (1988, 1990, 1992, 1997) took a different approach in investigating the function of episodic memory by developing a model outlining how memory can be utilised in the development of expert performance and automaticity in skill acquisition. In Logan's Instance Theory of Automaticity (ITA), episodic memory functions as a learning mechanism. Experience with a task builds separate memory traces that can then be retrieved when the task is repeated (Logan, 1997). Logan argues that task performance becomes automatic when it is based on the memory retrieval of past solutions to a problem. Thus when these solutions become reliable enough, performance can be based entirely on episodic memory retrieval (Logan, 1997).

More recently, work by Yonelinas (1994) has focussed on recognition in episodic memory. The Dual-Process Signal Detection Model (DPSD) differentiates between recollection and familiarity in memory. The model asserts that recollection and familiarity differ in relation to the type of memory information that they provide (Yonelinas, Aly, Wang, & Koen, 2010). Familiarity reflects "quantitative" memory strength and emulates a signal detection process where new items produce a Gaussian distribution of familiarity values. Accordingly, old items are therefore recognised as being more familiar than new items. In contrast to this, recollection is viewed as a threshold retrieval process in which "qualitative" information about a previous event is retrieved (Yonelinas et al., 2010). If *recollective strength* falls below a threshold then recollection will fail to produce any discerning evidence that an item has been encountered previously. When this occurs, individuals will be unable to retrieve information that discriminates between old and new items (Yonelinas et al., 2010).

Building on this past research, our retrieval fluency hypothesis also draws on the importance of retrieval strength in memory. In particular we were influenced by the work of Gann and Barr (2014) who speculated that when determining how to encode a referent, speakers might use the *strength of the memory signal* associated with a particular linguistic expression as an index of its contextual appropriateness. The assumption that memory signals correlate with informational adequacy is derived from the *encoding specificity principle* of episodic memory (Tulving & Thomson, 1973). According to this principle,

events are encoded into a deeper memory representation which includes the context the item was in during initial encoding (Ashcraft & Radvansky, 2010). Thus the strength of a memory signal is a function of the *similarity* between encoding and retrieval contexts (Tulving & Thomson, 1973).

Evidence in support of the encoding specificity principle comes from a range of studies in the memory literature. For example, in Godden and Baddeley's (1975) famous scuba diving study, participants learned a list of words either in water or on land. Half of the participants recalled words in the same context that they had learned the words in, whereas the other half of participants recalled words in the alternative context. Crucially, the authors found that recall was better when participants were in the same context the information was originally encoded in (Godden & Baddeley, 1975). More recently, fMRI research has also provided evidence in support of encoding specificity principle. Vaidya, Zhao, Desmond, & Gabrieli (2002) found that the cortical areas which are initially involved in the perception of a visual experience become part of the long term memory trace for that particular experience, thus suggesting a neural basis for encoding specificity in memory (Vaidya et al., 2002).

In line with this evidence, Gann and Barr (2014) have applied the encoding specificity principle to audience design performance. Accordingly, the fluency with which a speaker's expressions are retrieved should depend upon the degree that the referent and the retrieval context match the original encoding context (see Tulving & Thomson, 1973). Gann & Barr (2014) propose that memory retrieval may influence the speaker's propensity to engage in audience design when producing utterances for the addressee. The authors argue that expressions with a *strong* memory signal would be more likely to be deemed *contextually appropriate* by the speaker, resulting in less consideration of context and less delay in production, relative to expressions yielding weak memory signals.

Gann and Barr's (2014) retrieval fluency proposal is similar to Horton and Gerrig's (2005a) memory-based theory. Crucial to Horton and Gerrig's (2005a) theory, is the idea that memory acts as a *proxy* for common ground. Accordingly, the "effects typically ascribed to conversational common ground are emergent properties of ordinary memory processes acting on ordinary memory representations" (p. 2). Horton and Gerrig (2005a) argue that memories that are frequently associated with a particular cue will become most readily available for the speaker when that cue is presented. Importantly, resonance is influential to the extent that the relevant cue is available within the context – with enough strength to reach threshold (Horton & Gerrig 2005a).

Similarly to the retrieval fluency proposal Horton and Gerrig (2005a) argue that the strength of memory associations can impact upon judgments of common ground in audience design. Thus the overall collection of memories encoded with a particular addressee (as well as the *strength* of these memories) will influence the probability that speakers will be compelled to take on strategic control of both message formation and commonality assessment (Horton & Gerrig, 2016). Accordingly, in the memory-based account, Horton and Gerrig (2005a) argue that when associations between interlocutors and other information are weak commonality assessment will be likely to fail.

2.3 – Instance Theory of Automaticity

Crucially, Gann and Barr's (2014) proposal draws on Logan's (1988) *Instance Theory of Automaticity* (ITA). In the section above we briefly highlighted Logan's key idea. Logan (1988) argues that *automaticity is memory retrieval* – that performance becomes automatic when it is grounded in directly accessed memory retrieval of past solutions. Logan's (1988) theory suggests that individuals start with a general algorithm that adequately completes the task at hand. Individuals gain experience of specific solutions to a problem these are then retrieved when the same problem occurs on a separate occasion. Automization is therefore reflected in the switch from "algorithm-based performance to memory-based performance" (Logan, 1988, p. 493). For example, when an individual is first asked to solve a maths problem - "What is 13 x 21?" they may take a few seconds or so to compute their answer. Following Logan's logic once they have figured out the solution (13 x 21 = 273) they are likely to switch to a memory-based approach and retrieve their previous answer if presented with the same problem again at a later date.

Logan (1988) argues that both encoding and retrieval are connected through attention - thus the same act of attention that produces encoding also produces retrieval. The ITA has three important assumptions: (1) memory encoding is an unavoidable, obligatory consequence of attention, (2) retrieval from memory is also an unavoidable, compulsory consequence of attention and that (3) each time an individual encounters a stimulus their experience is encoded, stored and retrieved separately. Thus following ITA theory, Gann and Barr (2014) suggested that speakers store *episodic representations in memory* involving a referent, a context, and an expression.

The authors outline how the ITA can be applied to audience design - when the speaker first encounters a referent they are likely to adopt a reasoned approach in an attempt to find an adequate description which separates it from alternative referents. In turn, the chosen description then becomes linked to the cognitive antecedent conditions which represented

the original referential process. Accordingly this "processing episode" is then stored in the speaker's memory (Gann & Barr, 2014). Subsequently, when the same antecedent conditions appear again this will prompt the obligatory retrieval of the previous description.

Logan's theory argues that when an individual is attempting to complete a goal within the same context as they were previously, they can choose how to respond. They can do this either by opting to recall information from memory or they can run off an algorithm which computes a response to the task at hand. Logan (1988) views this choice as a "race" between memory and the algorithm and suggests that eventually memory will always dominate the algorithm, as over time more and more memory instances will join the race. This framework also suggests that each stored episode in memory races against other encoded episodes. Accordingly, the interlocutor can respond using their memory immediately after the first episode is retrieved (Logan 1988).

It is predicted that in a communicative environment, the greater the similarity between the original context and the current setting - the more likely the speaker will re-use their previous description (Gann & Barr, 2014; Tulving & Thomson, 1973). As mentioned previously, this theory proposes that speakers' will utilise *the strength of the memory signal* associated with a particular context as a way of determining how much consideration they need to apply when planning their description. Gann & Barr (2014) suggest that the strength of the memory signal obtained - otherwise known as *retrieval fluency*, acts as a heuristic for audience design. When speakers experience a strong signal (*greater retrieval fluency*) it indicates that their previous description is likely to be contextually adequate – resulting in less effort being allocated to utterance planning. Thus the strength of the memory signal helps to gauge the need for further planning before the interlocutor begins to speak. When the signal is *highly fluent* in memory the speaker is likely to begin their description before they fully engage in audience design (Gann & Barr, 2014). Thus when interlocutors experience *greater levels of retrieval fluency* they will be more likely to provide descriptions that may appear to be egocentric in nature. When the memory signal is weaker (*less fluency in memory*) speakers' will give more consideration to their utterance and engage more fully in audience design before beginning their description.

Importantly, the more effort the speaker allocates to a referential description, the more likely they are to monitor the current context and check that their description is sufficient. Conversely, less checking of the current conversational context would mean that the

speaker is more reliant on the *strength of the retrieval fluency signal* they experience which may in turn lead to more descriptive errors in audience design.

2.4 – Retrieval Fluency as a Heuristic

Our retrieval fluency hypothesis seeks to build upon Gann and Barr's (2014) theory. We follow the suggestion that the *strength of the memory cue* plays an important role in audience design performance and propose that the algorithmic vs. memory retrieval route (Logan, 1988) need not be considered as a "race". Our hypothesis suggests that if the memory signal associated with a particular expression crosses a threshold then this will be likely to cue the previous description used in that context. Thus rather than fully engaging in audience design (using common ground to tailor descriptions to the listener's specific needs) speakers will be likely to re-use previously established descriptions formed with the addressee. Crucially, this will pre-empt a "race" between memory and the algorithmic route and prevent a thorough search of common ground for a contextually relevant descriptive term.

We argue that the retrieval fluency heuristic is likely to be used as part of a default process that is largely performed on an unconscious level by the speaker. However, we note that whilst fluency can be used routinely as a useful heuristic, it is not an obligatory process. Occasionally, speakers may opt to consciously override the fluency effects that they experience and engage more fully in the process of audience design. Since the likelihood of using retrieval fluency as a heuristic is influenced by the *strength of the memory cue* available, the retrieval fluency hypothesis reflects an individual's propensity to engage in the audience design process. Thus when the speaker experiences a weaker memory signal they will be less likely to use the retrieval fluency heuristic as a substitute for audience design.

When the memory signal is weaker or alternatively when the speaker is confronted with a scenario in which no previous description comes to mind, we would expect participants to provide generic-listener adaptations for the listener. As outlined by Dell and Brown (1991), these adaptations are designed to benefit comprehension for a generic listener and are formed by consulting a model of the generic listener in the language community (Barr & Keysar, 2006). Should speakers have to rely on this approach we would expect them to engage in a form of monitoring and adjustment (Horton & Keysar, 1996) in an attempt to ensure that they provide an adequate description to the addressee. Only in circumstances where monitoring and adjustment fails to produce an adequate description would we then expect speakers to engage in full audience design by using their knowledge of their

common ground with the addressee to design a suitable utterance. Consistent with previous models (Horton & Keysar, 1996; Keysar, Barr, Balin, et al., 1998) we argue that in these instances common ground is likely to function as a correction mechanism in language production.

We note that our retrieval fluency hypothesis could be consistent with the Interactive Alignment Model proposed by Pickering and Garrod (2004). As outlined in Chapter 1, Pickering and Garrod's theory argues that over time interlocutors align situation models during dialogue. This alignment is the result of communicators producing and interpreting expressions in a similar fashion to their conversational partner (Pickering & Garrod, 2006). We believe that our retrieval fluency hypothesis could help to explain how alignment is facilitated. Through conversation speakers entrain on particular descriptions of objects. As these descriptions are re-used speakers form stronger memory traces for these utterances, resulting in *greater levels of retrieval fluency*, which makes them more likely to be recalled during later interactions. This idea is consistent with work by Knutsen & Le Bigot (2012) who argue that reference re-use depends upon accessibility in memory, with more accessible references being more likely to be used again. This greater re-use of descriptions (due to a stronger memory signal) may help to facilitate alignment by increasing the likelihood that interlocutors will become more familiar with each other's utterances.

In summary, accounting for the background literature reviewed above, our retrieval fluency hypothesis has two interesting theoretical components: (1) that speakers store *"referring episodes"* that link together referents, contexts, and expressions; and (2) that speakers make use of the *strength* with which referents and contexts *cue retrieval of expressions* as one index of the extent to which such expressions are contextually appropriate. In the section below I detail each of our three experiments and our attempt to investigate the retrieval fluency hypothesis.

2.5 – Experiment Overview: Testing the Retrieval Fluency Hypothesis

The work contained in this doctoral thesis is intended as a direct follow-up to Gann and Barr's (2014) study and serves to further test the memory-based model first put forward by Horton and Gerrig (2005a). The three experiments outlined in the following chapters document our attempt to test the *retrieval fluency hypothesis*: that speakers use retrieval fluency as a heuristic for audience design in referential communication. A key feature of both Experiments 1 and 2 was the manipulation of a communicatively irrelevant aspect of the context that stimuli items appeared in. In this way, we de-confounded memory from common ground use. This enabled us to test whether visual features in the environment

acted as a cue for memory during audience design performance. Importantly, Horton and Gerrig (2005a) highlight that resonance is a key part of the retrieval process in their memory-based model. It produces a parallel search of memory which enables a wide range of associated information to become accessible to the interlocutor (Horton, 2008). Therefore if the memory-based model is correct, altering the visual context should have an impact on the way in which participants access encoded information – to the extent that a more similar context should produce successful retrieval of previous descriptions for the listener. Our first two experiments tested this assumption.

In *Chapter 3*, I outline the methodology and rationale behind Experiment 1. In this study, we presented participants with a grid containing letters of various sizes and colours. Participants played the role of "Director" and were tasked with describing a highlighted target letter to the "Matcher" confederate. Crucially, we manipulated the appearance of the "competitor" and "foil" items which alternated between training and test trials in such a way that participants would have to adapt their descriptions at the test phase in order to avoid misspecifying descriptions. For example, participants were shown a target letter "A" during training but were presented with two contrasting letters during the test phase – "A" vs. "a". In this instance they would have to modify their description (e.g. "the *big* A") in order to provide an adequate description to the addressee. We expected participants to experience greater retrieval fluency when the test trial configuration was highly similar to the training trial configuration, leading them to continue to use the same description and therefore make more descriptive errors than when presented with configurations that were dissimilar between the training and test phase. The results of this study failed to significantly support the retrieval fluency hypothesis. However, there was some suggestion of a potential effect of fluency on audience design, which prompted the motivation for our second experiment.

Chapter 4 details Experiment 2. In this study we made some minor adjustments to the configuration and sequencing of objects and altered the stimuli presented to participants. In this experiment our results offered weak statistical support for the retrieval fluency hypothesis for audience design and indicated that participants relied on the strength of the memory signal present when constructing descriptions for the listener. However, the effect we detected was small and merited further investigation. Thus we opted to carry out one additional experiment which aimed to test the retrieval fluency hypothesis in a more communicatively relevant setting.

Chapter 5 outlines Experiment 3, which enabled us to apply our theory to practice. In this study, we had participants describe target items to one of two Matchers (both confederates) using an interactive webcam design. This enabled us to further test the concept of partner specificity (Horton, 2007; Horton & Gerrig, 2005a) whilst also assessing audience design using a task which de-confounded the effects of memory from the effects of common ground. At the test phase in this task the visual experience of the Director (participant) was controlled independently of the pragmatic situation, so that who the Director saw and who the Director was speaking to did not always coincide. Our design was fully interactive with participants developing their own descriptions for target objects with one of the two Matchers during the training phase. This set-up enabled us to test whether the speaker used the conversational partner they spoke to during training as a memory cue when providing descriptions at the test phase (e.g. Horton, 2007; Horton & Gerrig, 2005a). Crucially, this study allowed us address concerns raised by Brown-Schmidt (2009) and Brown-Schmidt et al. (2015) regarding a lack of live interaction in previous experiments which failed to find evidence in support of partner specificity (e.g. Barr & Keysar, 2002; Kronmüller & Barr, 2007). Our results in this experiment were in the opposite direction predicted and failed to support the retrieval fluency hypothesis. These findings have important implications for the retrieval fluency hypothesis and challenge the key assumptions of Horton and Gerrig's (2005a) memory-based model for common ground and audience design.

Finally*, Chapter 6* provides a summary of the key findings from all three experiments, final remarks and an outline of future directions for the study of audience design in referential communication.

Chapter 3 – Experiment 1

3.1 – Background

3.1.1 – Audience Design in Language Production

Chapter 1 outlined the idea that speakers choose their descriptions based on information that is more readily accessible to themselves, rather than their addressee. This is supported by research showing egocentric tendencies in speech production. Egocentrism in language is often demonstrated through misspecified descriptions, providing more or less information than the listener needs - with speakers more likely to overspecify than underspecify utterances for listeners (Deutsch & Pechman, 1982; Ferreira et al., 2005). As noted previously, Engelhardt et al. (2006) found that participants provided unnecessary, overspecified descriptions to a confederate in almost one third of trials. The authors note that speakers will overspecify when their expression encapsulates the relevant situation from their perspective. This means that they will fail to engage in audience design and will not attempt the process of adjusting their description to make it suitable for the listener. Further evidence shows that speakers are often unable to prevent themselves from providing addressees with privileged information when delivering referential descriptions - even when it results in a loss of points during an experimental game (Wardlow Lane et al., 2006). These findings indicate that speakers' failure to take into consideration their own unique perspective when providing descriptions is caused by autonomous, low-level processes which result in privileged knowledge becoming unintentionally incorporated into utterances (Wardlow Lane et al., 2006).

Wardlow Lane and Liersch (2012) replicated this finding and showed that even when speakers are offered a monetary reward for concealing privileged information from addressees, they were unable to do so. Similarly to Deutsch and Pechmann (1982), the authors argue that overspecification may have communicative benefits - by reducing privileged information and increasing common ground between interlocutors. However, they also note that overspecified descriptions can also lead to referential errors and confusion for the listener, a conclusion which is supported by Engelhardt et al. (2006) and Sedivy et al. (1999). In addition to this, Engelhardt, Demiral and Ferreira's (2011) found evidence that reaction times were significantly longer when addressees heard descriptions that contained overspecifications. ERPs indicated a centroparietal negativity (N400) that appeared 200-300ms after modifier onset suggesting that unnecessary pre-nominal modifying expressions had a negative effect on listeners' comprehension. Nevertheless,

this remains a contentious issue with some evidence indicating that overspecification does have communicative benefits (e.g. Nadig & Sedivy, 2002; Paraboni, Masthoff, & van Deemter, 2006; Sonnenschein, 1984; Sonnenschein & Whitehurst, 1982).

In a series of experiments, Wardlow Lane and Ferreira (2008) further tested the effect of privileged information on speaker descriptions. In their study, two naive participants played as the speaker and the addressee in a referential communication task. The participants were shown sets of four line drawings that consisted of pairs of objects that differed only in size (e.g. big vs. small triangle) and single objects that did not have a partner. In this study, both participants could see three of the items in the set. The speaker was instructed to occlude the fourth item, thus creating an object that was in their privileged ground. Participants were then presented with *contrasting* trials (where the target object was the same type of item as the privileged object) and *non-contrasting* trials (where the target was distinctive and did not form part of a pair). The authors measured the percentage of trials where participants used size-modifying descriptions in the contrasting vs. non-contrasting conditions.

The results were consistent with Wardlow Lane et al. (2006) and Wardlow Lane and Liersch (2012) – when privileged information was more salient for the speaker they found it harder to avoid using that information in their descriptions (Wardlow Lane & Ferreira, 2008). This effect further demonstrates egocentrism in language production as participants continued to use privileged information in their utterances even though it hindered their attempts to provide a referentially successful description. The finding that speakers use descriptions which are not optimal for addressees' understanding is further supported by research suggesting that speakers will often fail to include optional words in descriptions which would have helped to prevent temporary ambiguity for the addressee (Ferreira & Dell, 2000). This finding emphasises a tendency for speakers to adopt descriptive terminology which suits their own conversational needs rather than the needs of the listener.

More recently, Gann and Barr (2014) assessed audience design performance in partner adaptation. In this paper the authors viewed successful speech adaptation as a type of expert performance "in which skilled behaviour is the result of an interplay between memory and attention" (p. 744). Gann and Barr (2014) had participants play the role of speaker in a referential communication game. In this game, half of the participants played with one additional participant who was the addressee and the other half played with two extra participants who took turns at playing as the listener. Participants saw five pictures

that were shown at the corners of an imaginary pentagon. The speaker was privately informed which object was the target item to be described and was instructed to describe the object to the listener without providing details of its location on the computer monitor. Participants played through a series of "training" and "test" trial blocks. The training phase enabled participants to develop experience in describing each of the target items. In the test block speakers referred to these items again in addition to some new objects. Speakers referred to two types of target item: *conventional* (normal everyday objects) and *unconventional* items (abstract figures). These items provided speakers with opportunities to underspecify and overspecify target descriptions.

During the training phase conventional items (e.g. candle) always appeared alongside a less prototypical version of the object (e.g. unmelted candle). Crucially, speakers were required to provide a description that distinguished between these two items. At the test phase the competitor item was not included in the display – meaning that using a previously modified description (e.g. "the unmelted candle") would result in an overspecified utterance. Gann and Barr (2014) were interested in whether participants adapted their speech for new addressees during the test phase. The crucial question was: would speakers continue to use abbreviated descriptions formulated with the previous listener or would they adapt their utterances to suit the current listener? Additionally, the authors were interested in whether the overspecification rate would differ depending on the identity of the conversational partner.

Results indicated that participants were much more likely to overspecify than underspecify referents. When describing unconventional items speakers successfully shortened descriptions but were also able to adapt these utterances for new addressees who were unfamiliar with the target object (Gann & Barr, 2014). Notably, when providing descriptions speakers relied on feedback from the addressee when it was permitted and relied on their own judgements when feedback was unavailable. Gann and Barr (2014) found that speakers overspecified old objects at similar rates for both old and new addressees. In line with Engelhardt et al. (2011) the authors found that addressees experienced more difficulty understanding overspecified descriptions in comparison to adequately described utterances (Gann & Barr, 2014). Together, these findings give us an insight into the difficulties speakers experience while attempting to engage in audience design. As the evidence suggests, speakers often fail to abide by Grice's (1975) *Cooperative Principle* by providing overspecified, and potentially confusing, descriptions to the listener.

3.1.2 – Configuration of the Retrieval Fluency Experiment

Whilst we have a general understanding of the processes involved in successful audience design, our knowledge is far from being complete. This is partially due to the fact that previous research has treated representational and processing issues separately (Gann & Barr, 2014). Thus memory-based models (e.g. Horton & Gerrig 2005a) largely focus on representational issues but lack consideration of how these representations are deployed, whereas Monitoring/Perspective Adjustment Models (e.g. Horton & Keysar, 1996; Keysar et al., 1998) emphasise the importance of processing issues whilst assuming the existence of suitably structured representations (Gann & Barr, 2014). Our focus on the impact of *retrieval fluency* on audience design addresses this issue by considering how expressions are structured in memory whilst also addressing the issue of how speakers may process these stored expressions to mediate their descriptions to the listener.

In our first experiment, we attempted to test the *retrieval fluency* hypothesis by manipulating the level of fluency that the "Director" (participant) experienced whilst providing descriptions to the "Matcher" (experimenter). In this study each participant played in an interactive communication game - the participant and experimenter both faced away from one another and looked at separate computer screens. Each screen showed a grid containing various letters of varying colours and font sizes (see Figure 2 and Figure 3 for examples). As in previous studies (Brennan & Clark, 1996; Gann & Barr, 2014) the Directors were entrained on particular ways of describing referents and then were presented with a test display in which the context had changed so that the entrained-upon description would no longer be appropriate.

In our study, the objects being discussed were not everyday objects, but rather letters of the alphabet of varying colour and font size that were embedded in a display of other letters. Speakers entrained on descriptions either requiring a bare noun (*"the u"*) or a noun phrase with a size modifier (*"the little u"*, to distinguish it from a larger U in the display). In the test trial, the context changed in a way that invalidated the entrained-upon description (for example, the "large U" disappeared during the test phase, rendering the description *"the little u"* inadequate). Our main question was whether speakers would adapt their descriptions, and whether the likelihood of this adaptation depended upon the *fluency* with which context cued the entrained-upon description.

We attempted to alter the fluency with which Directors retrieved descriptions for a particular target object by manipulating how much the context varies each time the description was used. The key idea was that Directors who entrained on a description

within a *highly variable context* would experience *less fluent retrieval* of that description than Directors who entrained on that same description within a *low variability context*. We attempted to do this by altering the *Context Variability* of the grid that our target items appeared in. This enabled us to test whether people were better at tailoring their descriptions to a listener's informational needs when *retrieval fluency processing levels were low* compared to when retrieval fluency levels were *high*.

In addition to this, we also incorporated a *Shift Direction* factor in our study. This factor was included to vary the amount of information that Directors would have to provide in test trials relative to training. Thus in some test trials, participants had to provide more information to the Matcher and in others, less. This variation was intended to prevent a situation in which Directors would learn that they need to alter the information at test in only one direction (e.g., always increase rather than reduce information):

Context Variability Factor

The trials were presented in two blocked sequences (with the order counterbalanced across participants): a "Low Context Variability" level and a "High Context Variability" level. The Low Context Variability trials contained filler letters within the grid which were arranged relatively consistently with the previous trials presented. In this level only two or three letters were varied at random and they were only moved to one adjacent square on the grid (see Figure 3). As the context was very similar to previous trials, it was expected that participants would experience *greater retrieval fluency* at this level. In the High Context Variability level, the filler letters within the grid were arranged inconsistently – appearing in completely random locations which ensured that they were relatively dissimilar to previous trials. It was expected that participants should experience *weaker retrieval fluency* at this level.

Note that we opted to manipulate the position and colour of the filler letters in each display. These features were deemed to be salient to the Director (and thus impact retrieval) but were *communicatively irrelevant*. In particular, these features were chosen because they would not affect the description of the target item - speakers were made aware that the position of letters in each grid were set out in a different arrangement for the Matcher than the arrangement they saw. Additionally, the colour of the filler letters was not relevant to the descriptions of the target item (see section *4.2.4 – Materials* for further details of the configuration of the stimuli included in each display). As highlighted in Chapter 2, this manipulation enabled us to de-confound potential effects of memory from

common ground and test whether visual configuration is an influential cue that affects memory in language production.

Shift Direction Factor

In each display, the target appeared with a "critical" letter, whose identity formed the second factor of *Shift Direction*. This factor refers to whether speakers entrained upon unmodified descriptions ("the *u*") and were tested in a context requiring a size modifier ("the *small u*") or vice versa. In the former level (Singleton-Contrast level; see Figure 3 for example), the critical letter during training was a letter of the same colour but different identity from the target (e.g., if the target was a yellow "*u"*, the critical letter might be a yellow "p"), leading Directors to entrain upon a bare noun phrase ("*the u*"). We refer to this non-competitor letter 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 was clearly not the same letter (see Figure 3 for an example of the stimuli). For the test trial in this level, the foil letter was changed to have the same identity as the target but was of a different size (e.g., "a *small u*" vs. "a large U"), thus requiring the introduction of the modifier "*small*". In the Contrast-Singleton level this order was reversed: the critical object during the training trials was the competitor (see Figure 3). The competitor had the same identity as the target letter but contrasted in size during training (e.g., "a *small u*" vs. "a large U"), leading speakers to entrain upon a size-modified expression. This competitor was then replaced with the foil at the test phase, meaning that that the Director was no longer required to include a size modifier in their description.

If participants follow Grice's (1975) Cooperative Principle then we would expect Directors to adapt their description to suit the Matcher's referential needs. Previous research has informed us that interlocutors are more likely to overspecify than underspecify their descriptions (Deutsch & Pechmann, 1982; Ferreira et al., 2005) and so the introduction of both Singleton-Contrast and Contrast-Singleton levels allow us to test for this effect. During the test phase it was predicted that there would higher misspecification in the Contrast-Singleton level compared to the Singleton-Contrast level. Thus in line with previous research, it was expected that the rate of overspecification (in the Contrast-Singleton level) would be greater than the rate of underspecification (in the Singleton-Contrast level).

3.1.3 – Pilot Study and Pre-registered Predictions

The basis for our predictions was a pilot study containing 22 participants (with 24 sequences per participant, whereas our main study contained 48 sequences). This pilot study is available on the github site for the experiment (https://github.com/dalejbarr/EESP2) as well as in our files on the Open Science Framework (OSF: https://osf.io/4akir/). We pre-registered all our predictions on the OSF (outlined in section 3.3.4). Our main prediction was that speakers would be more likely to misspecify referents in the Low Context Variability level than in the High Context Variability level; in other words, we predicted a main effect of *Context Variability*.

3.2 – Method

3.2.1 – Participants

In total 36 subjects completed the experiment (24 Females, M=24.1 years). All subjects were recruited from the campus at the University of Glasgow. Participants were paid £6 or received 4 "participation credits" (course credits) for taking part in the study. Eleven participants in total had to be replaced. Ten were replaced due to the use of ineffective descriptions during the task (continuously failing to adapt their utterances for the listener, please see Section 3.3.3 – *Exclusion Criteria for Participant Responses* for more details). One additional participant was replaced due to the use of excessively long descriptions on each trial. Subjects gave written informed consent before beginning the experiment and were fully debriefed after the experiment had finished. Our procedures fully complied with the ethical code of conduct of the British Psychological Association.

3.2.2 – Experimental Setup and Task

The experiment was 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 sat in different areas of the testing room and looked at separate computer monitors throughout the experiment. Both were seated facing in opposite directions so that they were unable to see each other's display (please see Figure 2 for an example of the set-up). In each trial, the Director was asked to describe a highlighted target letter, which appeared on their monitor, to the Matcher. The Matcher then identified this letter on his own screen and selected it using a computer mouse. The target letter appeared on the Director's screen within a grid among other 'filler' letters (see Figure 2 and Figure 3). The Director was informed that in each trial the listener would have the same letters on their

monitor but that they may be arranged in a different format compared to the grid that appeared on their screen.

3.2.3 – Design

There were two factors in the design, Context Variability (Low and High) and Direction of Shift (Singleton-Contrast and Contrast-Singleton), forming a full-factorial 2x2 within-participant design.

3.2.4 – Materials

The parameters governing each display in the experiment are defined in the sqlite3 database EESP2.db in the github repository (https://github.com/dalejbarr/EESP2).

Each display consisted of a five-by-four grid containing uppercase letters (A-Z) of different font size and colour (see Figure 3 for examples). All letters appeared in Arial font. The font sizes were randomly generated for each trial and we describe them as either 'small' (font size varying 64 - 96pts) or 'large' (font size always 32pts higher than the smaller letter in a pair, maximum size was 128pts).

The experiment contained 48 "sequences" of trials, each consisting of a number of training trials followed by a single test trial. The term "sequence" is used to refer to the collection of training and test trials all associated with a single target/competitor pair. Twenty-four sequences appeared in the Low Variability Context level, and the remaining 24 in the High Variability Context level. Each participant was given a unique set of randomly generated displays; in other words, displays did not repeat across participants (thus obviating a by-items analysis). For each training sequence, the number of trials was randomly selected, with a range from 6 to 9. The motivation for varying training sequence length was to make the occurrence of the test trial unpredictable. Given these parameters, each experimental session could have contained between 336 (7 x 48) and 480 (10 x 48) trials.

The sequences for each of the 36 sessions were randomly generated in advance. Each sequence for each session was based on a randomly generated original "prototype" display, which was used as the test trial. The training trials were all distortions of this prototype. Each sequence had a target letter whose identity, colour, and size were fixed across all displays. The identity of the target letter for each sequence was chosen randomly, with the constraint that the same letter could not be used as target more than once within each block of 24 sequences formed by the Context Variability factor. After the selection of the target for a given sequence, a "foil" letter which acted as a competitor was selected from the remaining set of letters, with the probability of selection inversely proportional to its

similarity to the target, as derived by norms given in (Simpson, Mousikou, Montoya, & Defior, 2013).

By biasing the selection toward visually similar letters, we attempted to increase the likelihood that Directors would fail to detect the difference between a letter with the same identity (e.g., target="O", competitor="Q"). The random selection process also meant that each participant would get mostly distinct letter pairs, which allows us to treat items as a fixed effect in our analyses (Clark, 1973). The pairings for each session are stored in the table `LetterPairs` table within the `EESP2.db` database (available on github).

The target/competitor letters always appeared in the same colour and position across all training and test displays. In addition to these two letters, there were three sets of "distractor" letters scattered among the other squares in the grid. The distractor letters were randomly chosen from the set of letters excluding the target and competitor. Each set in each sequence had letters of a different colour, each randomly chosen (without replacement) from a palette of ten colours. The first set was of the same colour as the target and competitor, and had either four or five letters. The second set was of a different colour and also had either four or five letters. The third set was also of a different colour and had one or two letters. The sizes of the "distractor" letters that appeared within the grid were randomly generated (between 64-128pts). The information used to generate each prototype and sequence is stored in the table `SeriesInfo` in `EESP2.db`.

Next, the letters for each prototype were assigned positions within the display. The assignment of the target and competitor positions was random, with the constraint that they must be at least four spaces apart (using a city-block metric). The positions of the distractor letters were assigned randomly. The prototypes are contained in the table `Prototypes` in `EESP2.db`.

The training trials were created for each sequence by distorting the prototype, with the number of distortions randomly selected from a uniform distribution of integers from six to nine. In the Low Context Variability level, the distortion was created by randomly selecting two to three distractor letters, and moving them in the grid to an adjacent empty space. Any letter that was "locked in" (i.e., all surrounding spaces occupied) was never selected to move.

In the High Context Variability level, the positions of all of the distractor letters were randomly reassigned. Also, in this level the colours of two of the distractor sets could be swapped. There is an entry for each created display in the table `SessionGrids`, with the corresponding parameters for generating each display in the table Grids. These parameters

were used by a script written in the programming language PHP (`imgcreate.php`) to generate the actual image files that were displayed to participants. The matchers' grids were created simply by randomizing the positions of the letters in the director's grids. Thus, while the locations of the target/competitor were fixed within each series for the director, they varied from trial to trial for the matcher.

Finally, we wanted to make it more difficult for directors to identify the competitor letter using peripheral vision. To this end, we added a slight Gaussian blur to the directors' images using the `convert` command within the ImageMagick suite of command-line tools (version 8:6.7.7.1, www.imagemagick.org), with the sigma parameter set to 8 and radius set to 0 (0x8).

3.2.5 – Apparatus

The experimental stimuli were presented on a 19" LCD Dell desktop computer monitor (4:3 aspect ratio, resolution 1024 x 768 pixels). Participants were seated 45-55cm away from the monitor. A microphone was placed above the participant's computer monitor to record their descriptions of the *target* letter for each trial. The audio was tagged using Audacity 2.0.6 software. Eye movements were recorded during each trial using an Eyelink 1000 (SR Research) eye tracker (sampling rate 500Hz).

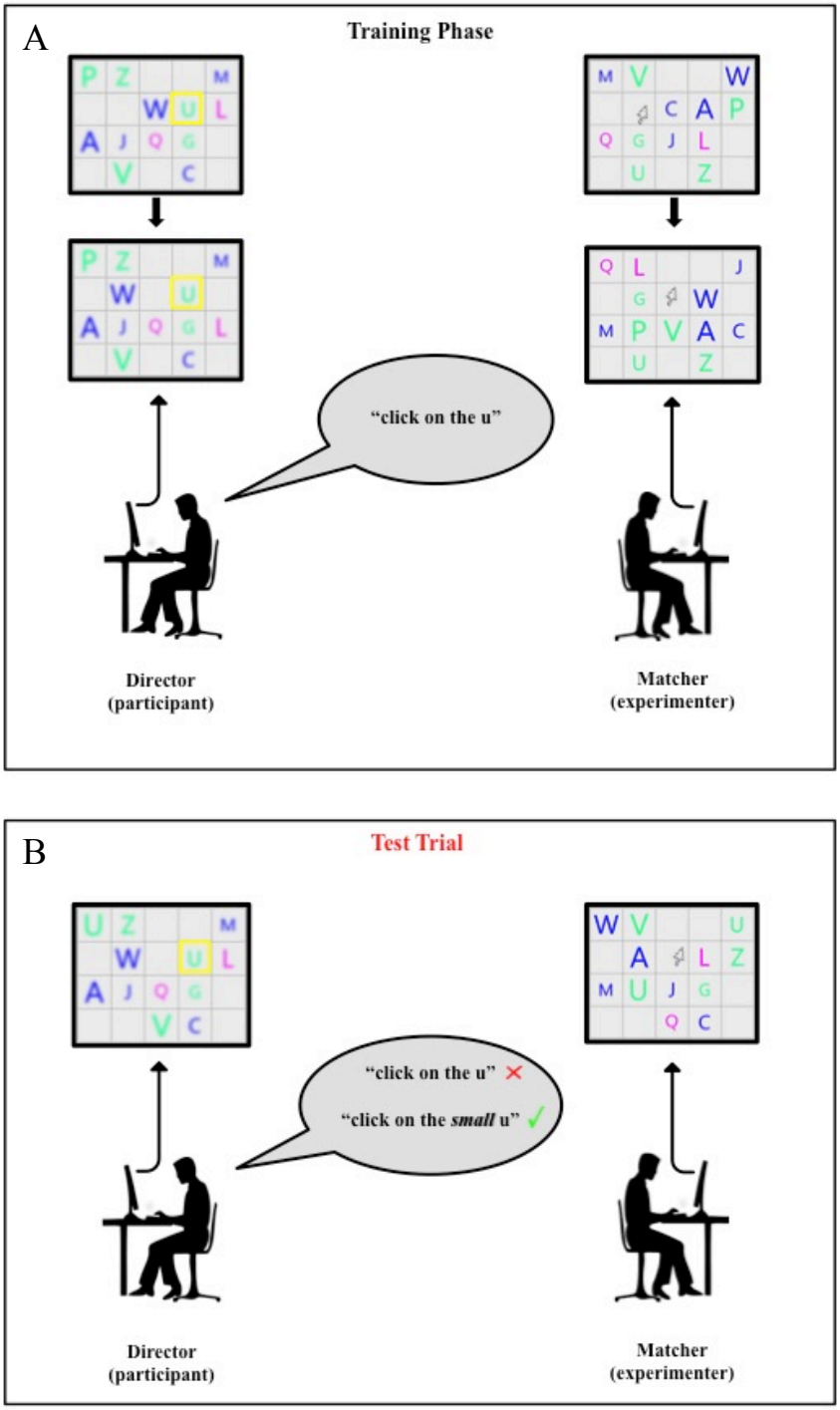


Figure 2: Outline of experimental set up and procedure for Experiment 1.

Panel A shows the Director (participant) and Matcher (experimenter) during the training phase (6-9 trials). Each grid of letters presented on the left-hand side shows an example of a training trial from the Director's perspective. Each of the grids shown on the right-hand side show the corresponding trials viewed by the Matcher. Both the Director and Matcher face in opposite directions looking at separate computer monitors. Once the Director has provided a description (*"click on the u"*) the Matcher will select the appropriate target letter on their screen in order to move onto the next trial. Panel B shows the test phase of the sequence - a competitor letter appears (the *large* U) and the Director is required to adapt their description to the Matcher. Stimuli are shown at the Low Context Variability and Singleton-Contrast levels.

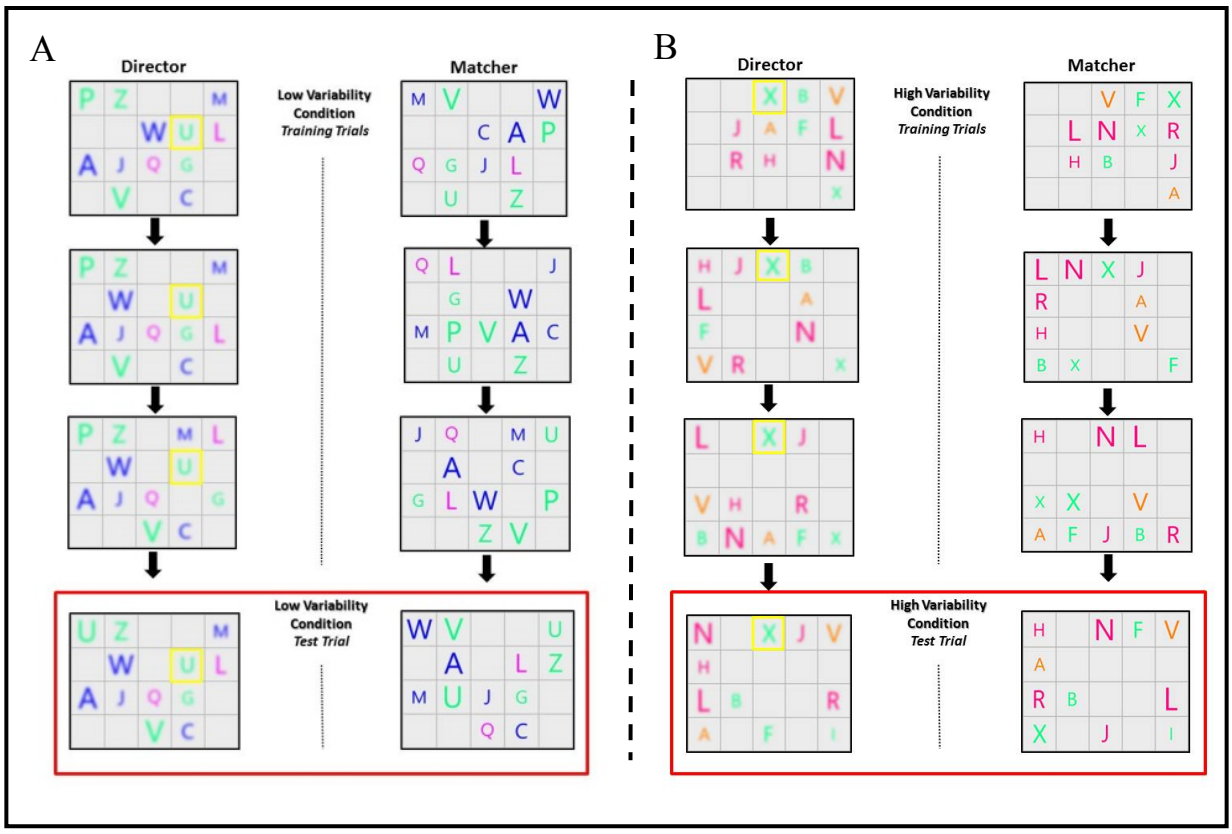


Figure 3: Overview of trials in both Context Variability and Shift Direction Factors.

Panel A shows an example of 3 training trials followed by a single test trial in the *Singleton-Contrast* level. On the left hand side we can see view of the Director (participant). The right hand side shows the view of the Matcher (experimenter). During the training trials the Director is shown each grid with the target letter highlighted in a yellow rectangle – in this case the letter "***u***". The Director is presented with 6-9 training trials before being shown the test trial. The test trial (bottom grid on Panel A) presents participants with the target letter "***u***" again but also introduces a *new* larger "***U***" letter. This test trial may prompt the Director to *underspecify* their description of the target letter to the Matcher - '*click on the* ***u***' whereas the description "*click on the* ***small u***" would be more appropriate in this instance. These trials show stimuli in the *Low Context Variability* level - only 2-3 filler letters on each grid are varied. Panel B shows a similar set up in the *Contrast-Singleton* level. The target letter in this case is the letter "***X***". Note that there is also a contrasting letter present in the grid – the "***small x***". The test trial presents participants with the target letter "***X***" again, but unlike the training trials the "***small x***" is no longer present. In the test trial the Director may *overspecify* their description of the target letter to the Matcher – "*click on the* ***big X***". The description "click on the ***X***" would be more appropriate in this instance. These trials show stimuli in the *High Context Variability* level – filler letters arranged in a completely random order. Note that although the letters are arranged differently for the Director and Matcher, the same letters appear on both grids in each trial.

3.2.6 – 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.

3.2.7 – Procedure

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

In each trial the Director was asked to verbally describe a *target* letter so that the Matcher could identify the item on their monitor and select it using a mouse. In order to discriminate the *target letter* from the filler letters, the *target* was highlighted within a yellow square in the Director's display (see Figure 2 and Figure 3). As the Matcher's

display was not identical to that of the Director the speaker had to describe the features of the highlighted letter, rather than use the *target*'s grid location as a description.

At the start of any given trial, an empty grid appeared on the Director's screen, with a yellow square marking the location for where the target would appear. After one second, the preview screen was replaced with the main display. Audio recording of the Director's response began simultaneously with the presentation of the main display. The trial ended when the Matcher selected the object designated by the Director. The Director could not see the Matcher's screen or mouse pointer, and received no feedback regarding whether the trial was completed correctly. If the Director failed to provide sufficient information to identify the target, the Matcher asked the director for clarification *("Which one do you mean?"*). Any such clarification exchanges appeared in the audio recording for the trial and were noted during later transcription.

Each block of trials (alternating between Low Variability Context and High Variability Context) contained both training and test phases. The training phase consisted of 6 – 9 trials where the *target* letter used in the test phase, appeared 4 – 5 times. The test phase comprised of a single trial. The order of the test trials was randomly generated by a computer script at the beginning of the experiment. Of the 48 test trials shown, 24 featured in the Low Variability Context (12 in Singleton-Contrast level, 12 in Contrast-Singleton level) and 24 featured in the High Variability Context (12 in Singleton-Contrast level, 12 in Contrast-Singleton level).

3.3 – Predictions and Data Analysis

3.3.1 – Main Measurements

Our analysis focussed on three categories of measurements: (1) speech content; in particular, use of a size modifier (big/small) 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.

3.3.2 – Transcription and Coding of Audio Files

For each of the 48 sequences for each Director, we transcribed and coded the audio recordings for two trials: (1) the last trial of the training sequence; and (2) the test trial. The last training trial was needed in order to provide baseline data for the speech onset latency in the test trial. Each trial was transcribed and coded for fluency and adjective use. Fluency was coded into one of four categories, as shown in the Table 1 below:

Speech Code	Description	Example(s)
FL	Fluent speech	"the small Z", "the Z", "Z"
UP	Unfilled pause (occurring after speech onset)	"the... big Z"
FP	Filled pause (um/uh)	"um... big Z"
RE	Repaired utterance	"Z... yeah Z", "Z... uh small Z"

Table 1: Outline of speech fluency categories

Furthermore, we coded whether or not a size modifier was used by the speaker, defined by the following categories:

Modifier Code	Description	Example(s)
NO	No size modifier	"Z", "the Z", "the red Z"
PR	Pre-nominal modifier	"small Z", "large Z"
PO	Post-nominal modifier	"Z that is small", "Z, big"
DE	Deleted adjective	"sm-- uh just the Z"
AS	Addition due to self-repair	"Z... Big Z"
AO	Addition due to other-repair	"Z..." [Matcher: "Which one?"] "Oh…the larger one"

Table 2: Outline of size modifier categories

Onset times of utterances were identified and entered into a data table in milliseconds (ms). The following criteria were applied when identifying utterance onsets:

1. Trials were discarded if the speech was unidentifiable.
2. Any filled pauses or articles were ignored (um, uh, the); speech onset was 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 blue Z"* onset was taken to be at the onset of the word "*blue*").
3. If Directors corrected themselves after an error (e.g. *"pink Z...eh sorry red Z"*) onset of the correction (i.e. *"blue"*) was recorded. However, such repaired utterances were not used in the analysis of speech onset.

3.3.3 – Exclusion Criteria for Participant Responses

One concern was that some Directors may have opted for a "lazy" strategy of always using a size modifier regardless of whether or not there was another letter of the same identity in the display. Indeed 3 of our 22 pilot participants did this. The problem with this behaviour is that on test trials in the Singleton-Contrast level, Directors could simply continue using the modified description, which would then spuriously appear to be appropriately specified. We identified these participants by coding whether or not they inappropriately used size modifiers in the final training trials for each sequence in the Singleton-Contrast level. We removed all data from speakers who did this on more than half of these trials and

replaced these participants. A list of the subjects removed (and their percentages of inappropriately used modifiers) is provided in Appendix 2.

For all remaining participants, we also excluded on a trial-by-trial basis any test trials in the Singleton-Contrast level where speakers used a size modifier on the last training trial. In the Contrast-Singleton level, this was less of an issue because speakers *must* use size modifiers during training or the addressee will be unable to resolve the reference; however when speakers repaired an utterance (for example "*the U... uh the small U*") in the last training trial for this level, we discarded the following test trial.

3.3.4 – Pre-registered Analysis and Predictions

Our pre-registration document specified that we would fit a generalized linear mixed model with maximal random effects, including a logit link and assumption of binomially distributed error variance, using the "bobyqa" optimizer. From our pilot data, we estimated the conditional odds of overspecification as being 1.763 times higher in the Low Variability level ($z = 1.436$, two-tailed $p = .151$). As power is so much lower for binary data than for continuous data, we *pre-registered a one-tailed* and not a two-tailed test for misspecification rate in the main study, with the alpha level for this test set at .05. We conducted power analyses for a difference of the observed size by simulating new datasets based on the model estimates, with 24, 36, or 48 participants (1,000 simulations for each N). Results are in the Table 3 below:

	N=24	**N=36**	**N=48**
one-tailed	.684	.854	.939
two-tailed	.572	.767	.893

Table 3: Power analysis for difference of the observed size for misspecification rate.

A one-tailed test with N=36 yields approximate power of .854 (linear interpolation).

Our second main prediction concerned the differential speech onset latency for appropriately specified descriptions. Onset latency is defined as the time taken to produce the first content word as measured from the onset of the display. Our prediction was that speakers would experience more difficulty shifting from the entrained description to a more contextually appropriate description in the Low Context Variability level than in the High Context Variability level, due to a more fluent retrieval of the entrained response. This analysis excluded trials where the size aspect of the target was misspecified (e.g., using a size adjective when it was unneeded, or failing to use it when needed). Parameters were estimated under maximum likelihood (REML=FALSE) using a linear mixed effects

model with identity link and Gaussian variance. The dependent variable was 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 pilot data suggested that speakers were about 97 milliseconds slower to begin speaking in the Low Context Variability level than in the High Context Variability level (z=1.803, two-tailed p=.0714). A power analysis of these data yielded the following estimates:

	N=24	**N=36**	**N=48**
one-tailed	.889	.964	.996
two-tailed	.808	.929	.983

Table 4: Power analysis for difference of the observed size for differential onset latency

We used a two-tailed test on these data with N=36; estimated power (linear interpolation) is .929.

For the eyetracking data, we predicted a lower proportion of gazes to non-target letters in the grid prior to the onset of speech in the Low Context Variability level than in the High Context Variability level on test trials; this would reflect less consideration of context due to a strong memory signal. Note that we analysed eye tracking data from trials that were appropriately specified (speech that contained no misspecifications). We did not have any pilot eye tracking data for this task, and so it was difficult (and fairly arbitrary) to estimate power.

In sum, we had two key predictions that were pre-registered on the OSF:

(1) A greater misspecification rate in the Low vs. High Context Variability level, alpha=.05, one-tailed;

(2) A greater increase (relative to the last training trials) in speech onset latency for the Low Variability level relative to High Variability, alpha=.05, two-tailed;

We also made two additional (less critical) predictions:

(3) Higher misspecification in the Contrast-Singleton level than in the Singleton-Contrast level (main effect of Shift Direction), alpha=.05, two-tailed;

(4) Fewer non-target fixations prior to speech onset in the Low Variability level than in the High Variability level, alpha=.05, two-tailed.

3.4 – Results

3.4.1 – Statistical Analysis

The statistical analysis for the production data (modifier use and speech onset) was performed using linear mixed-effects models with Directors (subjects) as a random factor (Baayen, Davidson, & Bates, 2008). All analyses attempted to use the maximal random effects structure justified by the design (Barr, Levy, Scheepers, & Tily, 2013), which implies by-subject random intercepts and by-subject random slopes for both main effects (Context Variability and Shift Direction) and their interaction. Item effects are not needed as the items were not repeated across participants (Clark, 1973). We derived p-values using the *t-to-z* heuristic (i.e., deriving p-values from the standard normal distribution for the t statistic), as that enabled us to perform one-tailed tests. Models were estimated using the lme4 package in R (version 1.1-7 or higher). Our analysis of the eye-tracking data used a Poisson regression model to analyse non-target fixations prior to speech onset. Similarly to the production analysis, we used the maximal random effects structure justified by the design. By-subject random intercepts and by-subject random slopes were used for both main effects (Context Variability and Shift Direction) and their interactions. Directors (subjects) were treated as a random factor in this model. The formula for each of our analysis models can be viewed in our pre-registration files on the Open Science Framework: https://osf.io/4akir/.

3.4.2 – Misspecification Rate

The results of the linear mixed-effects model did not reveal any evidence indicating that Directors followed a retrieval fluency heuristic, pre-registered one-tailed test: $z = -1.05$, $p = 0.15$ (see Figure 4). The overall misspecification rate was the same in both the Low Context Variability (17%) and High Context Variability (17%) levels (see Table 5 for the grand means of misspecification rate (%) broken down by Shift Direction and Context Variability). However, contrary to previous findings (Deutsch & Pechmann, 1982; Ferreira et al., 2005; Gann & Barr, 2014) speakers were more likely to underspecify referents in the test trial than overspecify. Thus our predication that there would be higher misspecification in the Contrast-Singleton level than in the Singleton- Contrast level was not supported. This surprising finding resulted in a main effect of Shift Direction in the opposite direction than we had predicted. In the Contrast-Singleton level participants entrained on descriptions (e.g. “the *big* X”) and then overspecified in the test trial (where the modifier “big” is not necessary) at a rate of 9%.

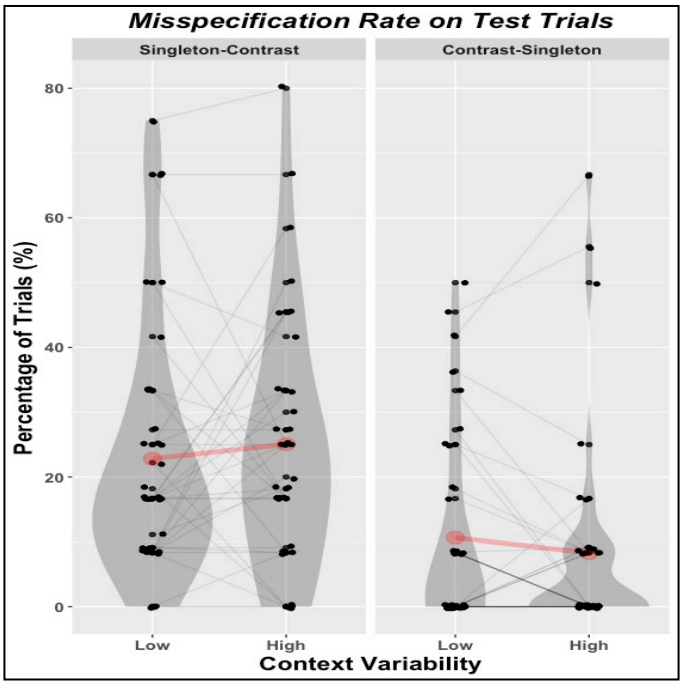


Figure 4: Misspecification rate (%) on test trials shown in both Shift Direction and Context Variability factors. Note that each grey line represents a single participant. The red circles represent the grand means across each level of the Shift Direction and Context Variability factors.

Shift Direction	Context Variability	Misspecification Rate (%)
Singleton-Contrast	Low Variability	22.8
Singleton-Contrast	High Variability	25.1
Contrast-Singleton	Low Variability	10.6
Contrast-Singleton	High Variability	8.2

Table 5: Grand mean misspecification rate (%) by Shift Direction and Context Variability factors.

This was significantly lower than the underspecification rate of 24% in the Singleton-Contrast level where participants entrained upon unmodified descriptions (e.g. "the u") and then encountered a test trial which required a modifier in the description (e.g. "the *small u*"), $z = 4.67$, $p < 0.01$ (Table 6 shows the rate of misspecification (%) broken down by Shift Direction and Modifier Code). Analysis revealed no significant interaction between Context Variability and Shift Direction, $z = 1.73$, $p = 0.08$.

Shift Direction	Modifier Code	Misspecification Rate (%)
Singleton-Contrast	Addition due to Self-repair	60.3
Singleton-Contrast	Addition due to Other-repair	14.6
Singleton-Contrast	No Size Modifier	7.5
Singleton-Contrast	Deleted Adjective	17.6
Contrast-Singleton	Addition due to Self-repair	5.1
Contrast-Singleton	Post-Nominal Modifier	16.5
Contrast-Singleton	Pre-Nominal Modifier	54.4
Contrast-Singleton	Deleted Adjective	24.1

Table 6: Misspecification rate (%) by Shift Direction and type of modifier.

3.4.3 – Speech Fluency Analysis

Fluent speech (FL) is categorised as speech that does not contain any misspecifications or filled/unfilled pauses. Our analysis revealed that there was a similar mean percentage of fluent trials in both the Low Context Variability (94%) and High Context Variability (92%) levels. Table 7 displays the fluent trials (%) broken down by Shift Direction and Context Variability. Whilst there was no significant effect of Context Variability (Low vs. High) on speech fluency, z = -0.45, p = 0.65 there was a significant effect of Shift Direction on speech fluency.

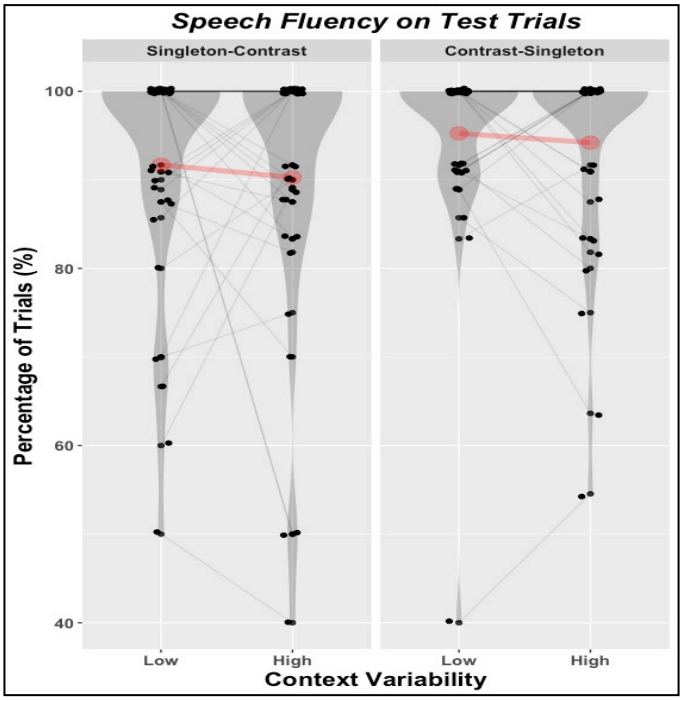


Figure 5: Fluent trials (%) in both Shift Direction and Context Variability factors. Note that each grey line represents a single participant. The red circles represent the average percentage across each level of the Shift Direction and Context Variability factors.

Shift Direction	Context Variability	Fluent Trials (%)
Singleton-Contrast	Low Variability	91.7
Singleton-Contrast	High Variability	90.3
Contrast-Singleton	Low Variability	95.2
Contrast-Singleton	High Variability	94.2

Table 7: Fluent trails (%) by Shift Direction and Context Variability**.**

Results indicated that participants were significantly less fluent in the Singleton-Contrast level (91%) compared to the Contrast-Singleton level (95%), z = -1.97, p = 0.05. Table 8 displays the percentage of trials (%) broken down by speech code. Figure 5 displays the fluent trials (%) across both Shift Direction and Context Variability factors. There was no significant interaction between Context Variability and Shift Direction, z = -0.13, p = 0.89.

Shift Direction	Speech Code	Percentage of Trials (%)
Singleton-Contrast	Fluent Speech	91.0
Singleton-Contrast	Filled Pause	7.9
Singleton-Contrast	Unfilled Pause	1.1
Contrast-Singleton	Fluent Speech	94.7
Contrast-Singleton	Filled Pause	4.6
Contrast-Singleton	Unfilled Pause	0.1
Contrast-Singleton	Other	0.5

Table 8: Percentage of trials (%) for each category of speech code in the Shift Direction factor.

3.4.4 – Differential Speech Onset Latency

Analysis of the differential speech onset latency (mean test trial onset – mean onset of final training trial of non-misspecified trials) did not produce any significant main effects. Thus we did not find any evidence supporting our prediction that there would be a greater increase (relative to the last training trials) in speech onset latency in the Low Context Variability level (average 174.6ms) relative to the High Context Variability level (average 173.5ms), t = -0.50, p = 0.62 (see Figure 6). Further analysis also revealed no significant effect of Shift Direction on onset latency, Singleton-Contrast (134.3ms) vs. Contrast-Singleton (207.2ms), t = -1.28, p = 0.2. Table 9 displays the mean onset change for each condition of Context Variability and Shift Direction. Finally, there was no significant interaction between Context Variability and Shift Direction, t = -1.65, p = 0.1.

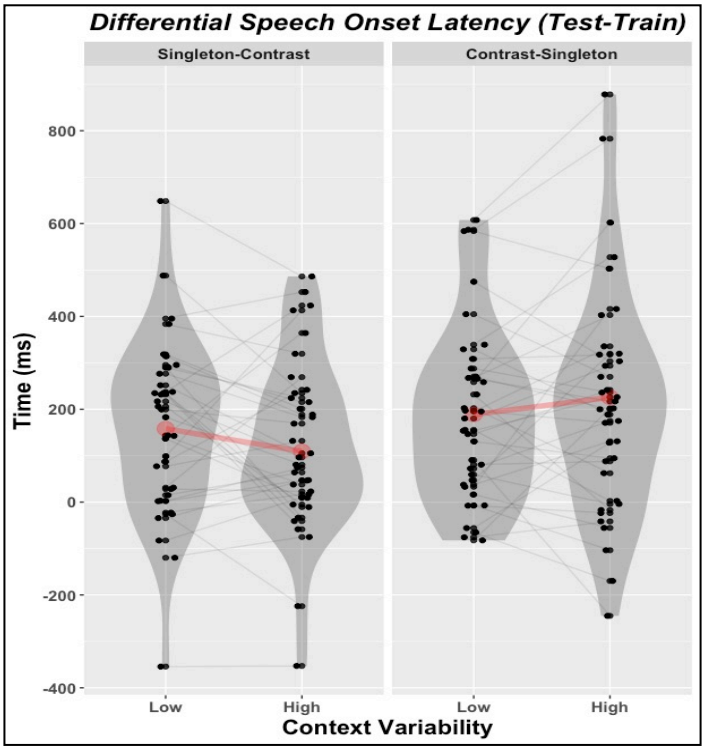


Figure 6: Differential speech onset latency (ms) in both the Shift Direction and Context Variability factors. Note that each grey line represents a single participant. The red circles represent the grand means across each level of the Shift Direction and Context Variability factors.

No. Trials	Shift Direction	Context Variability	Training Onset	Test Onset	Mean Onset Change (ms)
325	Singleton-Contrast	Low Variability	1128.4	1286.5	158.1
308	Singleton-Contrast	High Variability	1153.6	1262.9	109.2
378	Contrast-Singleton	Low Variability	1092.7	1281.4	188.7
379	Contrast-Singleton	High Variability	1101.8	1327.4	225.7

Table 9: Mean onset change (ms) by Shift Direction and Context Variability**.**

3.4.5 – Eye Tracking Analysis

Analysis of the eye-tracking data focussed on non-misspecified descriptions on test trials in the experiment. There was no significant effect of Context Variability on non-target fixations prior to speech onset. We did however, find a slight trend in the direction predicted, with fewer non-target fixations in the Low Variability level (mean = 3.66) than in the High Variability level (mean = 3.76), $z = 0.66$, $p = 0.51$. There was a significant effect of Shift Direction on non-target fixations with participants fixating more on non-target items in the Contrast-Singleton level (mean = 3.93) compared to the Singleton-Contrast level (mean = 3.46), $z = -4.7$, $p < 0.01$. Figure 7 displays the number of fixations across both Shift Direction and Context Variability factors. Table 10 displays the mean number of non-target fixations broken down by Shift Direction and Context Variability. There was no significant interaction between Context Variability and Shift Direction, $z = -0.57$, $p = 0.57$.

Shift Direction	Context Variability	Mean Number of Non-Target Fixations
Singleton-Contrast	Low Variability	3.47
Singleton-Contrast	High Variability	3.45
Contrast-Singleton	Low Variability	3.83
Contrast-Singleton	High Variability	4.02

Table 10: Mean number of fixations by Shift Direction and Context Variability.

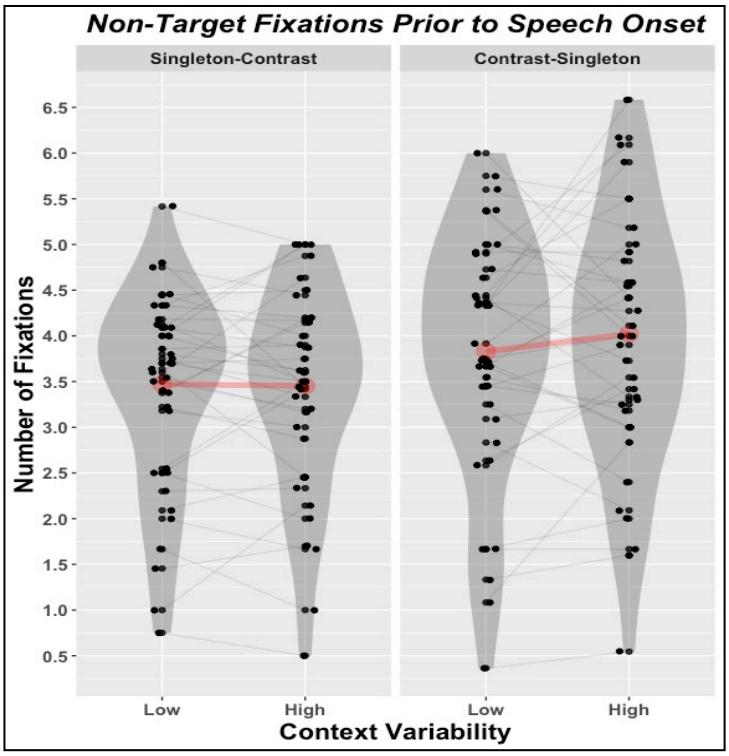


Figure 7: Non-target fixations prior to speech onset in both the Shift Direction and Context Variability factors. Note that each grey line represents a single participant. The red circles represent the grand means across each level of the Shift Direction and Context Variability factors.

3.5 – Discussion

Experiment 1 sought to test the *retrieval fluency hypothesis*: that speakers' use retrieval fluency as a heuristic for audience design in referential communication. We attempted to manipulate the fluency with which Directors retrieved descriptions for target objects by altering how much the context varied each time the description was used. In our study we used the Context Variability (Low vs. High) factor in order to test whether speakers used retrieval fluency as a heuristic when generating descriptions for the addressee. The key idea in this experiment was that Directors who entrained on a description within a *highly variable context* would experience *less fluent retrieval* of that description than Directors who entrained on that same description within a *low variability context*. Our main prediction was that this would cause greater misspecifications in the Low Context Variability level compared to the High Context Variability level.

Our analysis did not reveal evidence that participants followed a retrieval fluency heuristic. The overall misspecification rate was numerically same in both the Low Context Variability (17%) and High Context Variability (17%) levels. Nevertheless, this does suggest that speakers did rely on their memory of previously encoded descriptions to a

certain extent. Had speakers not been influenced by their memory, it is unlikely that we would have seen such misspecification at the test phase. Our eye tracking analysis only considered data for non-misspecified descriptions on test trials in each block of trials. Analysis revealed that speakers made fewer non-target fixations prior to speech onset in the Low Variability level (mean = 3.66) compared to the High Context Variability level (mean = 3.76), however this was not a statistically significant difference. There was no significant difference in onset change between the Low (174.6ms) and High (173.5ms) Context Variability levels.

Further analysis of the eye tracking data showed that participants made significantly fewer non-target fixations prior to speech onset in the Singleton-Contrast level (mean = 3.46) compared to the Contrast-Singleton level (mean = 3.93). Since participants were already primed to check the context for a competitor letter at the training phase in the Contrast-Singleton level, it is likely that this result reflects similar behaviour at the test phase. Participants' speech was significantly less fluent in the Singleton-Contrast level (91%) compared to the Contrast-Singleton level (95%). However, it should be noted that due to the low effect size of this speech fluency effect we are reluctant to draw any firm conclusions from this result. Our results revealed that Directors were significantly more likely to *underspecify* referents in the test trial than overspecify. Participants underspecified at a rate of 24% (Singleton-Contrast level) compared to an overspecification rate of 9% (Contrast-Singleton level).

Although it contradicts Grice's (1975) Cooperative Principle, speakers are often likely to overspecify their descriptions by providing more information than is required to identify the target object (Koolen, Gatt, Goudbeek, & Krahmer, 2011). The traditional cognitive view of referential behaviour argues that speakers will design their utterances in order to enable the addressee to efficiently locate the target object (Arnold, 2008). Thus speakers may adopt an "addressee oriented" approach to referential descriptions. In line with this argument speakers will overspecify in order to enable the addressee to find referent objects more quickly (Koolen et al., 2011). As mentioned previously, this view is supported by a number of studies which show that addressees find it easier to identify an object when the speaker overspecifies their description (Nadig & Sedivy, 2002; Paraboni, Masthoff, & van Deemter, 2006; Sonnenschein, 1984; Sonnenschein & Whitehurst, 1982).

It was therefore surprising that we found the opposite effect in our study with underspecifications occurring more frequently than overspecifications. It was expected that if participants misspecified in their descriptions then overspecification would have been

more likely as underspecifying can often confuse addressees (Horton, 2008). Note that participants' speech was also less fluent in the Singleton-Contrast level. The underspecification effect is underlined by the finding that participants' speech contained more misspecifications and filled/unfilled pauses in this level of the Shift Direction factor. In the Singleton-Contrast level participants are expected to provide additional information in their description at the test trial (e.g. going from "the u" - > "the *small* u"). Notably, participants altered their own descriptions (AS) at rate of 60% in this level. Furthermore, the Matcher requested additional information (AO) at a rate of 15%. In comparison in the Contrast-Singleton level where participants are expected to shorten their descriptions (e.g. going from "the *big* X" - > "the X") participants altered their own descriptions (AS) at a rate of 5% and were never asked for additional information (AO) from the Matcher. In the Singleton-Contrast level participants' speech was less fluent (FL = 91%) and contained more filled pauses (FP = 8%) compared to the Contrast-Singleton level where fluency was higher (FL = 95%) and filled pauses (FP = 5 %) were at a lower rate.

We are uncertain as to why we found a significant underspecification effect. It is possible that due to the nature of the stimuli in the experiment (all items were letters of varying font size and colour) participants became overly familiar with contents of the grid in each trial presentation. Perhaps this led participants to adopt a lackadaisical approach when describing target items causing them to put less effort into their descriptions in the Singleton-Contrast level. It is also possible that the reduced use of post-nominal modifiers (in comparison to pre-nominal modifiers) in this experiment had an impact on the misspecification rate – this is an aspect we discuss more thoroughly in Chapter 6.

Furthermore, the stimuli set used in this experiment may have hindered the development of retrieval fluency associations within memory. Participants may have failed to develop strong memory associations with the descriptions they used for each target item. This may have resulted in an overall lack of retrieval fluency effect which would explain why the misspecification rate was similar in both the Low Context and High Context Variability levels.

An additional concern for our experiment was that the order and sequencing of training trials may have prevented participants from developing stronger memory associations with target objects. Each sub-block of trials contained a mix of stimuli in both Low Context and High Context Variability levels. It is possible that this mix of trials in each sub-block counteracted each other, resulting in an overall lack of effect of the Context Variability

factor on misspecification rate. This could have reduced the level of retrieval fluency participants experienced at the Low Context Variability level.

In order to address some of these concerns, we decided to revaluate the design of our study for Experiment 2. We opted to change a number of aspects of the design and presentation of our experiment. Most notably we altered our stimuli set and decided to use more distinguishable target objects as opposed to letters in each grid. We also altered the sequencing of trials in the training phase of the experiment. In the following chapter, I will outline our alternative design which formed our second attempt to test the retrieval fluency hypothesis.

Chapter 4 – Experiment 2

4.1 – Background

4.1.1 – Retesting the Retrieval Fluency Hypothesis

Experiment 2 marks our second attempt to test the retrieval fluency hypothesis: that attending to a referent with the goal of referential encoding elicits retrieval of previous referential expressions used for a particular referent, proportionally to the match between encoding and retrieval contexts. Accordingly, this hypothesis proposes that speakers use the strength/fluency of these memory signals as a cue to their informational adequacy in the current communicative situation. As outlined in Chapter 3, 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. We also draw on Logan's (1988) Instance theory of Automaticity (ITA) which argues that performance becomes automatic when it is grounded on directly accessed memory retrieval of past solutions. Thus 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 weaker fluency.

As this was our second attempt to test the retrieval fluency hypothesis we made several changes to the design and presentation of the experiment. Firstly, we replaced the stimuli set with an entirely new collection of target items (with matched foil and competitor item pairs). These were normed by a separate group of volunteers beforehand (see section *4.2.2 - Norming of Test Items* for more details). Similarly, to previous studies (e.g. Engelhardt, Bailey, & Ferreira, 2006; Gann & Barr, 2014; Keysar, Barr, Balin, & Brauner, 2000) we opted to use everyday objects as our stimuli (e.g. car, apple, bat). We were concerned that one of the reasons that participants failed to demonstrate a retrieval fluency effect in Experiment 1 was due to the nature of the stimuli in the experiment (all target items and fillers were letters of varying font size and colour). Our new target objects were carefully selected to ensure that each object was unique in identity from other target items. As participants were encoding different types of objects we expected their memory traces for each item to be more distinctive. Directors were required to use different types of

modifiers in their utterances (e.g. "the *family* car" vs. "the *sports* car") in comparison to Experiment 1 where they were only ever required to provide a size modifier (e.g. "the small u" vs. "the big U"). Since these objects had more distinctive features we expected participants to build up stronger memory traces for their utterances thus creating more fluent memories of the descriptions used with each target item.

An additional alteration concerned the presentation sequence of trials in the Context Variability factor. Each sub-block in Experiment 1 contained a mix of trials from both the Low Context and High Context Variability levels. This may have prevented the development of a fluency effect in the Low Variability Context level and help to explain why we found no main effect of retrieval fluency on audience design in Experiment 1. In Experiment 2, we decided to alter this format. We replaced the Context Variability factor with a new *Training-Test Consistency* factor. This factor reconfigured the arrangement of trials in the training phase of the experiment. Details of all modifications are outlined in the section below.

4.1.2 – Formulation of Alternative Design

Similarly to our first experiment, Experiment 2 contained two factors in the design. Shift Direction (Singleton-Contrast vs. Contrast-Singleton) remained in the same format as the previous experiment and Training-Test Consistency (Training Consistent vs. Training Inconsistent) formed the second factor:

Shift Direction Factor

In each sequence, the target object appeared with a "critical" object, whose identity formed the factor of *Shift Direction*. 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 level (Singleton-Contrast; see Figure 9), 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, see Figure 8 and Figure 9 for an example of the foil). For the test trial in this level, the foil was replaced with the competitor object, which was another object from the same category as the target (e.g., a car) but differed in some critical way (e.g., a sports car), thus requiring speakers to modify their descriptions ("the car" -> "the *family* car").

In the Contrast-Singleton level (see Figure 9 for an example) this order was reversed: the critical object during training was the competitor (e.g., the "sports car"), leading speakers to entrain upon a modified expression during training. At test, the competitor was then replaced with the foil, such that participants were able to simplify their description of the target item ("the family car" -> "the car"). Similarly to the design implemented by Gann and Barr (2014), the Shift Direction factor enabled us to provide opportunities for participants to underspecify (Singleton-Contrast) or overspecify (Contrast-Singleton) descriptions on test trials.

In addition to the critical item, each display also contained other *filler* items (objects unrelated to the target item in each display). The relation of the arrangement of these items during training to their arrangement during test formed the critical manipulation of Training-Test Consistency.

Training-Test Consistency

The trials were presented in blocked sequences with the order counterbalanced across participants. Unlike our previous experiment, all training trials presented had 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 Training Consistent level (previously the "Low Context Variability" level in Experiment 1) the configuration of items in the display at training was highly similar to the configuration presented at test. In the Training Inconsistent level, the configuration of items in the training displays were highly dissimilar to test. In line with our hypothesis in Experiment 1, we reasoned that in attempting to referentially encode the target item at test, speakers in the Training 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) would 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 (see Figure 9 for an example).

4.1.3 – Pre-registered Predictions

As in Experiment 1 the basis for our predictions was a pilot study containing 22 participants (with 24 sequences per participant, whereas our main study contained 48 sequences). This pilot study is available on github (https://github.com/dalejbarr/EESP2) as

well as in our files for Experiment 2 on the Open Science Framework (OSF: https://osf.io/uq4k7/). We pre-registered all our predictions on the OSF (outlined in section *4.3.4 – Pre-registered Analysis and Predictions*). Our main prediction was that speakers would misspecify referents at a higher rate in the Consistent Training-Test level than in the Inconsistent Training-Test level.

4.2 – Method

4.2.1 – Participants

Thirty-six subjects completed the experiment (24 Females, M=23.2 years). All subjects were recruited from the campus at the University of Glasgow. Unlike Experiment 1, all subjects were Native English speakers. Participants were paid £6 or received 4 "participation credits" (course credits) for taking part in the study. One participant was replaced due to the use of ineffective descriptions during the task (continuously failing to adapt their utterances for the listener, please see Section 4.3.3 – *Exclusion Criteria for Participant Responses* for more details). Subjects gave written informed consent before beginning the experiment and were fully debriefed after the experiment had finished. Our procedures fully complied with the ethical code of conduct of the British Psychological Association.

4.2.2 – Norming of Test Items

Target and Competitor items were normed beforehand by 68 Native English speaking volunteers using the web-based surveyor SurveyMonkey. A number of items were updated or replaced based on our norming feedback. Four entirely new stimuli pairs were added to our original list (please see Appendix 3 for a complete list of the Target and Competitor objects used).

4.2.3 – Experimental Setup and Task

Similarly to our first study, the experiment was interactive with the participant playing the role of the "Director" (the speaker) and the experimenter playing the role of the "Matcher" (the listener). In this experiment, the Matcher was played by either a male or a female lab assistant. The Director and the Matcher sat in different areas of the testing room and looked at separate computer monitors throughout the experiment. Both were seated facing in opposite directions so that they were unable to see each other's display (see Figure 8). In each trial, the Director was asked to describe a target object which was highlighted on their monitor to the Matcher. The Matcher then identified this object on his/her own screen and

selected it using a computer mouse. The target object appeared on the Director's screen within a grid among other "filler" objects (see Figure 9). The Director was informed that in each trial the Matcher had the same objects on their monitor but that they may be arranged in a different format compared to the grid that appeared on their screen.

4.2.4 – Design

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

4.2.5 – 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 consisted of a five-by-four grid containing objects of different size and colour (see Figures 8 and 9). The experiment contained 48 "sequences" of trials, each consisting of a number of training trials followed by a single test trial (the term "sequence" to refers 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 design, counterbalanced across participants.

For each sequence, the number of training trials was randomly selected, with a range from 6 to 9. The motivation for varying training sequence length was to make the occurrence of the test trial unpredictable. Given these parameters, each experimental session could have contained between 336 (7 x 48) and 480 (10 x 48) trials. For each sequence, 7 to 10 filler items were randomly chosen from a database of stimulus images. The displays were then checked manually by two lab assistants to ensure that the filler items were sufficiently dissimilar to the target so as not to influence descriptions of the target.

4.2.6 – Apparatus

The experimental stimuli were presented on a 19" LCD Dell desktop computer monitor (4:3 aspect ratio, resolution 1024 x 768 pixels). Participants were seated 45-55cm away from the monitor. A microphone was placed above the participant's computer monitor to record their descriptions of the *target* object for each trial. The audio was tagged using Audacity 2.0.6 software. Eye movements were recorded during each trial using an Eyelink 1000 (SR Research) eye tracker (sampling rate 500Hz).

4.2.7 – 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.

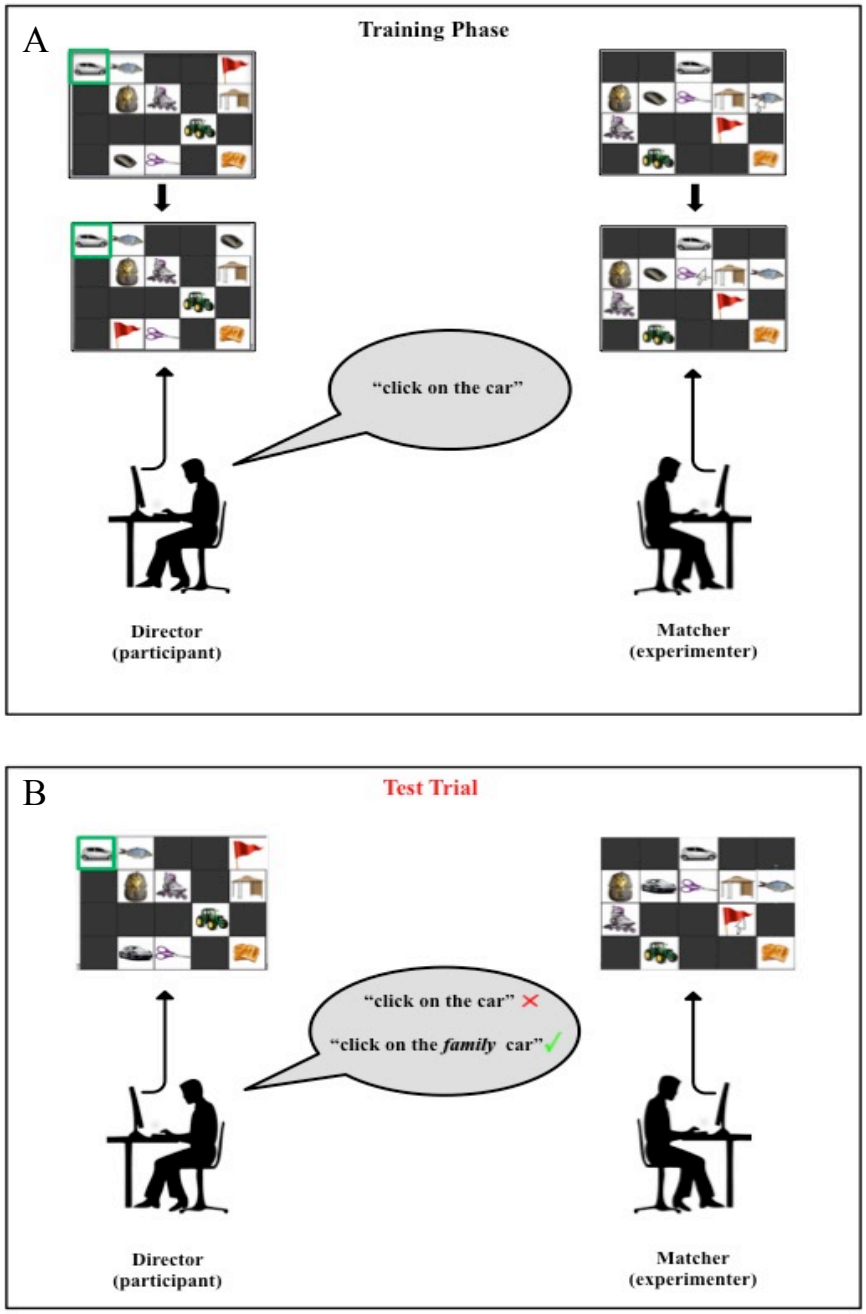


Figure 8: Outline of experimental set up and procedure for Experiment 2.

Panel A shows the Director (participant) and Matcher (experimenter) during the training phase (6-9 trials). Each grid of objects presented on the left-hand side shows an example of a training trial from the Director's perspective. Each of the grids shown on the right-hand side show the corresponding trials viewed by the Matcher. Both the Director and Matcher face in opposite directions looking at separate computer monitors. Once the Director has provided a description (*"click on the car"*) the Matcher will select the appropriate target letter on their screen in order to move onto the next trial. Panel B shows the test phase of the sequence - a competitor object appears (the *sports car*) and the Director is required to adapt their description for the Matcher. Stimuli are shown at the Training Consistent and Singleton-Contrast levels.

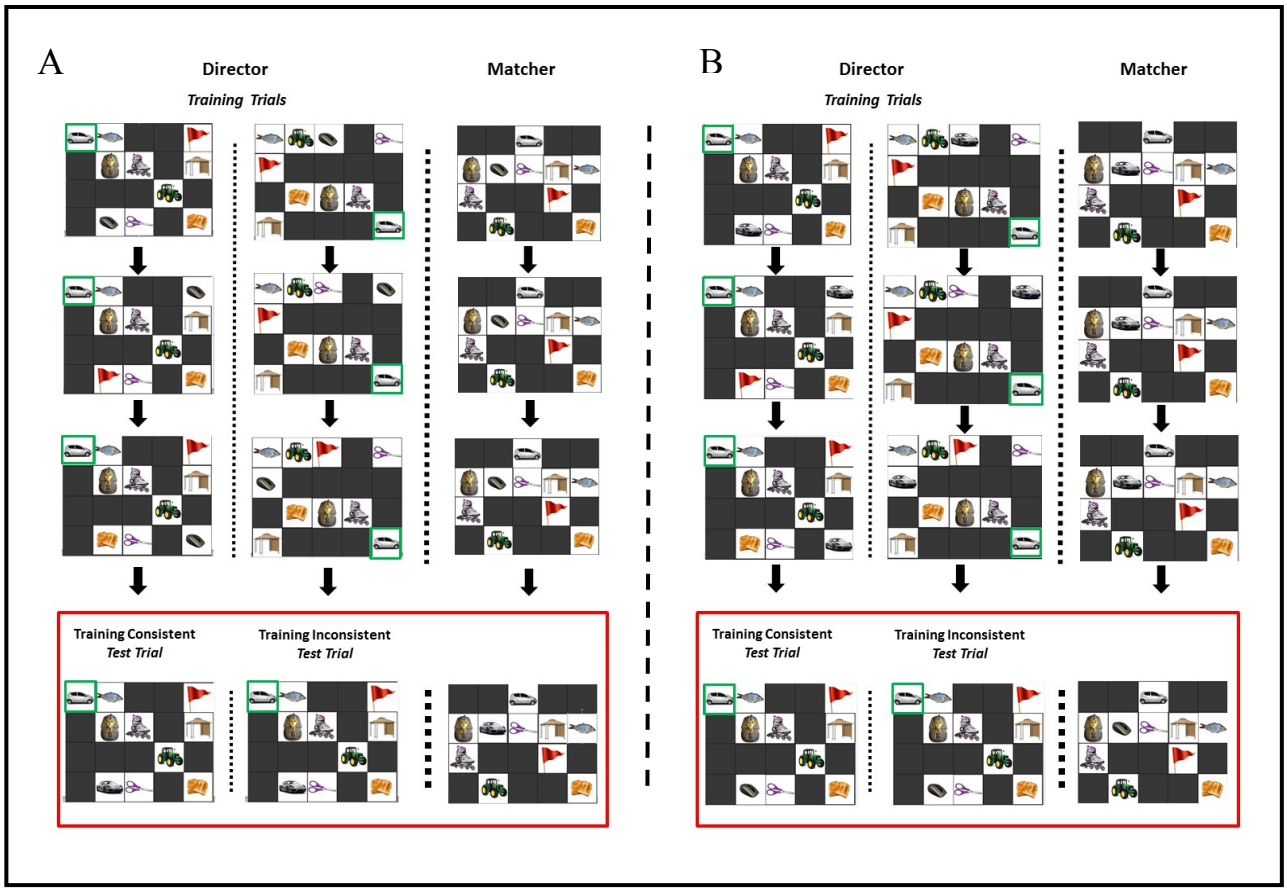


Figure 9: Overview of trials in both Training-Test Consistency and Shift Direction Factors.

Panel A shows an example of 3 training trials followed by a single test trial in the *Singleton-Contrast* level. 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 - the *Training Consistent* level. The middle column shows the alternative *Training Inconsistent* level. The column on the right shows the Matcher's view. 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 "*sports car*" object. This may prompt the Director to *underspecify* their description of the target object to the Matcher - *"click on the **car**"* whereas the description *"click on the **family car**"* would be more appropriate in this instance. Note that the training trials also present the "computer mouse" which acts as a foil for the *"sports car"* during the training phase. The Matcher's view remains fixed throughout the training and test phase with the *"sports car"* replacing the "computer mouse" in the test trial. Panel B shows an example of 3 training trials followed by a single test trial in the *Contrast-Singleton* level. The column on the left shows the Director's view of the stimuli at the *Training Consistent* level. The middle column shows the alternative *Training Inconsistent* level. The column on the right shows the Matcher's view. The training trials highlight the target object - *"the **car**"*. Note that the "*sports car"* is also present in the grid. Participants are likely to differentiate between the two cars during the training phase - "*click on the **family car**"*. The test trial presents participants with the target object *"the **car**"* again, but unlike the training trials the computer mouse foil replaces the *"sports car"*. This may prompt the Director to *overspecify* their description of the target object to the Matcher. The description *"click on the **car**"* would be sufficient in this instance. The Matcher's view remains fixed throughout the training and test phase with the "computer mouse" replacing the *"sports car"* in the test trial.

4.2.8 – Procedure

Upon arrival each participant was given an 'instruction' sheet detailing the task and their role during the experiment (see Appendix 4). Participants sat opposite the eye tracker and computer screen. The experimenter sat behind the participant facing a separate computer monitor. The layout of the room was designed so as to ensure that neither the participant nor the experimenter were able to see the other's monitor (please see Figure 8 for an example of the layout). The participant played the role of "Director" and the experimenter played the role of the "Matcher".

Similarly to Experiment 1, in each trial the Director was asked to verbally name the target object so that the Matcher could identify the item on their monitor and select it using a computer mouse. In order to discriminate the target object from the filler objects, the target was highlighted within a green square in the Director's display (see Figure 8 and Figure 9). The participant was told that the arrangement of images within the Matcher's grid would differ in an unpredictable way from the images shown on their own screen. Thus the Director was informed that he/she would have to describe the features of the highlighted target item, rather than use the target's grid location as a description. Unbeknown to the participant, the Matcher's view of the stimuli was fixed for each sequence so that the objects always appeared in the same location across the training and test trials – with the foil/competitor item trading places with each other on the test trial (see Figure 9). This alteration was influenced by performance in Experiment 1. Occasionally in the first experiment, the Matcher took longer to find the intended target item within the grid - even when the Director had provided an adequate description. We were concerned that this may have led speakers to incorrectly believe that they had provided an inadequate description to the Matcher. To prevent this from becoming a factor which influenced descriptions in Experiment 2, the Matcher's view was fixed to enable the experimenter to quickly identify the target object without disrupting the build-up of retrieval fluency effects experienced by the participant.

Unlike Experiment 1, which had a preview of the target location before the full set of images appeared, the location of the target object appeared at the same time as the rest of the images within the grid. Audio recording of the Director's response began simultaneously with the presentation of the main display. Each trial ended when the Matcher selected the object designated by the Director. The Director could not see the Matcher's screen or mouse pointer, and received no feedback regarding whether the trial was completed correctly. If the Director failed to provide sufficient information for the

Matcher to identify the target, the Matcher asked the Director for clarification (e.g. "which one do you mean?"). Any such clarification exchanges appeared in the audio recording for the trial and were noted during later transcription (see section *4.3.2 – Transcription and Coding of Audio Files* for details).

4.3 – Predictions and Data Analysis

4.3.1 – Main Measurements

Our analysis focussed 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.

4.3.2 – Transcription and Coding of Audio Files

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

Speech 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"

Table 11: Outline of speech fluency categories.

We also coded whether or not a descriptive modifier was used, defined by the following categories:

Modifier 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"

Table 12: Outline of item modifier categories.

* Note that a colour description was not coded as a modifier if it did not distinguish the target object from the competitor (for instance both the family car and the sports car were silver in colour).

Similarly to Experiment 1, onset times of utterances were measured in milliseconds (ms). The following criteria were applied when identifying utterance onsets:

- Trials were discarded if the speech was unidentifiable.
- Any filled pauses or articles were ignored (um, uh, the); speech onset was 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 as the onset of the word *"silver"*).
- If Directors corrected themselves after an error (e.g. *"white car...eh sorry silver car"*) onset of the correction (i.e. *"silver"*) was recorded. However, such repaired utterances were not used in the analysis of speech onset.

4.3.3 – Exclusion Criteria for Participant Responses

Similarly, to Experiment 1 we were concerned that some Directors may have opted for a strategy of "hyper-describing" target objects i.e. 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 level, Directors could simply continue using the modified description, which would then spuriously appear to be appropriately specified. We identified any participants doing this by coding whether or not in the final training trial for each sequence in the Singleton-Contrast level, they inappropriately described the modifier in a way that would have differentiated the target from the (absent) competitor. We removed all data from speakers who made this error on more than half of the training trials and replaced these participants. Unlike Experiment 1 where 11 participants were replaced, only 1 subject was replaced in this study (please see Appendix 5 for details).

For all of the remaining participants, we also excluded on a trial-by-trial basis any test trials in the Singleton-Contrast level where on the last training trial speakers used a modifier that distinguished the target from the competitor. In the Contrast-Singleton level, this was less of an issue because the speakers had to use size modifiers during training or the addressee would have been unable to resolve the reference. However, for any trials where the speaker repaired an utterance (for example "*the car, uh the family car*") in the last training trial for this condition, we discarded the following test trial from the analysis.

Furthermore, we also checked the quality of the materials to determine whether there were certain stimulus items that should be excluded. In particular, we considered the last training trial of each series for each item in which the critical object was a foil, and removed from the analysis any target item for which more than 50% of speakers used a description that distinguished it from the corresponding competitor. In total eight of our stimuli pairs (target and competitor items) were removed (please see Appendix 5 for a full list of the items).

4.3.4 – Pre-registered Analysis and Predictions

We pre-registered our analysis and predictions on the Open Science Framework. The basis for our estimate of a sample size of 36 participants (power = .85) was derived from our pilot study conducted prior to Experiment 1 (see section *3.3.4 - Pre-registered Analysis and Predictions* outlined in Chapter 3 for details). Similarly to our main prediction of Context Variability (a greater rate of misspecification in the Low vs. High Context Variability condition) in Experiment 1, our main prediction in this study concerned the Training-Test Consistency factor. Specifically, we predicted that speakers would misspecify referents at a higher rate in the Consistent Training-Test level than in the Inconsistent Training-Test level.

To maximize power (which was especially important given that the dependent variable for this analysis was binary, 1 = modifier used in description, 0 = no modifier), we opted to test for the main effect of Training-Test Consistency using a *one-tailed test* (see the methodology for Experiment 1 for further information about the power calculation). Although Experiment 1 was unsuccessful, we believed 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 made to the procedure to improve sensitivity we did not see any reason to increase our sample size for this experiment.

Similarly to Experiment 1, our second main prediction concerned the differential speech onset latency for appropriately specified descriptions. Our prediction was that speakers would experience more difficulty shifting from their entrained description to a more contextually appropriate description in the Consistent Training-Test level than in the Inconsistent Training-Test level, due to a more *fluent retrieval* of the entrained response. This analysis only included trials where the target was appropriately specified both at test as well as in the last training trial before test. The dependent variable was 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 previously conducted power analysis suggested .93 power for a two-tailed test with N = 36.

For the eye tracking data, we predicted a lower proportion of gazes to non-target images in the grid prior to the onset of speech in the Consistent Training-Test level than in the Inconsistent Training-Test level in the test phase; this would reflect less consideration of context due to a strong memory signal. Note that we analysed eye tracking data from trials that were appropriately specified (i.e. speech that contained no misspecifications).

In sum, we had two key predictions:
(1) A main effect of Training-Test Consistency on misspecification, with more frequent misspecification in the Consistent level, alpha=.05, one tailed;
(2) For appropriately specified descriptions, a main effect of Training-Test Consistency on differential onset latency (relative to the last training trial), with longer relative delays in the Consistent level, alpha=.05, two-tailed;

We also made two additional (less critical) predictions:
(3) Greater rate of underspecification than overspecification (based on the result from Experiment 1); in other words, a higher rate of misspecification in the Singleton-Contrast level than in the Contrast-Singleton level, alpha=.05, two-tailed;
(4) Fewer non-target fixations prior to speech onset in the Training-Test Consistent level than in the Inconsistent level, alpha=.05, two-tailed.

4.4 – Results

4.4.1 – Statistical Analysis

The statistical analysis for the production data (modifier use and speech onset) was performed using linear mixed-effects models with Directors (subjects) and sequence (stimuli items) as crossed random factors (Baayen et al., 2008). All analyses attempted to use the maximal random effects structure justified by the design (Barr, et al., 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 derived *p*-values using the t-to-z heuristic (i.e., deriving p-values from the standard normal distribution for the t statistic), as that enabled us to perform one-tailed tests. Models were estimated using the lme4 package in R (version 1.1-7 or higher). All independent variables were deviation coded. The analysis of modifier use assumed a logit link and binomial variance function, whereas the analysis of onset times used an identity link with a Gaussian variance function. Our analysis of the eye-tracking data used a Poisson regression model to analyse non-target fixations prior to speech onset. For this analysis we used the maximal random effects structure justified by the design. By-subject random intercepts and by-subject random slopes were used for both main effects (Training-Test Consistency and Shift Direction) and their interactions. Directors (subjects) and sequence (stimuli items) were treated as random factors in this model. The formula for each of our analysis models can be viewed in our pre-registration files on the Open Science Framework: https://osf.io/uq4k7/.

4.4.2 – Misspecification Rate

Analysis revealed a main effect of retrieval fluency on misspecification. The overall misspecification rate was considerably higher than in Experiment 1. Misspecification in the Consistent Training-Test level (previously Low Context Variability in Experiment 1) was at 85% compared to the Inconsistent Training-Test level (previously High Context Variability) which was at 80% (pre-registered one-tailed test: $z = 1.89$, $p = 0.03$). Figure 10 shows a breakdown of the misspecification rate by Shift Direction and Training-Test Consistency. This result suggests that Directors experienced greater levels of retrieval fluency in the Consistent Training-Test level causing them to make significantly more errors in their descriptions to the Matcher (see Table 13 for the grand means of misspecification rate (%) broken down by Shift Direction and Training-Test Consistency).

In line with previous research (Deutsch & Pechmann, 1982; Ferreira et al., 2005; Gann & Barr, 2014) we found that when participants misspecified they were more likely to overspecify descriptions than underspecify. This result was in contrast to Experiment 1. In the Contrast-Singleton level participants entrained on descriptions (e.g. *"the family car"*) and then overspecified in the test trial at a rate of 89% (e.g. where the utterance *"the car"* was adequate). This was significantly higher than the underspecification rate of 74% in the Singleton-Contrast level where speakers entrained upon unmodified descriptions (*"the car"*) and then encountered a test trial which required a modification ($z = 5.05$, $p < 0.01$). Thus our prediction (based on the results of Experiment 1) that there would be higher misspecification in the Singleton-Contrast level than in the Contrast-Singleton level was not supported.

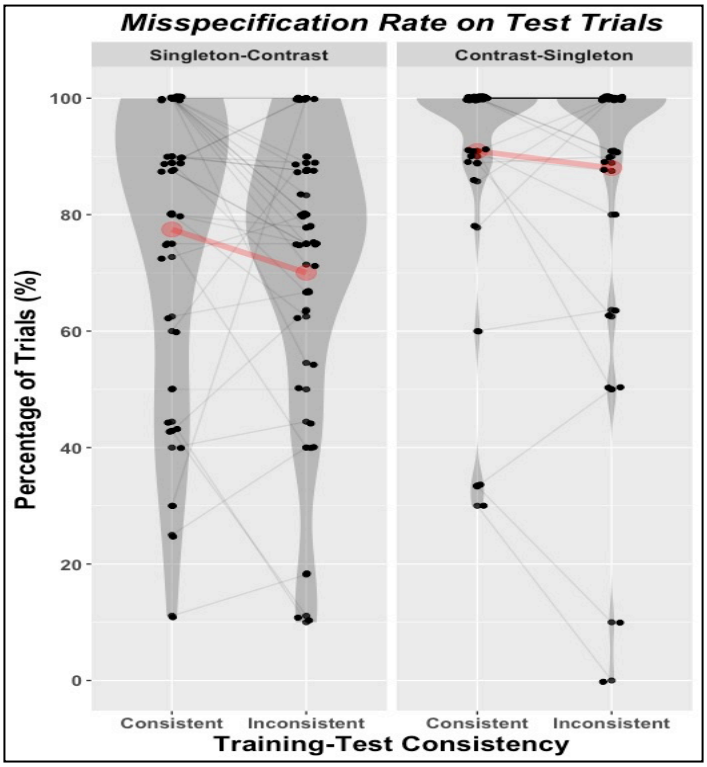


Figure 10: Misspecification rate (%) on test trials shown in both Shift Direction and Training-Test Consistency factors. Note that each grey line represents a single participant. The red circles represent the grand means across each level of the Shift Direction and Training-Test Consistency factors.

Shift Direction	Training-Test Consistency	Misspecification Rate (%)
Singleton-Contrast	Consistent	77.5
Singleton-Contrast	Inconsistent	70.0
Contrast-Singleton	Consistent	91.0
Contrast-Singleton	Inconsistent	88.0

Table 13: Grand mean misspecification rate (%) by Shift Direction and Training-Test Consistency factors.

Table 14 shows the rate of misspecification (%) broken down by Shift Direction and Modifier Code. Analysis revealed no significant interaction between Training-Test Consistency and Shift Direction, $z = -0.03$, $p = 0.97$.

Shift Direction	Modifier Code	Misspecification Rate (%)
Singleton-Contrast	Addition due to Self-repair	25.6
Singleton-Contrast	Addition due to Other-repair	66.7
Singleton-Contrast	Addition due to Other/Self	1.1
Singleton-Contrast	Post-Nominal Modifier	0.2
Singleton-Contrast	Pre-Nominal Modifier	0.2
Singleton-Contrast	No Modifier	4.9
Singleton-Contrast	Deleted Adjective	0.2
Singleton-Contrast	Deleted adjective/Addition self	0.2
Contrast-Singleton	Post-Nominal Modifier	13.4
Contrast-Singleton	Pre-Nominal Modifier	83.1
Contrast-Singleton	No Modifier	1.3
Contrast-Singleton	Deleted Adjective	1.0
Contrast-Singleton	Deleted Adjective/Addition Self	0.3
Contrast-Singleton	Other	0.8

Table 14: Misspecification rate (%) by Shift Direction and type of modifier.

4.4.3 – Speech Fluency Analysis

Fluent speech (FL) is categorised as speech which does not contain any misspecifications or filled/unfilled pauses. Our analysis revealed a similar mean number of fluent trials in both the Consistent (6%) and Inconsistent (8%) Training-Test Consistency levels. Table 15 displays the percentage of trials (%) broken down by speech code. Table 16 displays the fluent trials (%) broken down by Shift Direction and Training-Test Consistency. There was no significant effect of Training-Test Consistency (Consistent vs. Inconsistent) on speech fluency, $z = -0.83$, $p = 0.41$. However, results did show a significant effect of Shift Direction (Singleton-Contrast vs. Contrast-Singleton), $z = -2.16$, $p = 0.03$ on speech fluency. Participants were significantly more fluent in the Singleton-Contrast level (9%) compared to the Contrast-Singleton level (6%). Figure 11 displays the fluent trials (%) across both Singleton-Contrast and Contrast-Singleton conditions. Finally, there was no significant interaction between Training-Test Consistency and Shift Direction, $z = -0.43$, $p = 0.67$.

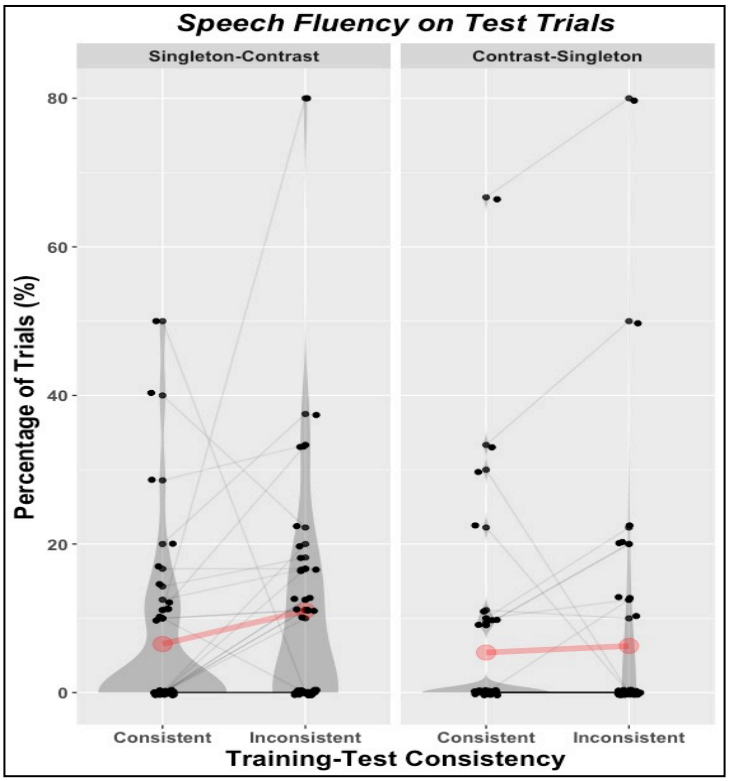


Figure 11: Fluent trials (%) in both Shift Direction and Context Variability factors. Note that each grey line represents a single participant. The red circles represent the average percentage across each level of the Shift Direction and Training-Test Consistency factors.

Shift Direction	Speech Code	Percentage of Trials (%)
Singleton-Contrast	Fluent Speech	33.5
Singleton-Contrast	Filled Pause	40.5
Singleton-Contrast	Filled Pause/Lengthened speech	1.3
Singleton-Contrast	Lengthened Speech	6.3
Singleton-Contrast	Lengthened Speech/Unfilled Pause	0.6
Singleton-Contrast	Unfilled Pause	3.8
Contrast-Singleton	Fluent Speech	55.4
Contrast-Singleton	Filled Pause	32.4
Contrast-Singleton	Unfilled Pause	4.1

Table 15: Percentage of trials (%) for each category of speech code in the Shift Direction factor.

Shift Direction	Training-Test Consistency	Percentage of Fluent Trials (%)
Singleton-Contrast	Consistent	6.5
Singleton-Contrast	Inconsistent	11.0
Contrast-Singleton	Consistent	5.4
Contrast-Singleton	Inconsistent	6.3

Table 16: Fluent trials (%) by Shift Direction and Training-Test Consistency.

4.4.4 – Differential Speech Onset Latency

Analysis of the differential speech onset latency (mean test trial onset – mean onset of final training trial) revealed a main effect of Training-Test Consistency, t = -2.09, p = 0.04. Participants took significantly longer in the Inconsistent level (average 948.2ms) compared to the Consistent level (average 672.2ms) to provide an adequate description to the addressee in the test phase. Further analysis also revealed a significant effect of Shift Direction (Singleton-Contrast vs. Contrast-Singleton) on onset latency, t = -4.44, p = <0.01. Participants took longer to begin their descriptions in the Singleton-Contrast level (1032ms) compared to the Contrast-Singleton level (420.5ms). Table 17 shows the mean onset change broken down by each level of Shift Direction and Training-Test Consistency.

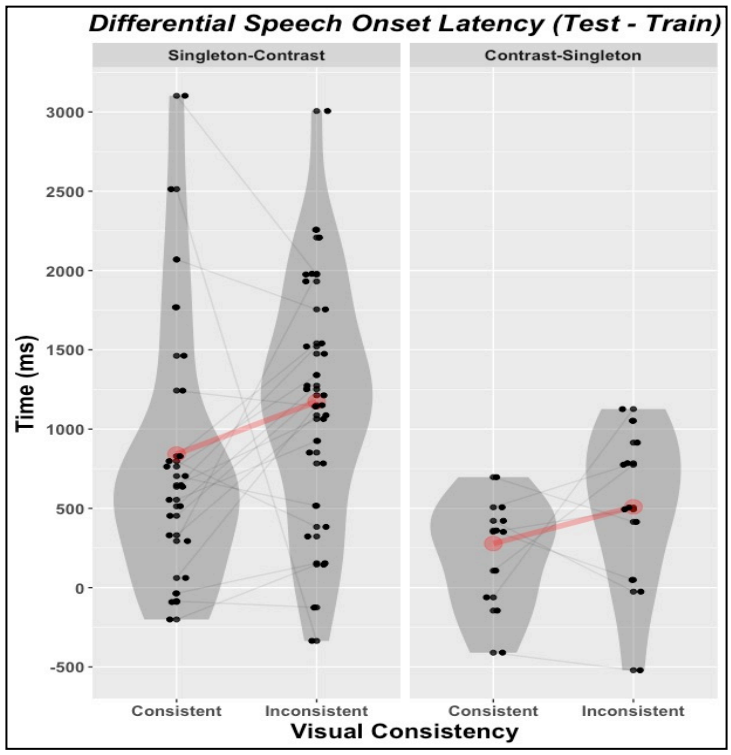


Figure 12: Differential speech onset latency (ms) in both Shift Direction and Context Variability factors. Note that each grey line represents a single participant. The red circles represent the grand means across each level of the Shift Direction and Training-Test Consistency factors.

Figure 12 displays the differential speech onset latency for each condition of Training-Test Consistency and Shift Direction. Finally, there was no significant interaction between Training-Test Consistency and Shift Direction, t = 0.22, p = 0.83.

No. Trials	Shift Direction	Training-Test Consistency	Training Onset (ms)	Test Onset (ms)	Differential Onset Latency (ms)
69	Singleton-Contrast	Consistent	1582.3	2457.5	842.1
89	Singleton-Contrast	Inconsistent	1514.3	2670.9	1173.2
32	Contrast-Singleton	Consistent	1749.7	2148.9	278.0
42	Contrast-Singleton	Inconsistent	1667.8	2210.1	509.5

Table 17: Mean onset change (ms) by Shift Direction and Training-Test Consistency.

4.4.5 – Eye Tracking Analysis

Analysis of the eye-tracking data revealed no significant effect of Training-Test Consistency on non-target fixations prior to speech onset. Similarly to Experiment 1, analysis focussed on non-misspecified descriptions at the test phase of the experiment. We found a slight trend in the direction predicted - with fewer non-target fixations in the Training Consistent level (mean = 6.09) than in the Training Inconsistent level (mean = 6.43), but this difference was not significant, $z = -0.72$, $p = 0.47$. There was no significant effect of Shift Direction on non-target fixations. However, speakers did fixate more on non-target items in the Contrast-Singleton level (mean = 6.79) compared to the Singleton-Contrast level (mean = 6.04), $z = 1.68$, $p = 0.09$. Figure 13 displays the number of fixations across both Shift Direction and Training-Test Consistency factors. Table 18 displays the mean number of non-target fixations broken down by Shift Direction and Training-Test Consistency. Finally, there was no significant interaction between Shift Direction and Training-Test Consistency $z = -0.52$, $p = 0.6$.

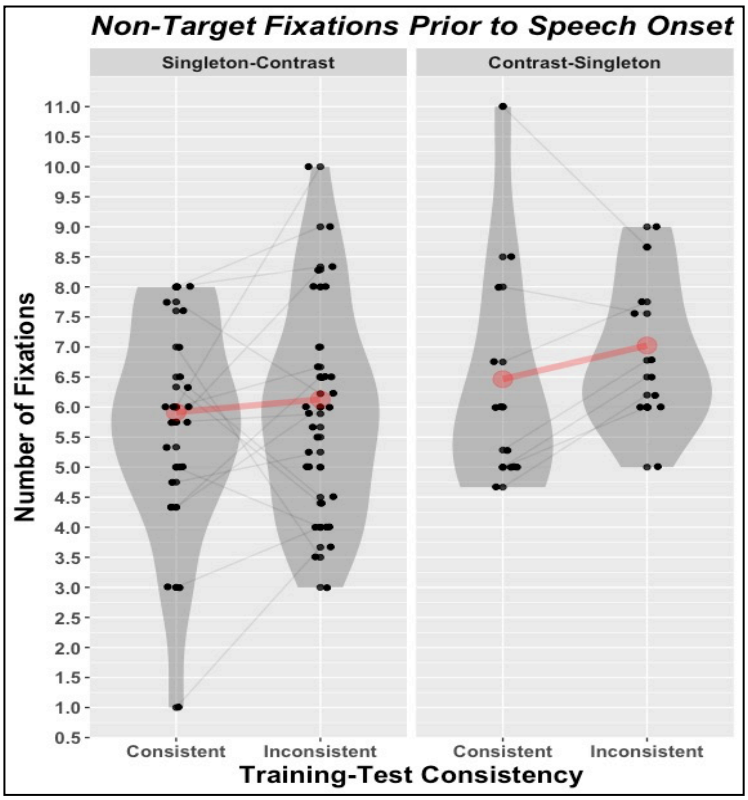


Figure 13: Non-target fixations prior to speech onset in both the Shift Direction and Training-Test Consistency factors. Note that each grey line represents a single participant. The red circles represent the grand means across each level of the Shift Direction and Training-Test Consistency factors.

Shift Direction	Training-Test Consistency	Mean Number of Non-Target Fixations
Singleton-Contrast	Consistent	5.91
Singleton-Contrast	Inconsistent	6.13
Contrast-Singleton	Consistent	6.46
Contrast-Singleton	Inconsistent	7.03

Table 18: Mean number of fixations by Shift Direction and Training-Test Consistency.

4.5 – Discussion

Experiment 2 was our second attempt to test the *retrieval fluency hypothesis*. This hypothesis proposes that attending to an object with the goal of referential encoding elicits retrieval of previous referential expressions used for that particular referent. Accordingly, speakers use the strength/fluency of these memory signals as a heuristic for audience design in referential communication. Our previous results from Experiment 1 did not reveal a main effect of fluency on misspecification rate and therefore failed to find evidence supporting the retrieval fluency hypothesis. Experiment 2 marked an improved attempt to test for the retrieval fluency effect.

In this experiment we made a number of important alterations in an attempt to increase the level of retrieval fluency experienced by participants. In particular, we made two significant modifications to the design – firstly we altered the stimuli set presented to participants. Instead of the letters (of varying colours and size) that were presented in Experiment 1, we introduced a new range of objects with more distinguishable features. The main idea behind this alteration was that these objects may have enabled participants to build up stronger, more fluent memories of the descriptions used with each target item. The second change was in relation to the presentation and sequencing of trials. In this experiment we replaced the Context Variability factor (Low vs. High Variability) from Experiment 1 with the Training-Test Consistency factor (Training Consistent vs. Inconsistent). This new factor reconfigured the arrangement of trials in the training phase of the experiment. In contrast to Experiment 1, the training trials remained relatively stable in presentation and we altered whether the training arrangement was similar (Consistent) or dissimilar (Inconsistent) during the test phase.

The results of Experiment 2 provided weak statistical support for the *retrieval fluency hypothesis*. There was a significant main effect of Training-Test Consistency on misspecification rate with participants misspecifying more frequently in the Consistent level (85%) compared to the Inconsistent level (80%). Participants struggled to adapt their descriptions to suit the conversational context when the training phase was contextually consistent with the test phase and often used the same description as before even though it was no longer contextually appropriate. This suggests that participants used a retrieval fluency heuristic and relied on their memory of their previous utterance to generate descriptions for the addressee.

The eye tracking analysis of non-misspecified test trials revealed that speakers made fewer non-target fixations in the Training-Test Consistent level (mean = 6.09) compared to the Training-Test Inconsistent level (mean = 6.43). However, this difference was not significant. We also found that participants made more non-target fixations in the Contrast-Singleton level (mean = 6.79) compared to the Singleton-Contrast level (mean = 6.04) of the Shift Direction factor. However, unlike Experiment 1, there was no statistically significant difference between these two levels. As mentioned in Chapter 4, it is likely that this pattern of results reflects the fact that the training trials had effectively primed participants to check the context for a competitor letter in the Contrast-Singleton level, unlike the training trials in the Singleton-Contrast level where there was no competitor present.

Notably, the misspecification rate in this experiment was considerably higher than the rate of misspecification in Experiment 1 (Low Context Variability, 17%; High Context Variability, 16%). Further analysis revealed that participants were more likely to overspecify than underspecify referents. Participants overspecified at a rate of 89% in the Contrast-Singleton level of the Shift Direction factor compared to a 74% rate of underspecification in the Singleton-Contrast level. Although we predicted the opposite result (based on our findings from Experiment 1) it was not surprising that we found this effect since overspecification is common in referential communication (Deutsch & Pechmann, 1982; Engelhardt et al., 2006; Gann & Barr, 2014; Horton & Keysar, 1996; Nadig & Sedivy, 2002). This result can perhaps be explained by the notion that when speakers misspecify utterances they usually prefer to overspecify than run the risk of underspecifying descriptions to listeners. Providing too little information may be considered more communicatively costly as it requires addressees to guess at the speaker's meaning – and potentially causes more confusion and misunderstanding for the listener (Gann & Barr, 2014).

Our second main prediction for this experiment focussed on the differential speech onset latency. We predicted that there would be a main effect of Training-Test Consistency on speech onset latency with participants experiencing more difficulty shifting from their entrained description to a more appropriate description in the Consistent Training-Test level. However, we obtained a significant effect in the opposite direction. Mean onset latency in the Inconsistent level was greater (948.2ms) than in the Consistent level (672.2ms). Although we obtained an unexpected effect in the Training-Test Consistency factor our results could nevertheless still be interpreted as support for the *retrieval fluency hypothesis* – on being presented with test trials which were inconsistent with the arrangement shown during the training phase participants experienced less fluent retrieval (weaker memory signals) of their previous description and therefore took longer to adapt their description to suit the conversational context. Whilst this is entirely possible, we are tentative about this result and would be cautious about interpreting this finding in such a way. Instead we acknowledge that this result perhaps reveals a flaw in our prediction as a significant effect in either direction could be interpreted as support for the fluency hypothesis.

Analysis also revealed an effect of Shift Direction on differential onset latency with participants taking significantly longer to provide descriptions in the test phase in the Singleton-Contrast level (1032ms) in comparison to the Contrast-Singleton level (420.5ms). This result reflects the likelihood that participants took longer to think carefully

and adapt their description (by adding a modifier to their speech) in the Singleton-Contrast level. In contrast, participants were quicker in the Contrast-Singleton level and gave less consideration to the content of their description. As mentioned previously, this is reflected in the higher rate of misspecification in the Contrast-Singleton level (89%) compared to the rate of misspecification in the Singleton-Contrast level (74%). Notably, we found no effect of Training-Test Consistency on speech fluency. However, there was a significant effect of Shift Direction on speech fluency. Participants were significantly more fluent (FL) in the Singleton-Contrast level (9%) in comparison to the Contrast-Singleton level (6%). This is in contrast to Experiment 1 where speech fluency was higher in the Contrast-Singleton level (95%) compared to Singleton-Contrast (91%).

Overall, these results provide weak evidence supporting the retrieval fluency hypothesis for audience design. However, due the low effect size of our main effect of Training-Test Consistency on misspecification (Consistent level 85% vs. Inconsistent level 80%) we were motivated to carry out an additional experiment which sought to further test the retrieval fluency hypothesis.

In our third study, we attempted to advance our current findings and develop our paradigm to reflect a more naturalistic conversational setting. To reach this goal, Experiment 3 was designed with a new completely new format. Specifically, we introduced an additional Matcher to the experimental design. This enabled us to introduce two new experimental factors (Pragmatic Consistency and Visual Consistency) to test how factors in the speaker's communicative environment affect how the speaker linguistically encodes referential information. Whilst the results of Experiment 2 offered some support for our hypothesis we acknowledge that the Training-Test Consistency factor in this study lacks communicative relevance in the context of a normal day-to-day interaction.

In Experiment 3 we attempted to address this issue by manipulating a more relevant cue - the appearance of the conversational partner that the Director spoke to. Thus in our final experiment we manipulated the visual consistency of the addressee in a further attempt to influence the level of retrieval fluency that the speaker experienced whilst providing referential descriptions. Importantly, by incorporating this manipulation into our design, Experiment 3 enabled us to further test the concept of partner specificity (the proposal that conversational partners can act as contextual cues for memory in common ground) advocated by Horton (2007) and Horton and Gerrig (2005a). The following chapter details the methodology and results of our final retrieval fluency experiment.

Chapter 5 – Experiment 3

5.1 – Background

5.1.1 – Does the Conversational Environment Affect Referential Encoding?

The results of Experiments 1 and 2 failed to provide compelling evidence indicating that the episodic effects of memory influence the speaker's production of referential descriptions during audience design. Although we did obtain a significant main effect of retrieval fluency on misspecification rate in Experiment 2, the effect size for this result was small. Our efforts to test the retrieval fluency hypothesis have thus far focussed on manipulating communicatively irrelevant cues in the speaker's environment (e.g. similarity between contexts in training vs. test phase, colour and position of letters in an array). We acknowledge that the manipulations implemented in our first two experiments lack communicative relevance in the context of a normal everyday interaction between two interlocutors. In this study, we sought to address this issue by manipulating a referential cue that is normally strongly correlated with common ground – the speaker's *perceptual experience* when addressing a listener.

Who the speaker is looking at during conversation can be considered to be an influential cue in helping the speaker to generate referential descriptions. As noted in Chapter 1, a key component of Horton and Gerrig's (2005a) memory-based model is the argument that conversational partners act as contextual cues for the automatic retrieval of information (Horton, 2007). This idea is supported by evidence indicating that common ground established with a specific partner can be considered in the early stages of language processing (e.g. Hanna & Tanenhaus, 2004; Metzing & Brennan, 2003). If the memory-based approach is correct, it implies that perceptual experiences can serve as a proxy for common ground. Thus the visual appearance of the addressee should act as a cue for partner specificity. However, in everyday conversation, pragmatic knowledge of the identity of the intended addressee is almost perfectly correlated with the perceptual experience of seeing the person one is speaking to. Studies which support the supposition of partner specificity in audience design (e.g. Gorman et al., 2013; Hanna et al., 2003; Horton, 2007) often fail to take this into account. Thus in our final experiment we attempted to de-confound the visual appearance of a potential addressee from the pragmatic knowledge of who the speaker was interacting with. In this way, our experiment

enabled us to further test Horton and Gerrig's (2005a) assumption of partner specificity in audience design.

Our design for this study was influenced by a gesture production experiment by Mol, Krahmer, Maes, and Swerts' (2011) that explored the idea of de-confounding the effects of seeing from being seen using webcam technology which simulated eye contact between communicators. Interestingly, Mol et al. (2011) found that speakers produced more gestures only when they knew that they were visible to the addressee. Recent research by Barr et al., (2014) also implemented a similar de-confounding technique. Taking advantage of the naturally occurring common ground that exits between university students, the authors recruited pairs of friends to play in a referential communication game along with a lab assistant. In this experiment one of the friends heard the name of a mutually known person and had to click on the corresponding photograph that appeared on a computer monitor. Crucially, on some trials Barr et al. (2014) de-coupled the voice that read out the name of the mutually known person from the actual designer of the message for the addressee. Results revealed that addressees looked at the target picture more quickly and reliably when the name of the target person was read out by the addressee's friend. This was irrespective of whether the name was selected by the friend or the lab assistant (Barr, 2014; Barr et al., 2014).

5.1.2 – Adapting the Retrieval Fluency Experiment

As with previous experiments in this line of research (e.g. Gann & Barr, 2014) the study was designed to enable the Director to build up experience describing a certain set of objects during a 'training' phase with one of the two Matchers. At a later test phase we assessed whether Directors drew upon this experience when describing the target object (depending on whether they spoke to the same or a different addressee and additionally, in the current case, who they saw on screen). As with Experiments 1 and 2, we aimed to assess the degree to which Directors rely on memory when providing descriptions by examining *referential misspecifications* during the test phase (i.e., whether Directors provided more or less information than is optimal for identifying the target within the current referential array). The extent to which Directors misspecify referents in the test phase indexes the degree to which they *are relying on remembered expressions* from the training phase rather than tailoring their expressions to the current visually available set of objects.

Similarly to Experiments 1 and 2, trials were organised into a series of blocks which were then each divided, in turn, into a "training" and "test" phase (this division into phases was

not readily apparent to participants). The purpose of the training phase was for Directors to entrain on particular referential expressions with a given addressee for particular targets in specific referential arrays. During the training phase in this study Directors always saw (through the webcam link) the same Matcher with whom that they were entraining on descriptions; the other Matcher was off-screen. The off-screen Matcher was not able to hear the Director's descriptions nor see the array of objects. This Matcher heard white noise in his/her headphones and wore a blindfold (see Section *5.2.7 – Procedure* for more details). In the test phase the same targets appeared in contexts requiring different descriptions enabling us to measure speakers' referential misspecifications.

Experiment 3 was designed as an interactive referential communication game with the participant playing the role of the Director (the speaker) and the experimenter and a lab assistant playing the role of Matcher 1 and Matcher 2 who interpreted the Director's descriptions. Building on the design from Experiments 1 and 2, we sought to dissociate perceptual cues (the visual image of a listener) from pragmatic cues (knowledge of the identity of the actual listener). Inspired by Mol et al. (2011) and Barr et al. (2014) we de-confounded pragmatic (Pragmatically Consistent vs. Inconsistent) and perceptual (Visually Consistent vs. Inconsistent) cues using a webcam communication setup where the visual experience of the Director was controlled independently of the pragmatic situation, so that who the Director saw and who the Director was speaking to did not always coincide. The Director sat in a separate testing room from both Matchers and viewed a separate computer monitor from the Matchers throughout the experiment. In the other room, both Matchers were seated next to each other and shared the same computer monitor.

The experimental setup de-confounded perceptual and pragmatic cues as follows. First, to control who could hear the Director's descriptions both Matchers wore headphones. The audio was configured so that only one of the two Matchers could hear the Director at a given time (please see Figure 14 for an overview of our design). The Director controlled which of the two Matchers was the addressee by manipulating an audio mixing board. Second, the Director was able to see into the Matchers' room through a webcam (but not vice versa). Independent to the audio manipulation, at any given time, only one of the two Matchers was on-screen; this on-screen matcher may or may not have been the intended addressee. In other words, Directors were occasionally confronted with a situation in which they were *seeing someone other than the person they were speaking to*.

We incorporated three main factors in this design:

Visual Consistency Factor

Throughout the experiment the participant viewed the stimuli overlain on a live webcam image of one of the two Matchers (see examples in Figures 14 and 17). The factor of Visual Consistency refers to whether or not the Matcher the participant saw at the test phase was the same one (Consistent) they saw at training; or a different one (Inconsistent). If Horton (2007) and Horton and Gerrig (2005a) are correct in their assumption that speakers use conversational partners as memory cues, then we would expect participants to misspecify at the test phase (due to greater retrieval fluency) in this factor, when looking at the same Matcher they described the target item to during the training phase.

Pragmatic Consistency Factor

This factor refers to whether the intended addressee at test is the same (Consistent) or different (Inconsistent) from the intended addressee of the training phase. Similarly to the Visual Consistency Factor, if speakers use addressees as memory cues for conversation then we would expect greater misspecification when participants speak to the same Matcher during the test phase as they spoke to during training. Note that this factor was manipulated *completely independently* from Visual Consistency.

Shift Direction Factor

As with our first two experiments, this factor was included to vary the amount of information that Directors would have to provide in test trials relative to training. This variation was to prevent a situation in which Directors would learn that they need to alter the information at test in only one direction (e.g., always increase rather than reduce information). As such, in this experiment it was not a factor of primary theoretical interest. In the Singleton-Contrast level speakers entrained upon descriptions for a target object in a context where modifiers were not required *("the car"*) and were then tested in a context requiring a modifier *("the family car"*). In the Contrast-Singleton level this order was reversed: the speaker was shown a competitor object as well as the target item during training (e.g., car vs. sports car), leading speakers to entrain upon a modified expression during training. At test, the competitor was then replaced with a foil item, such that participants were able to simplify their description of the target item ("*the family car" -> "the car*"). Please see the previous description of this factor in Chapter 4 for further details.

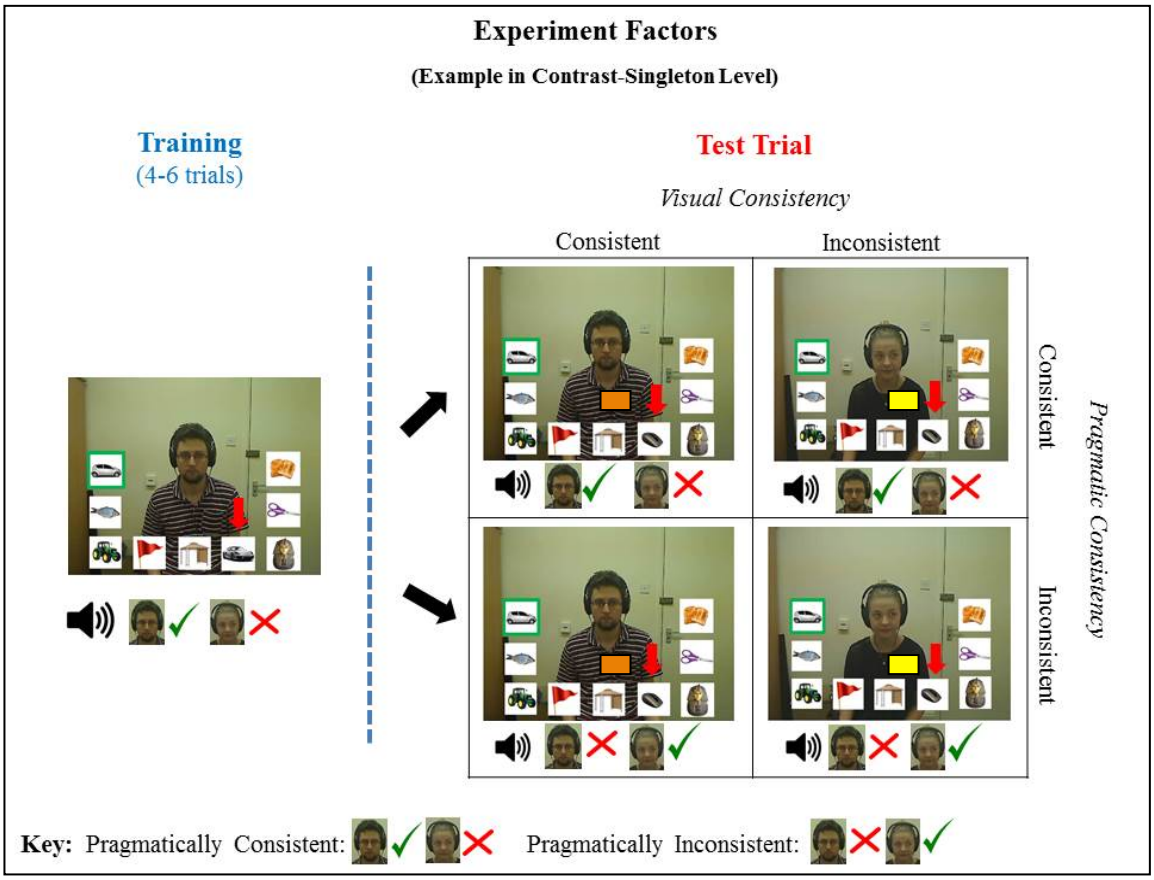


Figure 14: Overview of the Visual Consistency and Pragmatic Consistency factors

Under the Test Trial column above, we outline the different levels of the *Visual Consistency* and *Pragmatic Consistency* factors. During the training phase the Director (participant) always described the target item (e.g. *"the **car**"*) to the Matcher (yellow vs. orange) that he/she could see on the computer monitor while the off-screen Matcher wore a blindfold and listened to white noise. At the test trial we manipulated the Visual Consistency for the Director such that the participant either viewed the same Matcher as they saw during the training phase (Consistent) or the alternative Matcher who was off-screen during the training phase (Inconsistent). Additionally, we also manipulated the *Pragmatic Consistency* of the Matcher during the test trial: the Director either continued to describe the target item to the same Matcher as before (Consistent) or described the item to the alternative matcher (Inconsistent). Note that the red arrows highlighting the foil/competitor objects are for illustrative purposes and did not appear during the actual experiment. Similarly, the key shown for the Pragmatic Consistency factor in this figure is for illustrative purposes only.

An important aspect of this experiment was that it addressed the concerns raised by Brown-Schmidt et al. (2015) who highlighted that studies failing to find evidence in support of partner specificity in audience design are often characterised by a lack of partner interaction in their design (e.g. Barr & Keysar, 2002; Brown & Dell, 1987). The authors argue that experiments that have more extensive interactions between participants show greater partner specific effects (e.g. Lockridge & Brennan 2002; Hanna et al., 2003). Further, research suggests that when participants are unable to interact with their partner this results in a lack of partner specific bindings being formed in memory (Brown-Schmidt, 2009). To ensure that we accounted for these findings, live interaction with the addressee was a key feature of our experiment. Directors had the opportunity to engage directly with both Matcher 1 and Matcher 2 on different occasions throughout the experiment.

5.1.3 – Describing Unconventional vs. Conventional Referents

One potential issue with our design concerned the manipulation of Visual and Pragmatic Consistency. We acknowledge that it was unusual to decouple the speaker's pragmatic knowledge from their visual experience. A possible outcome from this decoupling was that we would see no evidence for an effect of Pragmatic Consistency on misspecification rate. Should this arise, a concern might be that perhaps speakers were simply inattentive to the identity of the current intended addressee (the Matcher who could hear the speech through the headphones). To check that speakers were indeed sensitive, along with our main target/competitor stimuli we included a set of unconventional filler trials for which, based on Gann and Barr (2014), we expected to see strong effects of Pragmatic Consistency.

These unconventional fillers included abstract drawings in grayscale that Directors would need to describe using complex descriptions (please see Figure 15 for an example of the stimuli). Since speakers lacked any experience describing these objects they would have to come up with their own descriptions which they could eventually shorten over time (Gann & Barr, 2014). A crucial test of whether speakers kept track of who they were interacting with on test trials was whether they continued to use a shortened description for an abstract object when talking to a new addressee who had never heard the description before. Following the methods of Gann and Barr, we measured description length in terms of number of words used to describe targets. Our hypothesis was that for a given target we would see a *greater increase in description length* from the last training trial to the test trial when the test addressee was not the same as the training addressee. Thus we expected to replicate the results of Gann and Barr (2014).

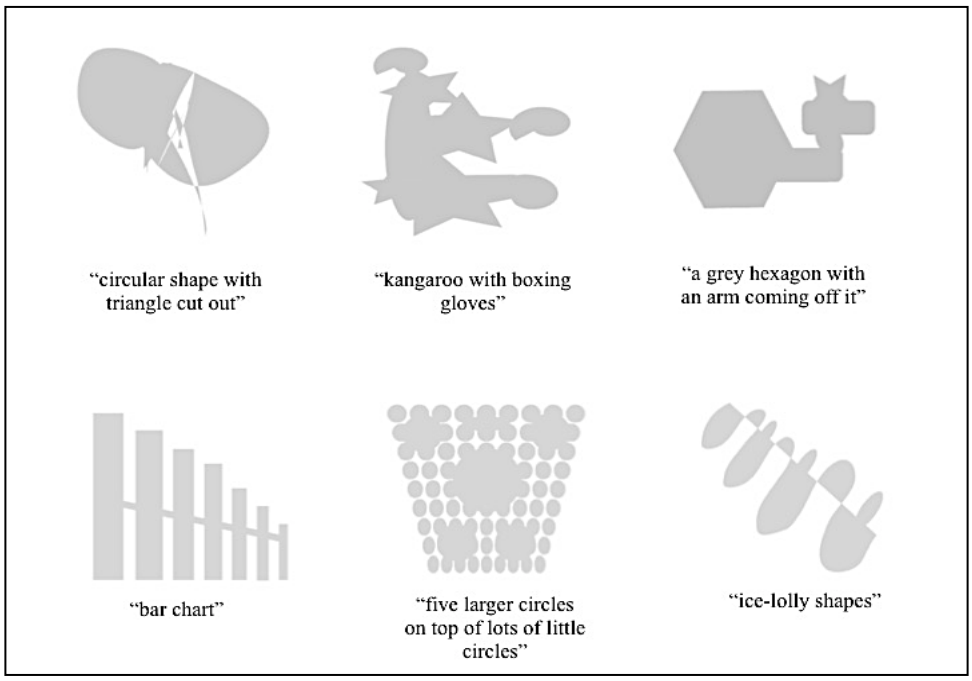


Figure 15: Example of six unconventional target items and the descriptions used by participants.

5.1.4 – Pre-registered Predictions

All our predictions were pre-registered on the OSF (outlined in section *5.3.4 – Pre-registered Analysis and Predictions*). Our main prediction concerned the Visual Consistency factor. We predicted that speakers would misspecify referents at a higher rate in the Visually Consistent level than in the Visually Inconsistent level.

5.2 – Method

5.2.1 – Participants

Forty subjects completed the experiment (31 Females, M=25.6 years). All subjects were recruited from the campus at the University of Glasgow. All subjects were Native English speakers. Subjects who were bilingual identified English as their first language. Participants were paid £6 for taking part in the study. One participant was replaced due to the use of ineffective descriptions during the task (continuously failing to adapt their utterances for the listener, please see Section *5.3.3 – Exclusion Criteria for Participant Responses* in Chapter 4 for more details). Subjects gave written informed consent before beginning the experiment and were fully debriefed after the experiment had finished. Our procedures fully complied with the ethical code of conduct of the British Psychological Association.

5.2.2 – Norming of Test Items

The target and competitor items originally used in Experiment 2 were used in this experiment. These items were previously normed by 68 Native English speaking volunteers using the web-based surveyor SurveyMonkey (please see the pre-registration for Experiment 2 for details: https://osf.io/uq4k7/). Based on performance in Experiment 2, nine items were replaced for this experiment (8 items which were over-described in the Singleton-Contrast level, at a rate of more than 50% during the training phase and 1 additional item which was replaced as participants had previously found it difficult to name the target). Nine new stimuli pairs were added to our original list (please see Appendix 6 for a complete list of the Target and Competitor objects used).

5.2.3 – Experimental Setup and Task

In each trial, the Director was tasked with describing a target object to a given Matcher so that the Matcher could then identify this object on his/her own screen. The intended Matcher (the "addressee") then selected the target from an array of objects by pressing a number key. The Director's view showed the target object within a grid containing images of other objects. In this experiment the grid was superimposed over a live webcam image of the Matcher visible behind the object images (see Figures 16 and 17 for examples). The Director was informed that in each trial both Matchers had the same objects on their monitor but that they were arranged in a different format to the grid that appeared on their screen. The Matchers' view consisted of a black background screen with each of the potential target items presented in 3x3 arrangement (see Figure 17 for an example of this layout and the response pad used by both Matchers).

5.2.4 – Design

There were three factors in the design, Direction of Shift (Singleton-Contrast and Contrast-Singleton), Visual Consistency (Consistent and Inconsistent) and Pragmatic Consistency (Consistent and Inconsistent) which formed a full-factorial 2x2x2 within-participant design. As explained previously, it was only the latter two factors (*Visual* and *Pragmatic Consistency*) that were of primary theoretical interest.

5.2.5 – Materials and Sequencing of Trials

We used the same Target and Competitor objects which were normed for Experiment 2 (see Section *5.2.2 – Norming of Test Items* for details on exceptions). Each display consisted of nine images of various objects displayed around the webcam image of the Matcher (see Figure 17 for an example of the layout). The experiment contained 12 blocks

of trials, each consisting of 4-6 training trials for each of four different target picture trials followed by a test phase with single test trials for each of the four targets. We use the term "sequence" to refer to the collection of training and test trials all associated with a single target/competitor/foil triplet. Thus there were 48 sequences, each of which appeared an equal number of times in all eight conditions of the 2x2x2 design, counterbalanced using eight stimulus lists.

For each sequence, the number of training trials was randomly selected, with a range from four to six. The motivation for varying training sequence length was to make the occurrence of the test trial unpredictable. Each experimental session had the same number of four-, five-, or six-length training sequences, and thus had a total of 240 training (=16 x (4 + 5 + 6)) and 48 test trials. As well as incorporating unconventional filler items into our design we also included conventional fillers that had targets much like the main trials. This type of sequence was included so that it was not always the case that the displays within a sequence predictably changed from training to test (i.e., through the substitution of the competitor for the foil or vice versa).

For the conventional fillers twelve sequences were included (one for each block) in which the display was identical from training to test. Six of these included a competitor so that the target must be described using a modifier. Each sequence included three training trials and one test trial, for a total of 48 trials. For the unconventional fillers there were three training trials and one test trial for twelve sequences (one for each block). Half of these test trials were presented when the speaker was talking to the same Matcher as the one they spoke to at the training phase (Pragmatically Consistent) and the other half were presented when the speaker was talking to a different Matcher from the one they spoke to during training (Pragmatically Inconsistent). Similarly, half of the test trials were shown in the Visually Consistent level (with the speaker looking at the same Matcher as the one that appeared during training) and the other half of these test trials were shown in the Visually Inconsistent level.

In sum, in each session there were 240 training trials, 72 filler training trials (36 conventional and 36 unconventional), 48 test trials, and 24 filler test trials (12 conventional and 12 unconventional), for a grand total of 384 trials.

5.2.6 – Apparatus

The experimental stimuli were presented on a 19" LCD Dell desktop computer monitor (4:3 aspect ratio, resolution 1024 x 768 pixels). Participants were seated 45-55cm away

from the monitor. A microphone was placed above the participant's computer monitor to record their descriptions of the *target* object for each trial. The audio was tagged using Audacity 2.0.6 software.

5.2.7 – Procedure

Upon arrival each participant was given an instruction sheet detailing the task and their role during the experiment (see Appendix 7). Both Matchers were set up in an adjoining room to the Director and faced a computer monitor (see Figure 16). The layout of the room was designed to ensure both Matchers were able to move in front of the webcam when prompted to appear on screen. During the experiment each Matcher was referred to by colour (yellow and orange) and both Matchers wore coloured tags to ensure the participant was able to differentiate them from one another. Before the experiment began participants took part in a practice session that consisted of twelve training trials and four test trials. This enabled the participant to familiarise themselves with their role as the Director as well experience the experiment from the Matchers' perspective.

In order to discriminate the target object from the filler objects, the target for a given trial was highlighted within a green square in the director's display (see Figure 16 and Figure 17). The participant was informed that as the arrangement of images within the Matcher's computer screen differed in an unpredictable way from that of the Director, they would have to describe the features of the highlighted item, rather than use the target's on-screen location as a description.

Before each block of training trials was presented, a notice was shown on-screen informing the Director which Matcher appeared on-screen (yellow or orange) and which Matcher was listening to their description (yellow or orange). This order was pre-determined and counterbalanced across participants. The notice also indicated that the off-screen Matcher was to put on the blindfold. The Director manipulated the audio channel using a crossfading slider on a mixing board. The Matcher who was not selected as the intended addressee heard white noise through his/her headphones to ensure that any speech from the Director was indecipherable. Half of the training phases were completed with the yellow Matcher as addressee, and the other half with the orange Matcher as addressee. The Matcher who was off-screen during training was always wearing a blindfold and could only hear white noise through their headphones.

Just prior to the test phase another on-screen notice appeared indicating that the blindfold was to be removed, and designating which Matcher was to appear on-screen and which

Matcher was to hear the Director's speech. The Director then switched the cross fader so that the indicated Matcher was able to hear the audio. Once the arrangements were completed one of the Matchers advanced to the first test trial. To avoid confounding the length of delay between training and test with the experimental factors, the on-screen notice appeared for a minimum of eighteen seconds.

Audio recording of the Director's response began simultaneously with the presentation of the main display. The trial ended when the Matcher listening to the description selected the object designated by the Director. After each individual trial the Matcher listening to the Director's descriptions was prompted to press a keyboard button to continue to the next trial. Note that the Director could not see the Matchers' screen and received no feedback regarding whether the trial was completed correctly. If the Director failed to provide sufficient information to identify the target, the Matcher was instructed to ask the Director for clarification (e.g. "which one do you mean?"). Any such clarification exchanges appeared in the audio recording for the trial and were noted during later transcription.

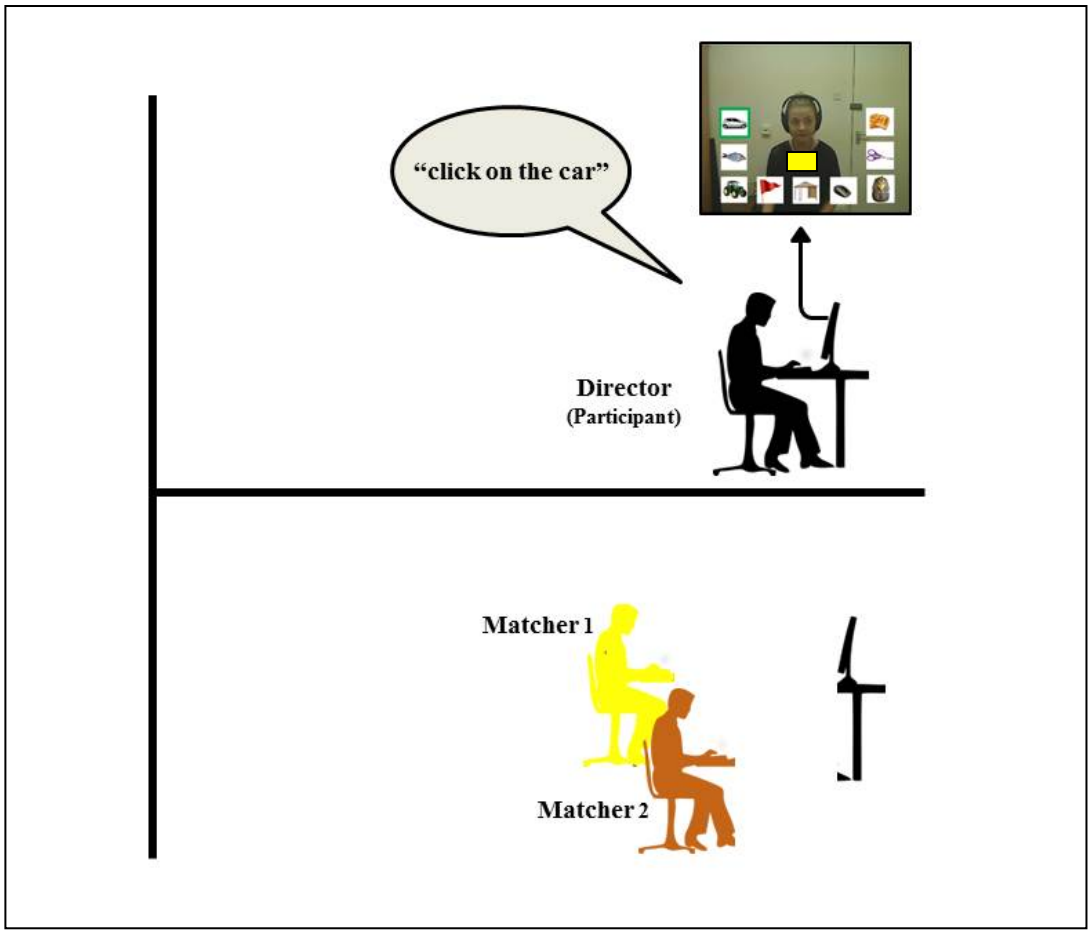


Figure 16: Outline of experimental set up and procedure for Experiment 3.

The Director sat in a separate room from the yellow and orange Matchers and viewed the stimuli on a separate computer monitor. During each block of trials the participant viewed a live webcam image of one of the Matchers in the background of the computer monitor and also communicated with one of the Matchers via a microphone and a set of headphones.

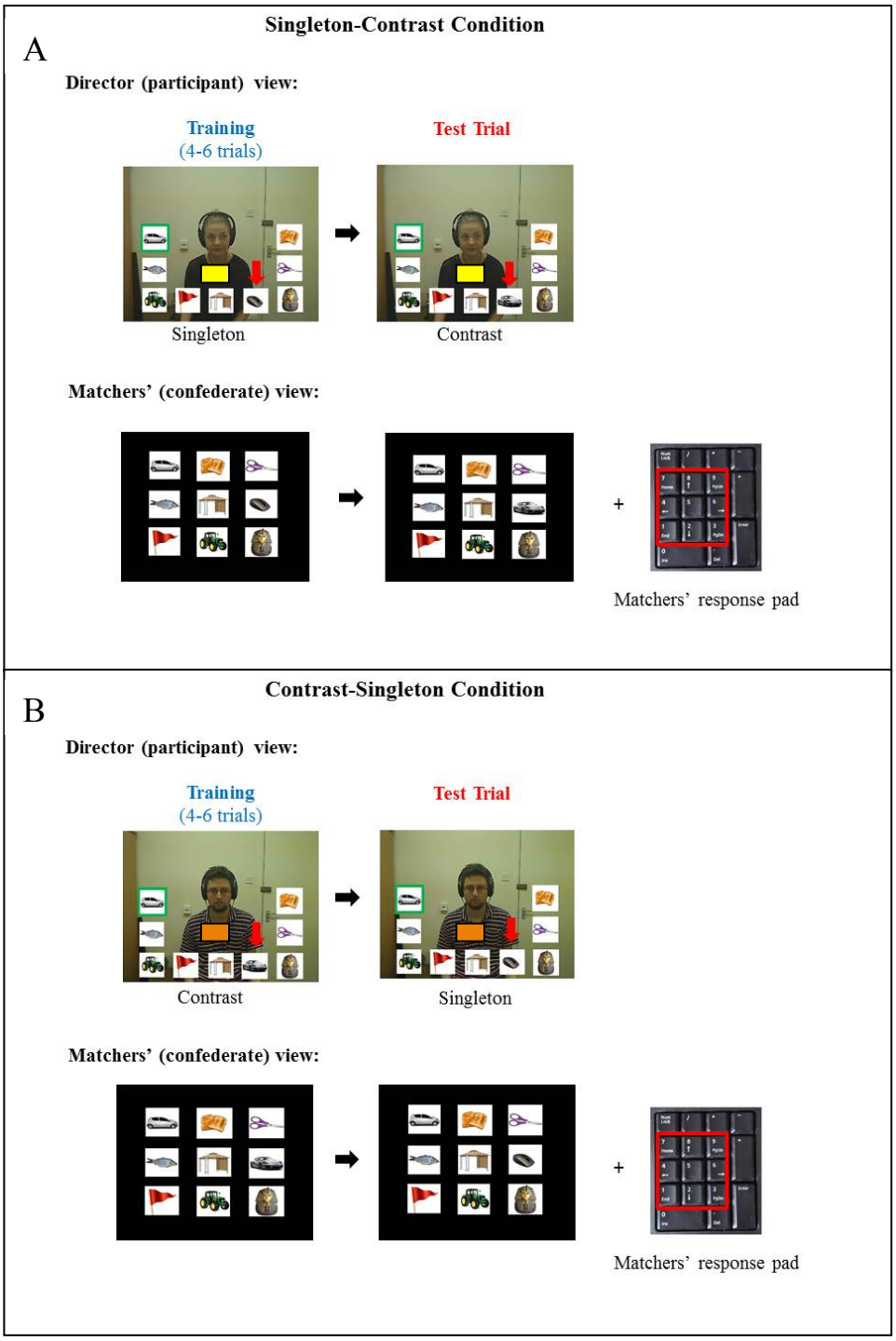


Figure 17: Overview of the training and test trials in the Shift Direction factor.

Panel A shows an example of the stimuli in the Singleton-Contrast level. The top half of the panel displays the Director's (participant) view of the stimuli. The stimuli objects are displayed around the image of the Matcher on the screen. Note that the target object appears in a green rectangle on the Director's screen (*"the **car**"*). After a series of training trials the test trial presents participants with the target object *"the **car**"* again, but unlike the training phase it also introduces a new *"**sports car**"* 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" (highlighted by the red arrow) which acts as a foil for the *"**sports car**"*. The "computer mouse" is replaced by the *"**sports car**"* in the test phase. The bottom of Panel A shows the Matchers' view of the stimuli during the training and test phase. Panel B shows the stimuli in the Contrast-Singleton level with the alternative matcher on-screen. Note that in this case the competitor object – *"the **sports car**"* (highlighted with the red arrow) is also present in the grid during training, while at test it has been replaced with the foil object - a "computer mouse". Please note that the red arrows highlighting the foil/competitor objects are for illustrative purposes and did not appear during the actual experiment.

5.3 – Predictions and Data Analysis

5.3.1 – Main Measurements

Our analysis focussed on two categories of measurements: (1) speech content and performance; in particular, use of a descriptive modifier and speech fluency; (2) differential onset latency, defined as the speech latency for the test trial minus the speech latency for the final training trial in each sequence.

5.3.2 – Transcription and Coding of Audio Files

For each of the 48 sequences for each Director, we transcribed and coded the audio recordings for two trials: (1) the last description of the target in the training phase; and (2) the test trial. The last training trial was needed in order to provide baseline data for the speech onset latency in the test trial, and to verify that speakers were not already misspecifying the referent during training. Each trial was transcribed and coded for fluency and adjective use.

Similarly to Experiment 2 fluency was coded into one of five categories, as shown in the table below:

Speech 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"

Table 19: Outline of speech fluency categories with examples.

We also coded whether or not a descriptive modifier was used, defined by the following categories:

Modifier 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"

Table 20: Outline of item modifier categories with examples.

Onset times of utterances were measured in milliseconds (ms). The following criteria were applied when identifying utterance onsets:

- Trials were discarded if the speech was unidentifiable.
- Any filled pauses or articles were ignored (um, uh, the); speech onset was 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 as the onset of the word *"silver"*).
- If Directors corrected themselves after an error (e.g. *"white car...eh sorry silver car"*) onset of the correction (i.e. *"silver"*) was recorded. However, such repaired utterances were not used in the analysis of speech onset.

* Note that a colour description was not coded as a modifier if it did not distinguish the target object from the competitor (for instance both the family car and the sports car were silver in colour).

5.3.3 – Exclusion Criteria for Participant Responses

We applied the same exclusion criteria for Experiment 3 as we implemented for Experiment 2. See Chapter 4, *Section 4.3.3 – Exclusion Criteria for Participant Responses* for a full outline of the criteria. Based on this criteria one subject and one stimulus pair (target and competitor items) were removed. Please see Appendix 8 for details.

5.3.4 – Pre-registered Analysis and Predictions

We pre-registered our analysis and predictions on the Open Science Framework. The basis for our estimate of a sample size of 40 participants and 48 items was derived from our pilot study which gave power of 85% for 36 participants and 48 items. This pilot study is available on github (https://github.com/dalejbarr/EESP2) and our pre-registration for Experiment 3 can be found on the Open Science Framework (OSF: https://osf.io/5yz3n/). As there were eight stimulus lists, the number of participants had to be a multiple of eight, and therefore we opted to move up to 40 participants. Given that for the current experiment, we improved our stimulus materials and used a more communicatively relevant memory cue (an image of the addressee, as opposed to the configuration of objects in the display), we assumed that the 85% estimate was a lower bound. Please see section *3.3.4 - Pre-registered Analysis and Predictions* outlined in Chapter 3 for details about the power calculation.

We made the following main predictions:

1) *Main effect of Visual Consistency:* It was predicted that speakers would misspecify targets at a higher rate in the Visually Consistent level (i.e., when looking at the same Matcher at test as during training) than in the Visually Inconsistent level (i.e., when looking at a different Matcher). This test was pre-registered as one-tailed, with alpha = .05, we assumed a lower bound of power of 85%. This prediction was of key theoretical interest, as it is directly related to the "retrieval fluency" hypothesis explored in Experiments 1 and 2.

2) *Main effect of Pragmatic Consistency*: It was predicted that speakers would misspecify targets at a higher rate in the Pragmatically Consistent level (i.e., when speaking to the same Matcher at test as during training) than in the Pragmatically Inconsistent level (i.e., when speaking to a different Matcher). This test was pre-registered as one-tailed, alpha = .05.

3) *Larger effect of Pragmatic Consistency than Visual Consistency:* We included this final prediction as it would enable us to determine whether speakers weigh pragmatic consistency differently from visual consistency. Should we see main effects of Visual and Pragmatic Consistency we would opt to use the *glht* function from the R package *multcomp* to test the null hypothesis that the two effects are equivalent (two-tailed, alpha = .05).

4) *Main effect of Pragmatic Inconsistency on unconventional items*: for this analysis (involving description of abstract objects), we predicted an interaction between Phase (training, test) and Pragmatic Consistency on description length, such that the effect of Phase would be larger in the Inconsistent level. We assessed this directional prediction for the interaction term using a one-tailed test with alpha = .05.

Finally, for an additional analysis, we also tested the three main predictions above for a different dependent variable: differential onset latency. As with our first two experiments, differential onset latency was defined as the time taken to produce the first content word as measured from the onset of the display. Our prediction was that in cases where speakers appropriately specify targets at test, they would experience more difficulty and thus exhibit longer speech onset latencies in the Visually Consistent level than in the Visually Inconsistent level. Likewise, we expected a similar pattern for the Pragmatic Consistency factor. We tested these hypotheses using a one-tailed test (alpha = .05). This analysis only included trials where the target was appropriately specified both at the test trial as well as

in the last training trial before test. The dependent variable was 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.

5.4 – Results

5.4.1 – Statistical Analysis

The statistical analysis for the production data (modifier use and speech onset) was performed using linear mixed-effects models with Directors (subjects) and sequence (item) as crossed random factors (Baayen, et al., 2008). All analyses attempted to use the maximal random effects structure justified by the design (Barr, et al., 2013), which implies by-subject and by-item random intercepts and by-subject and by-item random slopes for all three factors (Pragmatic Consistency, Visual Consistency, and Shift Direction) and their interactions. We derived *p*-values using the t-to-z heuristic (i.e., deriving p-values from the standard normal distribution for the t statistic), as this enabled us to perform one-tailed tests. Models were estimated using the lme4 package in R (version 1.1-7 or higher). All independent variables were deviation coded. The analysis of modifier use assumed a logit link and binomial variance function, whereas the analysis of onset times used an identity link with a Gaussian variance function.

5.4.2 – Misspecification Rate

Analysis of the misspecification data revealed no main effect of Visual Consistency. We found that manipulating whether the speaker was looking at the same Matcher at the test phase as they saw during training did not have a significant effect on misspecification rate. Indeed, the misspecification rate was in the opposite direction of our original prediction. Misspecification in the Visually Inconsistent level (looking at the alternative Matcher) was at 63% in comparison to the Visually Consistent level (looking at the same Matcher) which was 57%, pre-registered one-tailed test, $z = -2.69$, $p = 0.99$ Figure 18 shows the breakdown of misspecification rate (%) by Shift Direction for both Visual and Pragmatic Consistency. Our analysis did not reveal a significant effect of Pragmatic Consistency on misspecification rate. In this factor the rate of misspecification was the same in the Pragmatically Consistent level (60%) as it was in the Pragmatically Inconsistent level (60%), pre-registered one-tailed test, $z = -0.62$, $p = 0.27$. There was no significant interaction between Visual and Pragmatic Consistency, $z = 0.29$, $p = 0.77$.

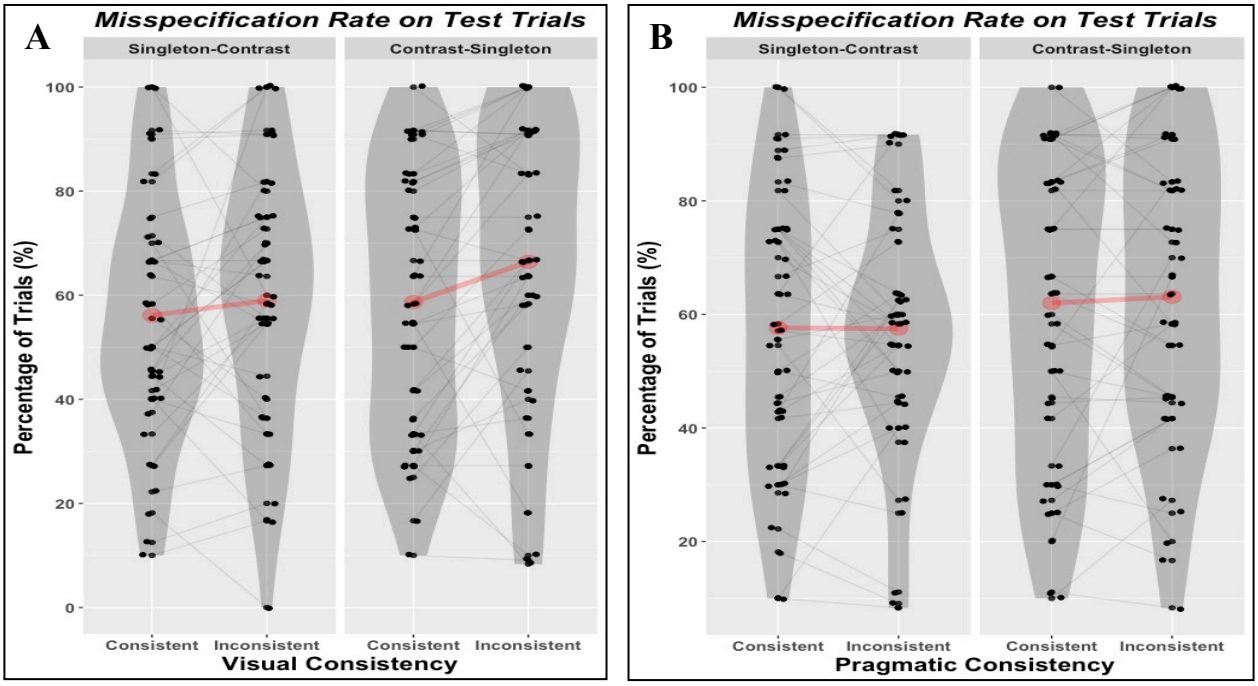


Figure 18: Panel A displays the percentage of fluent trials (%) by Shift Direction and Visual Consistency factors. Panel B shows the percentage of fluent trials (%) by Shift Direction and Pragmatic Consistency. Note that each grey line represents a single participant. The red circles represent the grand means across each level of the Shift Direction and Visual/Pragmatic Consistency factors.

Visual Consistency	Pragmatic Consistency	Shift Direction	Misspecification Rate (%)
Consistent	Consistent	Singleton-Contrast	56.2
Consistent	Consistent	Contrast-Singleton	58.7
Consistent	Inconsistent	Singleton-Contrast	56.1
Consistent	Inconsistent	Contrast-Singleton	58.7
Inconsistent	Inconsistent	Singleton-Contrast	58.9
Inconsistent	Inconsistent	Contrast-Singleton	67.6
Inconsistent	Consistent	Singleton-Contrast	59.1
Inconsistent	Consistent	Contrast-Singleton	65.3

Table 21: Misspecification rate (%) across Visual Consistency, Pragmatic Consistency and Shift Direction factors.

Shift Direction	Modifier Code	Misspecification Rate (%)
Singleton-Contrast	Addition due to Other-repair	58.2
Singleton-Contrast	Addition due to Other/Self	4.2
Singleton-Contrast	Pre-Nominal Modifier	0.2
Singleton-Contrast	No Modifier	4.2
Singleton-Contrast	Deleted Adjective	0.2
Contrast-Singleton	Addition due to Self-repair	0.4
Contrast-Singleton	Post-Nominal Modifier	16.7
Contrast-Singleton	Pre-Nominal Modifier	76.0
Contrast-Singleton	Pre/Post-Nominal Modifier	0.7
Contrast-Singleton	Pre-Nominal/No modifier	0.5
Contrast-Singleton	No Modifier	2.4
Contrast-Singleton	Deleted Adjective	3.2

Table 22: Misspecification rate (%) by Shift Direction and type of modifier.

Table 21 shows the misspecification rate (%) across all three main factors (Visual Consistency, Pragmatic Consistency and Shift Direction). Table 22 shows the misspecification rate (%) by Shift Direction and type of modifier. Results revealed no significant three-way interaction between Shift Direction x Visual Consistency x Pragmatic Consistency, $z = 0.33$, $p = 0.74$. Although Shift Direction was not of primary theoretical interest for this study it is worth noting that participants misspecified more frequently at the Contrast-Singleton level (62.6%) compared to the Singleton-Contrast level (57.6%). However, there was no significant effect of this factor on misspecification rate, $z = 1.12$, $p = 0.26$. Additionally, Shift Direction did not interact significantly with either of the Visual Consistency ($z = -1.03$, $p = 0.3$) or Pragmatic Consistency ($z = -0.65$, $p = 0.52$) factors.

5.4.3 – Differential Speech Onset Latency

Our prediction was that participants would produce longer onset latencies (an indication of greater difficulty altering the content of their description) in both the Visually Consistent level and the Pragmatically Consistent levels. Analysis of the differential onset latency (mean test trial onset – mean onset of final training trial) revealed no main effect of Visual Consistency, one-tailed test, $t = 0.68$, $p = 0.25$. Participants showed similar onset times for both the Visually Consistent level (average 897.1ms) and the Visually Inconsistent level (average 912.1ms). Analysis also revealed no significant effect of Pragmatic Consistency on differential onset latency, one-tailed test, $t = -0.29$, $p = 0.39$. Mean onset for Pragmatically Consistent level was 881.8ms compared to 926.3ms for the Pragmatically Inconsistent level. There was no significant interaction between Visual and Pragmatic Consistency, $t = -0.48$, $p = 0.63$. Table 23 provided a breakdown of the mean onset change for each level of the Visual and Pragmatic Consistency combinations. Figure 19 shows the differential onset latency broken down by Shift Direction for both the Visual and Pragmatic Consistency factors.

No. Trials	Visual Consistency	Pragmatic Consistency	Training Onset (ms)	Test Onset (ms)	Mean Onset Change (ms)
182	Consistent	Consistent	1406.2	2257.7	863.2
186	Consistent	Inconsistent	1418.5	2355.3	930.3
161	Inconsistent	Consistent	1408.8	2310.7	902.5
161	Inconsistent	Inconsistent	1472.1	2393.0	921.8

Table 23: Mean onset change across Visual and Pragmatic Consistency factors.

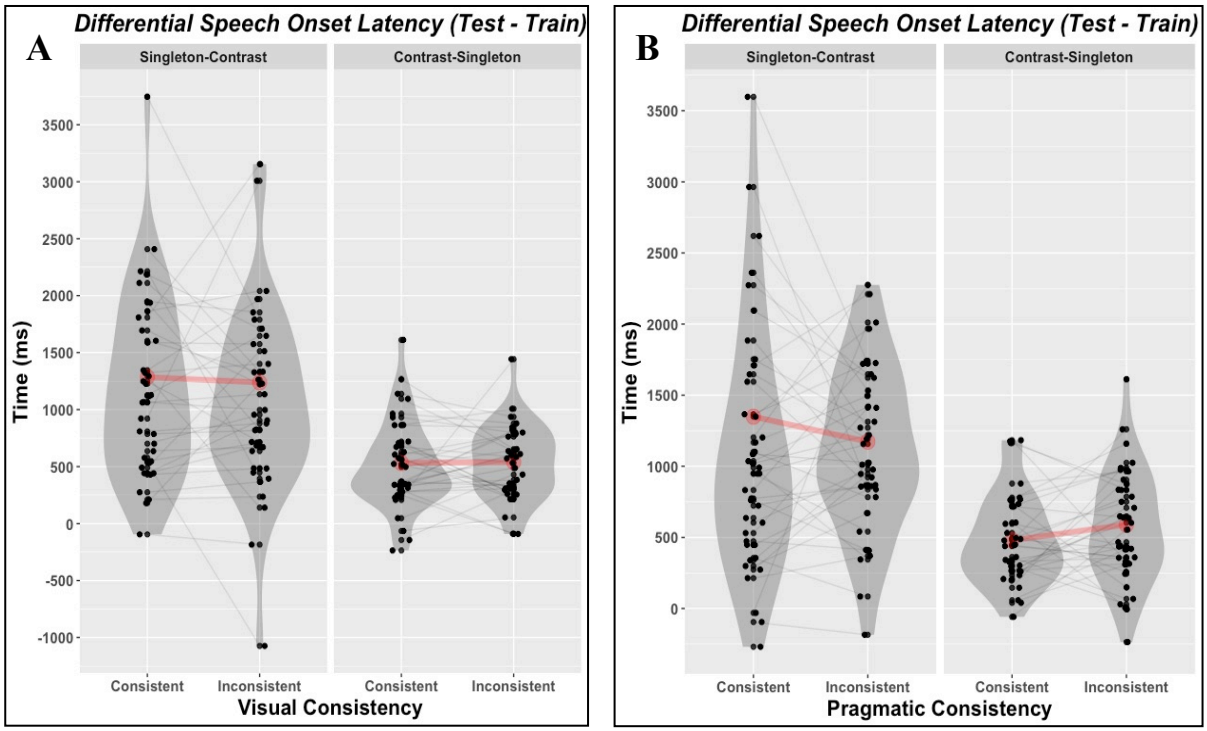


Figure 19: Panel A displays the differential speech onset latency (ms) for the by Shift Direction and Visual Consistency factors. Panel B show the differential latency for Shift Direction and Pragmatic Consistency factors. Note that each grey line represents a single participant. The red circles represent the grand means across each level of the Shift Direction and Visual/Pragmatic Consistency factors

Although not of primary concern in relation to our main predictions, we did find a significant effect of Shift Direction on onset latency, t = -6.08, p = <0.01. Participants took longer in the Singleton-Contrast level (1262.6ms) compared to the Contrast-Singleton level (535.9ms) to produce a relevant description for the listener. Furthermore, there were no significant interactions between Shift Direction and Visual Consistency (t = -0.51, p = 0.61) or Shift Direction and Pragmatic Consistency (t = 1.61, p = 0.11). Finally, there was no significant three-way interaction (Shift Direction x Visual Consistency x Pragmatic Consistency) on onset latency, t = -0.53, p = 0.59.

5.4.4 – Unconventional Referents Analysis

This manipulation involved trials where participants were prompted to describe unconventional, abstract objects. As mentioned previously, one possible outcome for the misspecification analysis was that we would see no evidence for an effect of Pragmatic Consistency on misspecification rate. Since it was unusual for us to decouple the speaker's pragmatic knowledge from their visual experience, one potential concern with this result was the ambiguity of whether speakers' failed to adapt their referential descriptions because they were inattentive to the identity of the addressee or whether they were aware of the addressee's identity but that the Pragmatic Consistency factor was not effective in

influencing their referential descriptions. Thus to check that speakers were actually sensitive to the identity of the current intended addressee (the Matcher who could hear the speech through the headphones) we opted to follow Gann & Barr (2014) in including a set of unconventional fillers to our study. We expected to find a main effect of Pragmatic Consistency with these unconventional items. Our hypothesis was that for a given target item we would see a greater increase in description length from the last training trial to the test trial when the test addressee was not the same as the training addressee (i.e. in the Pragmatically Inconsistent level).

In our analysis we measured description length in terms of number of words used to describe targets. Our analysis was performed using a linear mixed-effects model with Directors (subjects) and sequence (items) as crossed random factors (Baayen et al., 2008). Analysis used the maximum random effects structure justified by the design (Barr et al., 2013), which implied by-subject and by-item random intercepts and by-subject and by-item random slopes for our two main factors (Perceptual Consistency and Pragmatic Consistency). A Poisson link function was chosen to reflect the distribution of our dependent variable – count data (number of words).

Our analysis of the unconventional items did reveal a significant main effect of Pragmatic Consistency on description length, one-tailed test, z = -4.49, p < 0.01. Thus we successfully replicated the findings of Gann & Barr (2014). Participants used longer descriptions when the addressee was not the same at the test phase (mean = 10.6 words) compared to when they were speaking to the same addressee (mean = 8.2 words). Furthermore, our analysis revealed that 32 out of our 40 participants showed this main effect – indicating that the Pragmatic Consistency factor provided a strong, effective manipulation. We found no significant effect of Visual Consistency on description length of unconventional items, z = -1.19, p = 0.23. Participants showed a similar description length when looking at the same addressee (mean = 9.3 words) compared to looking at a different addressee (mean = 9.5 words) at the test phase.

Table 24 provides an overview of the mean word count broken down by both Visual Consistency and Pragmatic Consistency factors. Figure 20 displays the average description length on test trials in both the Visual Consistency and Pragmatic Consistency factors. Finally, we found no significant interaction between Pragmatic and Visual Consistency factors, z = -1.11, p = 0.27.

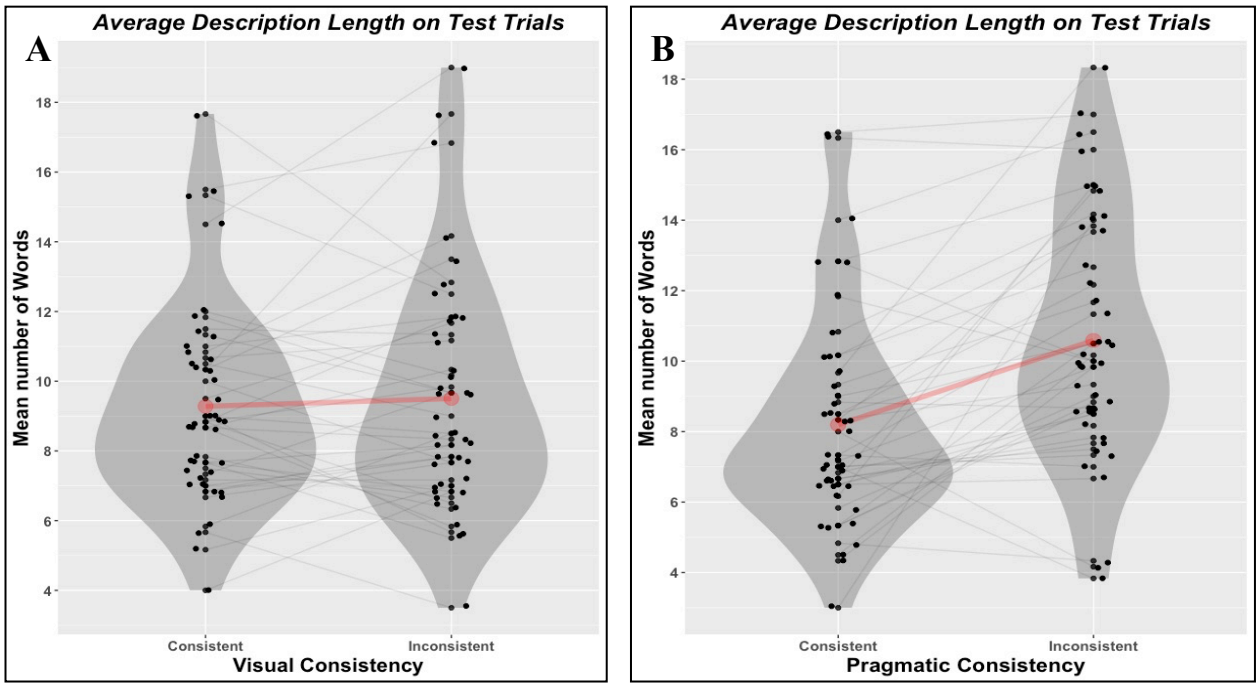


Figure 20: Panel A displays the average description length at the test phase in the Visual Consistency factor. Panel B shows the average description length on test trials in the Pragmatic Consistency factor. Note that each grey line represents a single participant. The red circles represent the grand means across each level of the Shift Direction and Visual/Pragmatic Consistency factors.

Visual Consistency	Pragmatic Consistency	Mean Train Word Count	Mean Test Word Count	Difference (Test-Train)
Consistent	Consistent	7.1	8.2	1.1
Consistent	Inconsistent	6.7	10.4	3.7
Inconsistent	Consistent	6.5	8.2	1.7
Inconsistent	Inconsistent	6.3	10.8	4.5

Table 24: Mean word count by Visual and Pragmatic Consistency factors.

5.5 – Discussion

In our third experiment we created a new paradigm to further investigate the *retrieval fluency hypothesis* for referential encoding. In this study we attempted to dissociate the perceptual cues (visual image of the listener) from the pragmatic cues (knowledge of the identity of the actual listener) that the speaker experienced whilst producing referential descriptions for the addressee in an interactive communication game. This design also enabled us to test the assumption of partner specificity – which forms a key component of

the memory-based model of referential communication (Horton, 2007; Horton & Gerrig, 2005a; Horton & Gerrig, 2016).

The results of Experiment 3 did not reveal any evidence suggesting that participants followed a retrieval fluency heuristic. We found no main effect of Visual Consistency, in fact there was a trend in the opposite direction of our prediction. We predicted that participants would experience greater levels of retrieval fluency when viewing the same Matcher (via the webcam video) at the test phase as they saw during the training phase. It was expected that when speakers were shown the visual image of the Matcher who had also appeared at the training phase this would cue the speaker to use the previous expression they used to describe the target item – even when that description was no longer communicatively relevant at the test phase (much like the Consistent level of the Training-Test Consistency factor in Experiment 2). However, speakers' rate of misspecification at test trials was numerically higher in the Inconsistent level (63%) compared to the Consistent level (57%) of the Visual Consistency factor.

In addition to this, we also failed to find a main effect of Pragmatic Consistency on misspecification. We expected participants to make more referential errors when they were describing the target item to the same Matcher at the test phase that they spoke to during the training phase. However, consistent with Gann and Barr (2014) participants misspecified at the same rate (60%) regardless of whether they were describing items at the Pragmatically Consistent or Inconsistent level. Shift Direction did not have a significant effect on misspecification rate – although speakers did overspecify more often (Contrast-Singleton level, 62.6%) than underspecify (Singleton-Contrast level, 57.6%).

Although it was not a primary component of our analysis we did find a significant effect of Shift Direction on differential onset latency. Similarly to Experiment 2, participants took longer to adapt their description in the Singleton-Contrast level (1262.6ms) compared to the Contrast-Singleton level (535.9ms). This result suggests that participants had greater difficulty adapting their descriptions (to add in additional referential detail) in the Singleton-Contrast level compared to the Contrast-Singleton level.

A key manipulation in this experiment was the implementation of unconventional filler trials that enabled us to test whether the Director was keeping track of the identity of the intended addressee during the test phase of the experiment. Following Gann and Barr (2014), we included abstract grayscale drawings that the Directors were required to describe. These items required participants to provide complex and often detailed descriptions in order for the listener to be able to correctly identify the target image.

Crucially, in our pre-registration, we noted that a non-significant effect in the Pragmatic Consistency factor may prompt the concern that the speaker was inattentive to the identity of the current intended addressee. The unconventional test trials enabled us to address this concern. We expected to see strong effects of Pragmatic Consistency for the unconventional items – with participants providing longer descriptions at the Inconsistent level (when there was a new Matcher at the test phase).

The results supported our hypothesis with speakers providing significantly longer descriptions for new addressees (mean = 10.6 words) compared to old addressees (8.2 words) at the test phase. This was an effective manipulation with 32 of our 40 participants showing this effect. Crucially, there was no significant effect of Visual Consistency on description length. Participants provided a similar description length when looking at the same addressee (mean = 9.3 words) compared to when looking at the alternative addressee (mean = 9.5 words). The significant effect of Pragmatic Consistency suggests that participants engaged in audience design and were aware when they were talking to a different Matcher at the test phase from the one they described target items to during training. This result mirrors Gann and Barr (2014) who note that "ideal speakers" will be sensitive to the addressee's informational needs – using the same amount of words for an old referent when speaking to a new addressee as they would do when describing a new target item to that same addressee.

Gann and Barr (2014) assessed onset latency of descriptions for old and new addressees and found no significant difference in onset. This suggests that speakers avoided underspecifying old referents to new addressees through a process of *monitoring and adjustment* rather than through additional planning. Whilst we did not test for this effect in our current experiment, it is possible that speakers' adopted a similar strategy when describing old referents to new addressees. Thus the successful adaptation of a shortened description could possibly be explained by the fact that speakers adapt their utterances incrementally (Pechmann, 1989). It is relatively easy for speakers to incrementally add additional information to a reduced description without having to undergo extra planning (Gann & Barr, 2014). Notably, since speech is incremental by nature, it is more difficult to avoid producing overspecified descriptions as it is not possible to incrementally delete information that has already been altered (Gann & Barr, 2014).

In this study we manipulated the speaker's perceptual experience (a referential cue that is usually strongly correlated with common ground), in an attempt to test whether additional episodic representations available in memory influence reference generation. Since our

first two experiments lacked communicative relevance we considered this to be a key test of our retrieval fluency hypothesis. The results obtained underline a lack of support for our theory. Although we did not find evidence in support of the retrieval fluency hypothesis we did obtain a significant effect with our unconventional filler trials. This result indicates that the majority of participants were aware of the instances in which they were talking to a different Matcher. Crucially, this knowledge did not improve the accuracy of their descriptions (misspecification rate was identical in the Pragmatically Consistent vs. Inconsistent level).

If the memory-based model is correct in its assumption of partner specificity, then we would have expected participants to be partner specific in their choice of description – using the Matcher at the training phase as a cue to generate descriptions. It therefore appears very unlikely (at least in the current experimental setting) that speakers' used the conversational partner they spoke to during training as a memory cue for designing referential utterances on test trials.

In the following chapter, I will outline the theoretical implications of this result in combination with an overview of the findings from each of our three experiments.

Chapter 6 – General Discussion

6.1 – Summary of Experimental Findings

The three experiments presented in this thesis set out to test the *retrieval fluency hypothesis* for reference generation. This hypothesis proposed that rather than repeatedly consulting common ground with a conversational partner, speakers make snap judgements regarding the contextual appropriateness of a referring expression using heuristic assessments. Crucially, our hypothesis proposed that speakers would judge the appropriateness of a referential expression as a function of retrieval fluency – the relative ease with which an expression comes to mind (Oppenheimer, 2008). As noted in Chapter 2 this hypothesis has two important theoretical components: (1) that speakers store "referring episodes" that link together referents, contexts, and expressions; and (2) that speakers make use of the strength with which referents and contexts cue the retrieval of expressions as an index of the extent to which such expressions are contextually appropriate.

Our hypothesis was influenced by the work of Horton and Gerrig (2005a) and also by Gann and Barr (2014) who applied the encoding specificity principle (Tulving & Thomson, 1973) to partner specificity. We suggested that the fluency with which a speaker's expressions are retrieved would be dependent upon the degree that the referent and the retrieval context match the original encoding context. Our hypothesis therefore proposed that expressions with *strong memory signals* would be more likely to be deemed contextually appropriate by the speaker – resulting in less consideration of context and a shorter delay in speech production, relative to expressions yielding weaker memory signals (see Gann & Barr, 2014 for original proposal).

A key aspect of our study was its potential to serve as a further test of Horton and Gerrig's (2005a) memory-based model for referential communication. This influential theory proposes that many apparent instances of audience design can be explained by automatic memory processes (Horton & Gerrig, 2016). An important feature of the memory-based model is the idea that conversational partners act as memory cues that prompt the retrieval of referential information. We noted that Horton and Gerrig's (2005a) proposal is similar to Gann and Barr (2014) in arguing that the strength of the memory encoded will influence how the speaker incorporates this information into production (see also Brennan and Clark, 1996).

We previously highlighted an overall lack of empirical support for Horton and Gerrig's model (Barr & Keysar, 2002; Brown-Schmidt & Horton, 2014; Brown & Dell, 1987; Kronmüller & Barr, 2007, 2015). Given that the memory-based model relies on the concept of resonance – a parallel search of memory which makes it possible for a range of associated information to become available on the basis of relatively local cues (Horton, 2008, Ratcliff, 1978), we find it surprising that studies that offer support for this model (e.g. Gorman, et al., 2013; Hanna, et al., 2003) often fail to account for the effect that these additional episodic representations will have on memory. In our series of experiments we set out to de-confound these additional contextual effects (e.g. visual similarity between past and present contexts, the colour of objects in an array) from common ground.

In Experiment 1 we had participants play the role of "Director" in a referential communication game. Participants were presented with a series of grids containing letters of various sizes and colours and were tasked with describing the highlighted target letter to the "Matcher" (experimenter). Along with the target letter we manipulated the appearance of the "competitor" and "foil" items that alternated between training and test trials. This Shift Direction factor (Singleton-Contrast vs. Contrast-Singleton) was implemented in order to force the speaker to vary the amount of information they had to provide in the test trials relative to the training trials.

A crucial aspect of this study concerned the presentation of the context that the target items appeared in. We attempted to de-confound the effects of memory from common ground by manipulating the variability of the visual configurations in which the targets appeared – a *communicatively irrelevant* aspect of the stimuli. Thus participants entrained upon descriptions in either the Low Variability Context or the High Variability Context. We expected that when speakers entrained upon descriptions in the Low Variability level (stimuli appeared in a very similar configuration across trials) they would experience a greater level of retrieval fluency in the test phase of the experiment – causing them to produce more referential misspecifications.

However, the results of Experiment 1 failed to provide evidence in support of the retrieval fluency hypothesis. We found no difference in misspecification rate (17%) across the levels of the Context Variability factor. Unexpectedly, when speakers misspecified a description in Experiment 1 they were significantly more likely to underspecify than overspecify their utterance. This result was particularly surprising given that most evidence indicates an effect in the opposite direction - with overspecification more common in referential communication (e.g. Deutsch & Pechmann 1982; Engelhardt et al., 2006).

Notably, participants also produced significantly less fluent descriptions (speech without pauses or misspecifications) in the Singleton-Contrast level compared to the Contrast-Singleton level – indicating that participants found it more difficult to add information to an established description than delete information from a previous utterance.

Experiment 2 served as a further test of the retrieval fluency hypothesis. A key change was the implementation of pictures of everyday objects as stimuli as opposed to the letters shown in Experiment 1. This alteration was made in an attempt to make each target object more distinctive in the speaker's memory – it was expected that this would enable the Director to build up stronger memory traces for their descriptions, creating more fluent retrieval of expressions at the test phase. We also introduced the Training-Test Consistency factor to this experiment. Unlike the Context Variability factor in Experiment 1, all training trials in this experiment were presented in a stable arrangement during the training phase. Instead we manipulated the consistency of the context between the training and test phase (Training-Test Consistent vs. Training-Test Inconsistent). Due to the higher similarly between training and test arrangements, we expected speakers to experience a stronger memory signal in the Training Consistent condition causing them to make more referential errors and engage in audience less effectively.

Results from Experiment 2 provided weak statistical support in favour of the retrieval fluency hypothesis. Speakers misspecified descriptions more often in the Consistent Training-Test level (85%) compared to the Inconsistent level (80%), although this effect barely reached significance in a one-tailed test with a small effect size. We also found that speakers were much more likely to overspecify than underspecify their descriptions. This significant result was in contrast to Experiment 1, where we obtained a statistically significant underspecification effect.

However, although these results were promising, we acknowledged that the cues that we manipulated in Experiment 2 perhaps lacked communicative relevance - making it difficult to apply these findings more broadly to everyday interactions. In Experiment 3, we sought to address this issue. Our final experiment sought to manipulate the level of retrieval fluency that the speaker experienced by using the conversational partner as a memory cue. This experiment differed from our first two studies as we included a second Matcher as part of our design. We implemented two main factors: Visual Consistency and Pragmatic Consistency and included the Shift Direction factor from Experiments 1 and 2. A crucial aim of this experiment was to increase communicative relevance by manipulating a referential cue that is normally strongly correlated with common ground. To do this we de-

confounded the perceptual cue (visual image of the listener) from pragmatic cues (knowledge of the identify of the actual listener).

Experiment 3 did not reveal any evidence that speakers followed a retrieval fluency heuristic when providing referential descriptions to addressees. We predicted that participants would experience stronger levels of retrieval fluency when viewing the same Matcher at the test phase as they saw during training. However, the trend was in the opposite direction from the predicted effect with participants producing more referential errors in the Visually Inconsistent level (63%) than the Visually Consistent level (57%) of the Visual Consistency factor. We also found no main effect of Pragmatic Consistency with speakers producing the same error rate (60%) for descriptions in both the Pragmatically Consistent and Inconsistent levels. Although the Shift Direction factor was not of primary interest in this study, we did find that speakers were more likely to overspecify descriptions (Contrast-Singleton level, 62.6%) than underspecify descriptions (Singleton-Contrast level, 57.6%). Unlike Experiment 2 however, this difference was not statistically significant.

We found no significant effect of Visual Consistency on description length for unconventional referents. We did however find a significant effect for Pragmatic Consistency on description length for unconventional referents – speakers lengthened previously shortened descriptions of items at the test phase for new listeners. This result indicates that speakers successfully adapted their speech to engage in audience design. Moreover, this effect demonstrated a successful manipulation with 32/40 participants adapting their speech. This underlines an important point – participants were clearly able to keep track of who the intended addressee was, ruling out the possibility that they were inattentive to the identity of the current addressee in our study.

Although we did not find evidence for a retrieval fluency effect in either Experiment 1 or Experiment 3 the fact that we obtained a high misspecification rate across all three experiments is an indication that speakers did indeed rely on their previously encoded memories of target item descriptions. In both Experiments 2 and 3, Directors overspecified descriptions during the test phase (e.g. using the description "*Eden white cheese*" when the word "cheese" would have been adequate or "*pipe with a wooden section*" when "pipe" would have been a sufficient description for the addressee to identify the target object). Had participants not relied on their previously encoded memories, there would have been an overall lower misspecification rate because the kinds of misspecifications we obtained would have been highly unlikely without the training experience.

We note that the underspecification effect obtained in Experiment 1 is in contrast to the rate of overspecification obtained in both Experiments 2 and 3. This result could be linked to the stimuli used in Experiment 1. Our first study presented participants with target letters as opposed to the target objects used as stimuli in Experiments 2 and 3. The use of target letters as stimuli in Experiment 1 tended to prompt speakers into using pre-nominal modifiers (e.g. "*the big A*") when they adapted their description on test trials in the Singleton-Contrast level of the Shift Direction factor. In Experiments 2 and 3 speakers produced a higher rate of post-nominal modifiers when describing objects at the test phase in the Contrast-Singleton level (e.g. "*candle, that's not been lit*"). It is possible that the nature of the stimuli in Experiments 2 and 3 made it easier for speakers to post-nominally modify their descriptions (leading to greater overspecification) compared to Experiment 1 where post-nominal descriptions (e.g. "*A, big*") were less common. Thus if the speaker initially failed to use a pre-nominal modifier on test trials in the Singleton-Contrast level of the Shift Direction factor in Experiment 1, it is likely that they continued to use their unmodified description (e.g. "*A*"). These unmodified descriptions would have been the same utterances originally generated during the training phase of the experiment. We believe that this may explain the significant underspecification effect obtained in Experiment 1.

6.2 – Theoretical Implications

The results of Experiment 1 and Experiment 3, coupled with the small effect size of our significant result in Experiment 2, provide little evidence in support of the retrieval fluency hypothesis. In addition to this, our results fail to support Horton and Gerrig's (2005a) assumption that episodic memory effects are a key source of partner specificity in reference production. It is possible that there was something specific about our design or study implementation that could explain the lack of evidence in favour of our hypothesis. As mentioned previously in Chapters 1 and 5, a common criticism of studies which fail to find support for partner specificity in audience design, is the perceived lack of interaction experienced by participants (Brown-Schmidt, 2009; Brown-Schmidt et al., 2015). However, this is a concern that cannot be levelled at our study since all three experiments involved extensive interaction between the Director (participant) and the Matcher (experimenter). Across all three experiments the minimum number of training trials participants completed before being shown a target item in the test phase was 4 trials (6-9 training trials for Experiments 1 and 2, 4-6 trials in Experiment 3). In Experiments 1 and 2 participants completed a minimum of 336 trials describing target items to the Matcher. In Experiment 3 all participants completed 384 trials (number of training sequences for

specific items was randomised across subjects). Therefore participants had ample opportunity to interact with the addressee(s) and generate their own descriptions for target items.

Furthermore, we would also point to Kronmüller and Barr (2015), who note that previous criticism from Brown-Schmidt (2009) has been particularly selective when identifying studies that fail to find partner specific effects due to a lack of interaction. Studies which have opportunities for participants to interactively establish common ground (Barr & Keysar, 2002; Brennan & Hanna, 2009; Shintel & Keysar, 2007) have also failed to find partner specific effects while other experiments that have found support for partner effects in memory (Horton & Slaten, 2012) have done so despite being non-interactive in nature.

Our lack of support for the memory-based model is compatible with previous research that has argued against the idea of partner specificity. For example, we previously highlighted the work of Barr and Keysar (2002), who studied the use of referential precedents in communication. The authors argued that precedents are frequently used in conversation because they are available in memory and can be implemented to solve referential ambiguity - not because they are partner specific. Barr and Keysar (2002) initially predicted that when addressees interacted with a new speaker they would be slower to gaze at and reach out for target objects in their experiment. However, they found no significant difference in reaction times when participants heard the old speaker compared to the new speaker. Barr and Keysar's results support the anchoring and adjustment model of referential communication – speakers and listeners use mutual knowledge only to identify and address coordination problems in communication (Barr & Keysar, 2002).

The results of Experiment 3 in our study support a similar "adjustment" model for reverential communication. Despite a lack of evidence in favour of the retrieval fluency heuristic we found that speakers engaged in audience design when describing unconventional target items (increasing previously shortened descriptions when speaking to a new listener). In Chapter 5, we suggested, based on similar results obtained by Gann and Barr (2014), that participants avoided underspecifying old referents to new addressees through a process of monitoring and adjustment. Unlike Gann and Barr (2014), who provided evidence for this claim by measuring speech onset latency (see Chapter 5 for details), we did not specifically test for monitoring and adjustment in our study. However, our result does offer some support for Monitoring and Adjustment Model (Horton & Keysar, 1996). It is possible that speakers were not accounting for common ground in the initial planning of descriptions but were adapting descriptions for the addressee if and

when it was deemed necessary. It is likely that speakers would have achieved this by incrementally adding additional information to their previously shortened descriptions (Gann & Barr, 2014; Pechmann, 1989).

It difficult to determine the extent to which our retrieval fluency hypothesis supports Pickering and Garrod's (2004) Interactive Alignment Model since our study did not provide a direct test of their theory. We highlighted in Chapter 2 that the retrieval fluency hypothesis may be compatible with the Interactive Alignment Model because retrieval fluency could help to facilitate alignment. Perhaps had we varied the role of the participant during each of our experiments we could have tested some of the main assumptions of the model. For example, we could have had participants play as the Matcher on some trials as well as playing as the Director. This would have enabled participants to experience both conversational roles and may have facilitated greater alignment between the participant and the confederate. Alternatively, had we included two naïve participants instead of using the experimenter as the Matcher in Experiments 1 and 2 then this may have generated greater (and more natural) alignment between both interlocutors. We could have then tested the retrieval fluency hypothesis within this framework.

However, we had clear methodological reasons for our study design. We opted to use the experimenter/lab assistant across all three experiments in order to ensure that participants interacted with a Matcher who knew when it was necessary to ask for additional descriptive information. Additionally, since the experimenter was playing the role of the Matcher they were able to provide quick feedback to the Director. We reasoned that fewer delays in response time from the Matcher would facilitate greater entrainment on descriptions for the Director (participant), which in turn, would help develop stronger memory traces of utterances. In theory, this would enable participants to experience stronger levels of retrieval fluency.

The other reason that we opted to have participants only play the role of the Director was to enable them to gain experience of describing target items during the training phase across a series of trials (between 6-9 trials in Experiments 1 and 2). Since research has indicated that self-generated descriptions are re-used more frequently and are remembered better (Knutsen & Le Bigot, 2012; Knutsen, Ros, & Bigot, 2016) we reasoned that this design would maximise the retrieval fluency effects that participants experienced across trials during the training phase of the experiments.

In Chapter 1, we noted our concern that support for Horton and Gerrig's memory-based model has frequently been based on Horton's (2007) study (e.g. Brown-Schmidt, 2009,

2012; Brown-Schmidt et al., 2015; Gorman et al., 2013; Horton, 2008; Horton & Slaten, 2012). In his 2007 paper, Horton found that subjects were quicker to name pictures with labels that were linked to the current partner at the test phase in comparison to naming items with labels that were associated with an alternative partner. Horton argues that this finding supports the concept of partner-specificity in memory. However, we would urge caution in drawing conclusions from this study. The difference in onset between the current and alternative partner in Experiment 1 in this study was relatively low at 87ms (a similar conclusion in Experiment 2 was based on a difference of 67ms). Moreover, despite this relatively small effect, the author states this result reflects a "significant effect of partner context" (p.1120) with a p value of "$p = < 0.06$". We also note that this difference in onset does not distinguish between a quicker onset time due to common ground between the speaker and the addressee or whether the quicker onset was due to the episodic priming of the picture labels associated with each addressee. Both of these factors are perfectly confounded in Horton's study (2007). Notably, the significant effect that we obtained in Experiment 2 in favour of the retrieval fluency hypothesis did offer some support to the memory-based model. However, similarly to Horton's (2007) study the effect size for own experiment was small (5% difference in misspecification rate between the Training Consistent and Training Inconsistent levels) with $p = 0.03$.

In addition to these results, Brown-Schmidt and Horton (2014) recently failed to replicate Horton's (2007) findings. This replication focused on the result of Experiment 1 and found that there was no significant difference between conversational partners. Participants were only 3ms faster to name pictures that were previously studied with the same partner ($p = 0.40$). The authors carried out two additional studies that sought to further explore their initial findings. Notably, the second of these additional experiments was conducted as a direct replication at 99% power and failed to support Horton's original work. Participants were 26ms slower to name items when they had studied labels with the same partner during training compared to the alternative partner ($p = 0.36$). This result further underlines an emerging lack of support for the memory-based model of communication.

However, despite this strong effect Brown-Schmidt and Horton (2014) suggest that they may have obtained a different result if the partner in the experiment had not been "incidental" to the task. Additionally, they suggest if the similarity between items from training to test has been greater they may have obtained alternative results. They argue that "establishing more clearly motivated partner-item associations could help increase the likelihood that the presence of a specific partner would reliably prompt retrievals of

relevant knowledge" (p.7). We believe that the methodology and design implemented in Experiment 3 of our study addresses this issue. In our study, speakers entrained upon item descriptions with one of two conversational partners at the training phase and were then tested in both Visual and Pragmatic Consistency factors at the test phase. Both partners actively interacted with the speaker throughout the experiment and were therefore a critical component of main manipulation. This was in contrast to Horton's (2007) original study where the partner simply read out object category clues to the participant from a worksheet. Although speakers took longer to provide descriptions at the Visually Inconsistent level than the Visually Consistent level (difference of 15.1ms) and also longer at the Pragmatically Inconsistent level than Pragmatically Consistent level (44.5ms) in our experiment, neither of these effects was significant (nor were there any significant interactions). Our study shows that even when the conversational partner took on a more significant role in the experiment (interacting frequently with the participant) there was a lack of evidence in favour of partner specificity. Furthermore, speakers generated their own descriptions for target items (rather than being cued as was the case in Horton's study) further increasing the partner-specific item associations that were formed during the training phase.

Although Brown-Schmidt and Horton's (2014) work has significant implications for the study of memory in referential communication the results of this study have been overlooked in recent review papers. For instance, Horton and Brennan (2016) provide an overview of the memory-based account in the context of metarepresentation without highlighting the null effect obtained by Brown-Schmidt and Horton (2014) or drawing on previous research that has failed to support the memory-based account. Similarly, Horton and Gerrig (2016) published a review article aimed at expanding their memory-based theory. The authors also used this article as an opportunity to comment on studies that have "weakened" their original claims. Despite using this paper to provide a comprehensive overview of their memory-based account the authors fail to fully address the implications of Brown-Schmidt and Horton's study – merely explaining the results as a consequence of "relatively-arbitrary partner-item associations" which may have been "too tenuous" to enable interlocutor identity to act as a cue in memory (Horton & Gerrig, 2016, p. 791). We believe we have addressed some of the concerns raised by Brown-Schmidt and Horton, (2014) and Horton and Gerrig, (2016) in our third experiment. As noted above we failed to find evidence in favour of partner specificity in memory when we controlled for the additional episodic effects experienced by speakers when providing referential descriptions.

6.3 – Limitations

Our experiments serve as a further test of Horton and Gerrig's (2005a) theory and have important implications for the underlying assumptions of the memory-based model. However, we acknowledge there are limitations with our study. Notably, we obtained no effect of Context Variability in Experiment 1 nor did we find significant effects of Pragmatic Consistency or Visual Consistency in Experiment 3. In fact, if anything there was a trend in the opposite direction of our main prediction for the Visual Consistency factor. In Experiment 2 we did find a main effect of Training-Test Consistency on memory with participants misspecifying more frequently in the Training Consistent level. However, the effect size for this result was relatively small (difference of 5% between conditions).

Nevertheless, we have shown evidence that participants experienced some episodic effects whilst providing descriptions to addressees during the test phase of our experiments. One indication of this was the rate of misspecification across all three experiments in our study. Speakers' consistently misspecified descriptions – either overspecifying by providing too much information in Experiments 2 and 3, or by underspecifying descriptions as was the case in Experiment 1. As we alluded to previously, these misspecification effects would not have been present had it not been for the episodic memories that were developed during the training phase of each experiment. Additionally, in line with this, we would also point to the data from our unconventional referent analysis in Experiment 3. Our results showed participants engaged in successful audience design – lengthening old descriptions of objects for new addressees who had not seen the target item before. This result implies that speakers had formed episodic memories of descriptions they had previously used and were able to recall and adjust these utterances when required.

One possible explanation for the lack of significant episodic effects in our study is that perhaps speakers only kept minimalistic information in their episodic traces during communication. Thus our experiments may have not been sensitive enough to identify episodic effects that were present for speakers. Perhaps if we were to re-test our hypothesis over a longer period of time we may obtain results that would be more favourable for the memory-based model (Horton & Gerrig, 2005a). In effect this would compare the episodic traces created in the lab in our current study to longer-term traces similar to those developed in everyday interactions.

We therefore suggest that focussing on the scope of the memory traces formed by the speaker over a prolonged period of time may provide a greater insight into whether interlocutors take advantage of partner specificity when engaging in referential

communication. There is some evidence in the literature that would support this idea. As mentioned previously, Barr et al.'s (2014) results supported the concept of partner specificity in memory - addressees looked more quickly and reliably at a target image when the addressee's friend read out the description compared to when the description was read out by the lab assistant. Crucially, Barr et al. (2014) took advantage of the common ground that was already established between pairs of friends that participated in this study. When the designer of descriptions was a friend, addressees were able to rely on a shared reference that was based on social familiarity and experience. In our study, if participants had been given a longer period of time to consolidate their memories of the item descriptions during the training phase, perhaps we would have seen more evidence of partner specificity and common ground for specific utterances. Future studies may wish to consider entraining speakers on referential descriptions across a series of testing sessions that occur on separate occasions before then testing whether the speaker relies on partner specificity when generating descriptions. This may enable speakers to build up a stronger retrieval fluency of memories and develop greater partner specific effects.

The idea of having multiple training sessions is supported by literature that shows the benefits of distributed practice in memory. The distribution of multiple study opportunities or practice sessions has shown a robust improvement in memory in word based tasks (e.g. Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Janiszewski & Sawyer, 2003) and also picture tasks (Hintzman & Rogers, 1973; see Benjamin & Tullis, 2010 for an overview of this literature). Furthermore, research shows a marked improvement in memory for participants who are allowed to sleep after processing new information. During sleep newly encoded memory traces (in addition to older related memories) are continually reactivated. In this way, new memories are progressively added to pre-existing knowledge networks (Born & Wilhelm, 2012). Evidence suggests that sleeping facilitates consolidation – strengthening and stabilizing memories formed before sleep onset (Maquet, 2001; Rasch & Born, 2013; Walker & Stickgold, 2006). Thus sleep appears to stimulate the re-processing of new memories and assists with their integration into long-term memory (Diekelmann & Born, 2010; Lewis & Durrant, 2011; Stickgold & Walker, 2013).

For example, in a language acquisition and processing study Dumay and Gaskell (2007) had participants learn fictitious words that overlapped with real words (e.g. "cathed*ruke* vs. cathedral") and compared groups of participants who learned the words in the evening (pm group – before sleep) or in the morning (am group – after sleep). The authors found that

lexical competition between the fictitious words and real words was not observed after immediate exposure nor after a full day awake. Participants' only experienced lexical competition when they had enjoyed a night's sleep after learning the competitor words. This finding underlines the impact sleep has on memory consolidation (Dumay & Gaskell, 2007) and is a factor that should be considered when testing the memory-based model.

In Chapter 2, we noted that previous research has distinguished between recollection and familiarity in memory (Yonelinas, 1994; Yonelinas et al., 2010). A potential limitation of our study is that we did not dissociate the metacognitive effect of retrieval fluency from the cognitive effects that may have been experienced due to familiarity in our experiments. Across all three experiments we attempted to manipulate the retrieval fluency experienced by participants. However we did not include an independent measurement to validate the extent of these fluency effects. Perhaps we could have included a memory test for participants at the end of the experiment to determine the strength of the memory formed. This could have taken the form of recognition test where participants were required to identify whether an object had been seen before or not (old vs. new). This would have enabled us to determine whether participants showed sensitivity to our experimental manipulations outwith the communicative context they were originally presented in. Nevertheless, as we have highlighted previously in this chapter, the high misspecification rate obtained across all three experiments is a strong indication that participants successfully formed episodic memories of the items they were shown.

Finally, we would also comment on the recent debate that has highlighted potential concerns of using confederates in language production and dialogue tasks (Kuhlen & Brennan, 2013). In all three of our studies, the experimenter or a lab assistant undertook the role of addressee. Whilst this is common practice in dialogue studies – we acknowledge that ideally we would have tested our hypotheses using speaker and listener who were both naïve to the purpose of our study.

One potential issue with this set-up was that the Matcher (experimenter) always knew which item within the array was the target object (although this was never actually revealed to the participant). As a result of this, the experimenter quickly became familiar with the descriptions participants commonly used to identify referents during the experiment. In some instances it was therefore possible that the Matcher was able to identify a target object from an inadequate description produced by the Director. Had the addressee been naïve they may have required additional information from the speaker. This factor may have affected the overall misspecification rate of descriptions even further than

the rate in our current study. Additionally, participants knew that the Matcher was the experimenter, which could have influenced their descriptions for target objects. For example, some speakers may have adopted a more lackadaisical approach to reference production – under the impression that the experimenter would ask for more information if they provided an inadequate description (the Matcher was instructed to say "*which one do you mean?*" if the description was insufficient). However, it is unlikely that these factors would have influenced the overall outcome of our study. Although there are pragmatic benefits of using confederates in language production experiments, future research should attempt to avoid doing so when possible.

6.4 – Future Directions/Closing Remarks

The research in this doctoral thesis tested for the retrieval fluency hypothesis for audience design. Across three experiments we found little evidence supporting our hypothesis. Crucially, our results also fail to support Horton and Gerrig's (2005a) memory-based model for referential communication. In Experiment 3 of our study we did not find evidence of partner specificity in memory – a key component of Horton and Gerrig's original model. Given that the memory-based model is a prominent theory in referential communication our results (coupled with Brown-Schmidt and Horton's 2014 recent failed replication) highlight a need for additional research to address some of the key assumptions of this model.

In light of our results, we would encourage others to attempt to replicate the original effects of partner specificity on memory (Horton, 2007; Horton & Gerrig, 2005a). In particular, in order to build upon our findings, it is crucial that researchers attempt to do so whilst also de-confounding the effects of memory from common ground. This will enable us to get a clearer idea of whether partner specificity acts as critical cue for memory or whether other aspects in the communicative environment also play an important role. As well as making new attempts to test Horton and Gerrig's (2005a) memory-based model we would encourage others to attempt to replicate the effects obtained in our retrieval fluency experiments – particularly the results obtained in Experiment 3 where we de-confounded both visual and pragmatic cues using two separate addressees.

Whilst we appreciate that Horton and Gerrig's (2005a) original paper was designed to promote further discussion of the role of memory in referential communication, we believe that future research should be more specific when generating hypotheses that test for memory effects in common ground. For example, whilst arguing that memory acts as a proxy for common ground, Horton and Gerrig (2005a) are generally non-specific in

hypothesising how or when this is likely to happen. This makes it relatively easy for researchers to claim support for the memory-based model. Since episodic memory is a crucial component of our everyday interactions – it is important to be more specific when hypothesising about its role in reference production. In the future, researchers should state more clearly how memory is expected to impact upon common ground and in what circumstances these effects would be likely to occur.

With this in mind we would encourage researchers to be as transparent as possible when generating their hypotheses. When designing our study we pre-registered our hypotheses and analysis procedures on the Open Science Framework (OSF) before beginning data collection. We believe this to be an important step in increasing clarity and promoting replication in psychology. In order to advance the study of audience design in communication we would encourage researchers to commit to pre-registration. Going forward, this will help to address any lack of transparency in the literature and aid attempts to replicate important findings that help shape our understanding of the impact of memory on referential communication. Clearly, we also believe future research should seek to address some of the additional issues that we have raised above. Whilst the idea of partner specificity in memory is appealing, our results highlight that there is still some ambiguity as to whether speakers use their communicative partner as a memory cue in audience design.

Appendices

Appendix 1:

Participant Instruction Sheet for Experiment 1:

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 letters. In each trial a single letter will be highlighted by a yellow outline. Your task is to verbally describe this letter so that the Matcher is able to identify it on a separate computer monitor (please see ***Fig. 1*** below). Although, the Matcher's monitor will contain the same letters as those that appear on your screen, they will be arranged in a completely random order. Therefore, it is unlikely that the letters will appear in the same locations as those shown on your screen. In order to provide an accurate instruction to the Matcher, you must avoid using the *spatial location* of the target letter in your description. You may, however, describe the letter in any other way that you think may help Matcher to locate the target item.

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 experiment. There will be a full debrief after the experiment is finished.

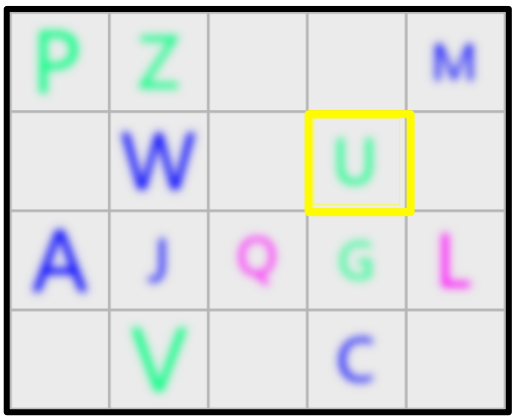


Figure 1: Example of the display on the *Director's screen*. The Director will describe the highlighted target letter ('u' in this example) to the Matcher. Once the Matcher has selected the letter, a new trial will begin.

Appendix 2:

Error rate of Participants removed from analysis in Experiment 1:

Session ID	**Shift Direction**	**Number of Invalid Trials**	**Error Rate (%)**
1	Singleton-Contrast	17/24	70.8
8	Singleton-Contrast	19/24	79.2
9	Singleton-Contrast	15/24	62.5
16	Singleton-Contrast	18/24	75.0
24	Singleton-Contrast	24/24	100.0
34	Singleton-Contrast	24/24	100.0
36	Singleton-Contrast	22/24	91.7
37	Contrast-Singleton	24/24	100.0
40	Singleton-Contrast	18/24	90.0
44	Singleton-Contrast	15/24	62.5

**Note that Session ID 43 was also removed due to continued use of overly long descriptions across all trials in the experiment.*

Appendix 3:

List of Target and Competitor/Foil Objects for Experiment 2:

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	Golf 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

Appendix 4:

Participant Instruction Sheet for Experiment 2:

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 →

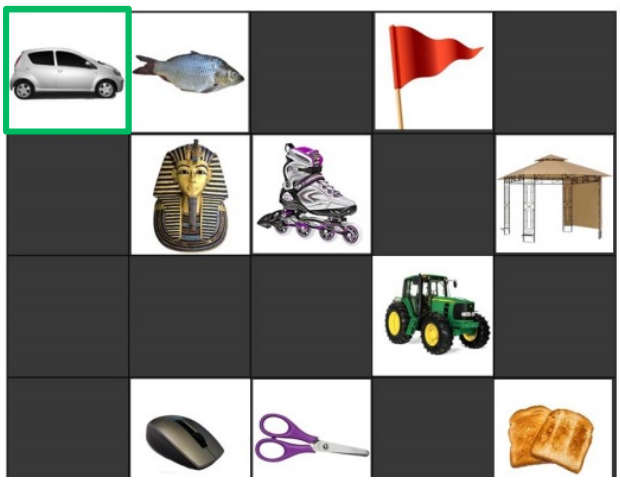


Figure 1: Example of the display on the *Participant's screen*. The Participant will identify the highlighted target object ("car" in this example) to the Experimenter. Once the Experimenter has selected the letter, a new trial will begin.

Appendix 5:

Error rate of Participant removed from analysis in Experiment 2:

Session ID	**Shift Direction**	**Number of Invalid Trials**	**Error Rate (%)**
4	Singleton-Contrast	13/24	54.2

List of Stimuli Items removed from analysis across all participants in Experiment 2:

Target Object	**Competitor**	**Foil**	**Shift Direction**	**Error Rate (%)**
Bedroom lamp	Lava lamp	Rocket	Singleton-Contrast	55.6
Bicycle helmet	Builders helmet	Mellon	Singleton-Contrast	68.4
Bumblebee	B letter	D letter	Singleton-Contrast	52.6
Computer mouse	Mouse	Squirrel	Singleton-Contrast	72.2
Mountain bike	Motor bike	'Go' Kart	Singleton-Contrast	57.9
Office phone	Mobile phone	Remote control	Singleton-Contrast	78.9
Reading glasses	Drinking glasses	Test beakers	Singleton-Contrast	55.6
Spoon	Wooden spoon	Wooden spatula	Singleton-Contrast	63.2

Appendix 6:

List of Target and Competitor/Foil Objects for Experiment 3:

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
Kitchen knife	Swiss army knife	USB stick
Leather glove	Boxing glove	Bean bag
Gold key	Car key	Ping pong bay
Riding saddle	Bicycle saddle	Putter
Camcorder	CCTV camera	Hairdryer
Orange	Orange slice	Sunset picture
Sun hat	Cowboy hat	Wooden bowl
Gun	Toy gun	Hook
AA battery	Car battery	Box
Money (notes)	Money (coins)	Bolts and screws
Boot	Car boot	Breadbin
Red apple	Green apple	Pear
Acoustic guitar	Electric guitar	Frying pan
Garden spade	Beach spade	Spatula
Horse	Rocking horse	Cradle

Mirror	Hand mirror	Wreath
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
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

Wrapping Bow	Crossbow	Hairpin
Crocodile	Crocodile Inflatable	Green surfboard
Carrots	Carrots chopped	Orange pegs
Wall plug	Sink plug	White cd
Bicycle helmet	Crash helmet	Bowling ball
Nail for hammer	Finger nail	Raw chicken breast
Vase	Vase with flowers	Grass tuft
Pizza	Pizza slice	Cake slice
Mouse wired	Mouse wireless	Black/silver ring

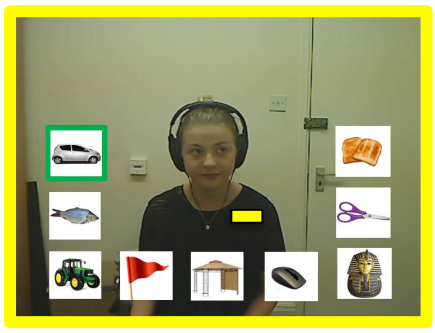
Appendix 7:

Participant Instruction Sheet for Experiment 3:

You will play the role of the "Director" and will verbally name the TARGET item (highlighted by a **green** outline) to one of two Matchers who will sit in a separate room from you. Figures 1 & 2 below show the two people who will be listening to your descriptions. They will interact with you through a live webcam video. Only one Matcher will be able to hear your description at a time. The Matcher who appears on the screen may not be the person listening to your description.

In the examples shown in Figures 1 & 2 the target item is the *car*. You would describe this item to the listener (e.g. *"Select the car")* so that they are able to identify it on their computer monitor. Figure 3 shows the view of the Matchers' screen. They will select the item you describe using the corresponding numbers on their keyboard.

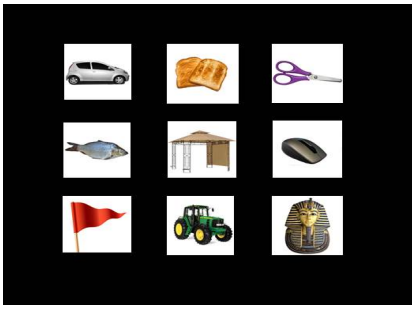
Before we start we will have a practice session!

1. 

Caitlyn will be the **YELLOW** Matcher.

2. 

Kieran will be the **ORANGE** Matcher.

3.  +

Appendix 8:

Error rate of Participant removed from analysis in Experiment 3:

Session ID	**Shift Direction**	**Number of Invalid Trials**	**Error Rate (%)**
4	Singleton-Contrast	13/24	54.2

Item removed from analysis across all participants in Experiment 3:

Target Object	**Competitor**	**Foil**	**Shift Direction**	**Error Rate (%)**
Wrapping Bow	Crossbow	Hairpin	Singleton-Contrast	94.1

References

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review : An Official Journal of the Society for Personality and Social Psychology, Inc*, *13*(3), 219–35. https://doi.org/10.1177/1088868309341564

Arnold, J. E. (2008). Reference production: Production-internal and addressee-oriented processes. *Language and Cognitive Processes*, *23*(4), 495–527. https://doi.org/10.1080/01690960801920099

Ashcraft, M. H., & Radvansky, G. A. (2010). *Cognition* (Fifth Edit). Pearson Education, Inc.

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*(2008), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Barr, D. J. (2008). Analyzing "visual world" eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*, 457–474. https://doi.org/10.1016/j.jml.2007.09.002

Barr, D. J. (2014). Perspective Taking and its Imposters in Language Use: Four Patterns of Deception. In T. M. Holtgraves (Ed.), *The Oxford Handbook of Language and Social Psychology.* (pp. 98–110). New York, USA: Oxford University Press.

Barr, D. J., Jackson, L., & Phillips, I. (2014). Using a voice to put a name to a face: The psycholinguistics of proper name comprehension. *Journal of Experimental Psychology. General*, *143*(1), 404–13. https://doi.org/10.1037/a0031813

Barr, D. J., & Keysar, B. (2002). Anchoring Comprehension in Linguistic Precedents. *Journal of Memory and Language*, *46*(2), 391–418. https://doi.org/10.1006/jmla.2001.2815

Barr, D. J., & Keysar, B. (2006). Perspective Taking and the Coordination of Meaning in Language Use. In J. Traxler, Matthew & A. Gernsbacher, Morton (Eds.), *Handbook of Psycholinguistics* (Second Edi, pp. 901–938). USA: Elsevier.

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. https://doi.org/doi.org/10.1016/j.jml.2012.11.001

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective ? *Cognitive Psychology*, *61*, 228–247. https://doi.org/10.1016/j.cogpsych.2010.05.004

Born, J., & Wilhelm, I. (2012). System consolidation of memory during sleep. *Psychological Research*, *76*, 192–203. https://doi.org/10.1007/s00426-011-0335-6

Bornstein, R. F., & Dagostino, P. R. (1992). Stimulus Recognition and the Mere Exposure Effect. *Journal of Personality and Social Psychology*, *63*(4), 545–552.

Bortfeld, H., & Brennan, S. E. (1997). Use and Acquisition of Idiomatic Expressions in Referring by Native and Non-Native Speakers. *Discourse Processes*, *23*, 119–147. https://doi.org/10.1080/01638537709544986

Brennan, S. E., & Clark, H. H. (1996). Conceptual Pacts and Lexical Choice in Conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1482–1493.

Brennan, S. E., & Hanna, J. E. (2009). Partner-Specific Adaptation in Dialog. *Topics in Cognitive Science*, *1*(2), 274–291. https://doi.org/10.1111/j.1756-8765.2009.01019.x

Brown-Schmidt, S. (2009). Partner-specific interpretation of maintained referential precedents during interactive dialogue. *Journal of Memory and Language*, *61*(2), 171–190. https://doi.org/10.1016/j.jml.2009.04.003.Partner-specific

Brown-Schmidt, S. (2012). Beyond common and privileged: Gradient representations of common ground in real-time language use. *Language and Cognitive Processes*, *27*(1), 62–89. https://doi.org/10.1080/01690965.2010.543363

Brown-Schmidt, S., & Horton, W. S. (2014). The Influence of Partner-Specific Memory Associations on Picture Naming: A Failure to Replicate Horton (2007). *PLoS ONE*, *9*(10), 1–8. https://doi.org/10.1371/journal.pone.0109035

Brown-Schmidt, S., Yoon, S. O., & Ryskin, R. A. (2015). *People as contexts in conversation*. *Psychology of Learning and Motivation - Advances in Research and Theory* (Vol. 62). Elsevier Ltd. https://doi.org/10.1016/bs.plm.2014.09.003

Brown, P. M., & Dell, G. S. (1987). Adapting production to comprehension: The explicit

mention of instruments. *Cognitive Psychology*, *19*(4), 441–472. https://doi.org/10.1016/0010-0285(87)90015-6

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed Practice in Verbal Recall Tasks : A Review and Quantitative Synthesis. *Psychological Bulletin*, *132*(3), 354–380. https://doi.org/10.1037/0033-2909.132.3.354

Clark, H. H. (1973). The Language-as-Fixed-Effect Fallacy: A Critique of Language Statistics in Psychological Research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335–359.

Clark, H. H. (1992). *Arenas of language use*. Chicago: University of Chicago Press.

Clark, H. H. (1996). *Using Language*. Cambridge, UK: Cambridge University Press.

Clark, H. H., & Carlson, T. B. (1981). Context for Comprehension. In J. Long & A. Baddeley (Eds.), *Attention and performance IX* (pp. 313–330). Lawrence Erlbaum Associates, Publishers, Hillsdale, New Jersey.

Clark, H. H., & Carlson, T. B. (1982). Speech acts and hearers' beliefs. In N. V. Smith (Ed.), *Mutual Knowledge* (Ed, pp. 1–36). New York: Academic Press.

Clark, H. H., & Marshall, C. R. (1978). Reference diaries. In D. L. Waltz (Ed.), *Theoretical issues in natural language processing* (Vol.2, pp. 57–63). New York: Association for Computing Machinery.

Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. . K. Joshe, B. L. Webber, & I. A. Sag (Eds.), *Elements of discourse understanding* (pp. 10–61). Cambridge, UK: Cambridge University Press.

Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. In J. Le Ny & W. Kintsch (Eds.), *Language and comprehension* (pp. 287–299). Amsterdam, The Netherlands: North Holland Publishing.

Clark, H. H., Schreuder, R., & Buttrick, S. (1983). Common Ground and the Understanding of Demonstrative Reference. *Journal of Verbal Learning and Verbal Behaviour*, *22*, 245–258.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*, 1–39.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of Processing: A Framework for Memory Research. *Journal of Verbal Learning and Verbal Behavior*, *11*, 671–684.

Dell, G. S., & Brown, P. M. (1991). Mechanisms for Listener-Adaptation in Language Production: Limiting the Role of the "Model of the Listener." In D. J. Napoli & J. A. Kegl (Eds.), *Bridges between psychology and linguistics: A Swarthmore festscrift for Lila Gleitman.* (Eds, pp. 105–129). Hillsdale, NJ: Erlbaum.

Deutsch, W., & Pechmann, T. (1982). Social interaction and the development of definite descriptions. *Cognition*, *11*(2), 159–184. https://doi.org/10.1016/0010-0277(82)90024-5

Dickerson, B. C., & Eichenbaum, H. (2010). The Episodic Memory System : Neurocircuitry and Disorders. *Neuropsychopharmacology REVIEWS*, *35*, 86–104. https://doi.org/10.1038/npp.2009.126

Diekelmann, S., & Born, J. (2010). The memory function of sleep. *Nat. Rev. Neurosci.*, *11*, 114–126. https://doi.org/10.1038/nrn2762

Dumay, N., & Gaskell, M. G. (2007). Sleep-Associated Changes in the Mental Representation of Spoken Words. *Psychological Science*, *18*(1), 35–39. https://doi.org/10.1111/j.1467-9280.2007.01845.x

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. https://doi.org/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*, 304–314. https://doi.org/10.1016/j.bandc.2011.07.004

Ferreira, V. S. (2008). Ambiguity, accessibility, and a division of labor for communicative success. *Learning Motiv*, *49*(1), 209–246. https://doi.org/doi:10.1016/S0079-7421(08)00006-6.

Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, *40*(4), 296–340. https://doi.org/10.1006/cogp.1999.0730

Ferreira, V. S., Slevc, L. R., & Rogers, E. S. (2005). How do speakers avoid ambiguous

linguistic expressions? *Cognition*, *96*(3), 263–84. https://doi.org/10.1016/j.cognition.2004.09.002

Fussell, S. R., & Krauss, R. M. (1989a). The effects of intended audience on message production and comprehension: Reference in a common ground framework. *Journal of Experimental Social Psychology*, *25*, 203–219. https://doi.org/10.1016/0022-1031(89)90019-X

Fussell, S. R., & Krauss, R. M. (1989b). Understanding friends and strangers: The effects of audience design on message comprehension. *European Journal of Social Psychology*, *19*, 509–525.

Fussell, S. R., & Krauss, R. M. (1992). Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. *Journal of Personailty and Social Psychology*, *62*(3), 378–391.

Gann, T. M., & Barr, D. J. (2014). Speaking from experience: audience design as expert performance. *Language, Cognition and Neuroscience*, *29*(6), 744–760. https://doi.org/10.1080/01690965.2011.641388

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordiation. *Cognition*, *27*, 181–218.

Garrod, S., & Doherty, G. (1994). Conversation, co-ordination and convention: an empirical investigation of how groups establish linguisitc conventions. *Cognition*, *53*, 181–215.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, *8*(1), 8–11. https://doi.org/10.1016/j.tics.2003.10.016

Gillund, G., & Shiffrin, R. M. (1984). A Retrieval Model for Both Recognition and Recall. *Psychological Review*, *91*(1), 1–67.

Glucksberg, S., Krauss, R. M., & Weisberg, R. (1966). Referential Communication in Nursery School Children: Method and Some Preliminary Findings. *Journal of Experimental Child Psychology*, *3*, 333–342.

Godden, D. R., & Baddeley, A. D. (1975). Context-Dependent Memory In Two Natural Environments: On Land And Underwater. *British Journal of Psychology*, *66*(3), 325–331. https://doi.org/10.1111/j.2044-8295.1975.tb01468.x

Goldinger, S. D. (1996). Words and Voices : Episodic Traces in Spoken Word Identification and Recognition Memory. *Journal of Experimental Psychology; Learning, Memory, and Cognition*, *22*(5), 1166–1183.

Gorman, K. S., Gregg-Harrison, W., Marsh, C. R., & Tanenhaus, M. K. (2013). What's Learned Together Stays Together: Speakers' Choice of Referring Expression Reflects Shared Experience. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*(3), 843–853. https://doi.org/doi:10.1037/a0029467

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

Hanna, J. E., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task : evidence from eye movements, *28*, 105–115. https://doi.org/10.1016/j.cogsci.2003.10.002

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*, *49*(1), 43–61. https://doi.org/10.1016/S0749-596X(03)00022-6

Heller, D., Grodner, D., & Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition*, *108*, 831–836. https://doi.org/10.1016/j.cognition.2008.04.008

Hintzman, D. L., & Rogers, M. K. (1973). Spacing effects in picture memory. *Memory and Cognition*, *1*(4), 430–434.

Horton, W. S. (2007). The influence of partner-specific memory associations on language production: Evidence from picture naming. *Language and Cognitive Processes*, *22*(7), 1114–1139.

Horton, W. S. (2008). A memory-based approach to common ground and audience design. In I. Kecskes (Ed.), *Intention, common ground, and the egocentric speaker-hearer* (pp. 189–222). Berlin: Mouton De Gruyter.

Horton, W. S., & Brennan, S. E. (2016). The Role of Metarepresentation in the Production and Resolution of Referring Expressions. *Frontiers in Psychology*, (July), 7:1111. https://doi.org/10.3389/fpsyg.2016.01111

Horton, W. S., & Gerrig, R. J. (2002). Speakers' experiences and audience design:

Knowing when and knowing how to adjust utterances to addressees. *Journal of Memory and Language*, *47*(4), 589–606. https://doi.org/10.1016/S0749-596X(02)00019-0

Horton, W. S., & Gerrig, R. J. (2005a). Conversational Common Ground and Memory Processes in Language Production. *Discourse Processes*, *40*(1), 1–35. https://doi.org/10.1207/s15326950dp4001

Horton, W. S., & Gerrig, R. J. (2005b). The impact of memory demands on audience design during language production. *Cognition*, *96*(2), 127–42. https://doi.org/10.1016/j.cognition.2004.07.001

Horton, W. S., & Gerrig, R. J. (2016). Revisiting the memory-based processing approach to common ground. *Topics in Cognitive Ccience*, (January 2016), 1–30. https://doi.org/10.1111/tops.12216

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition*, *59*, 91–117.

Horton, W. S., & Slaten, D. G. (2012). Anticipating who will say what: The influence of speaker-specific memory associations on reference resolution. *Memory & Cognition*, *40*, 113–126. https://doi.org/10.3758/s13421-011-0135-7

Isaacs, E. A., & Clark, H. H. (1987). References in Conversation Between Experts and Novices. *Journal of Experimental Psychology. General*, *116*(1), 26–37.

Jacoby, L. L., Woloshyn, V., & Kelley, C. (1989). Becoming famous without being recognized: Unconscious influences of memory produced by dividing attention. *Journal of Experimental Psychology: General*, *118*(2), 115–125. https://doi.org/10.1037//0096-3445.118.2.115

Janiszewski, C., & Sawyer, A. G. (2003). A Meta-analysis of the Spacing Effect in Verbal Learning : Implications for Research on Advertising Repetition and Consumer Memory. *Journal of Consumer Research*, *30*.

Johnson-Laird, P. N. (1982). Mutual ignorance: Comments on Clark and Carlson's paper. In N. . Smith (Ed.), *Mutual Knowledge* (Ed). London: Academic Press.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgement. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heurisitics*

and biases: The psychology of intuitive judgement (Eds, pp. 49–81). Cambridge, UK: Cambridge University Press.

Keysar, B. (1994). The Illusory Transparency of Intention: Linguistic Perspective Taking in text. *Cognitive Psychology*, *26*, 165–208.

Keysar, B. (1998). Language users as problem solvers: Just what ambiguity problem do they solve? In S. R. Fussell & R. Kreuz (Eds.), *Social and Cognitive approaches to Interpersonal Communication* (Eds, pp. 175–200). Mawah, NJ: Lawrence Erlbaum Associates.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking Perspective in Conversation: The Role of Mutual Knowledge in Comprehension. *Psychological Science*, *11*(1), 32–38. https://doi.org/10.1111/1467-9280.00211

Keysar, B., Barr, D. J., Balin, J. A., & Paek, T. S. (1998). Definite Reference and Mutual Knowledge: Process Models of Common Ground in Comprehension. *Journal of Memory and Language*, *39*(1), 1–20. https://doi.org/10.1006/jmla.1998.2563

Keysar, B., Barr, D. J., & Horton, W. S. (1998). The egocentric basis of language use: Insights from a processing approach. *Current Directions in Psychological Science*, *7*(2), 46–50. https://doi.org/10.1017/CBO9781107415324.004

Knutsen, D., & Le Bigot, L. (2012). Managing dialogue : How information availability affects collaborative reference production. *Journal of Memory and Language*, *67*(3), 326–341. https://doi.org/http://dx.doi.org/10.1016/j.jml.2012.06.001

Knutsen, D., Ros, C., & Bigot, L. (2016). Generating References in Naturalistic Face-to-Face and Phone-Mediated Dialog Settings. *Topics in Cognitive Science*, *8*(2016), 796–818. https://doi.org/10.1111/tops.12218

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

Kronmüller, E., & Barr, D. J. (2007). Perspective-free pragmatics: Broken precedents and the recovery-from-preemption hypothesis ☆ . *Journal of Memory and Language*, *56*(3), 436–455. https://doi.org/10.1016/j.jml.2006.05.002

Kronmüller, E., & Barr, D. J. (2015). Referential precedents in spoken language

comprehension: A review and meta-analysis. *Journal of Memory and Language*, *83*(2015), 1–19. https://doi.org/10.1016/j.jml.2015.03.008

Kuhlen, A. K., & Brennan, S. E. (2013). Language in dialogue : when confederates might be hazardous to your data. *Psychon*, *20*, 54–72. https://doi.org/10.3758/s13423-012-0341-8

Levison, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicatures.* Cambridge: MA: MIT Press.

Lewis, P. A., & Durrant, S. J. (2011). Overlapping memory replay during sleep builds cognitive schemata. *Trends in Cognitive Sciences*, *15*(8), 343–351. https://doi.org/10.1016/j.tics.2011.06.004

Lockridge, C. B., & Brennan, S. E. (2002). Addressees' needs influence speakers' early syntactic choices. *Psychonomic Bulletin and Review*, *9*(3), 550–557.

Logan, G. D. (1988). Toward an Instance Theory of Automization. *Psychological Review*, *95*(4), 492–527.

Logan, G. D. (1990). Repetition priming and automaticity: Common underlying mechanisms? *Cognitive Psychology*, *22*, 1–35.

Logan, G. D. (1992). Shapes of reaction-time distributions and shapes of learning curves: A test of the instance theory of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 883–914.

Logan, G. D. (1997). Automaticity and reading: Perspectives from the instance theory of automatization. *Reading & Writing Quarterly: Overcoming Learning Difficulties*, *13*, 123–146. https://doi.org/10.1080/1057356970130203

Maquet, P. (2001). The Role of Sleep in Learning and Memory. *Science*, *294*, 1048–1053.

McGlone, M. S., & Tofighbakhsh, J. (2000). Birds of a feather flock conjointly (?): Rhyme as reason in aphorisms. *Psychological Science*, *11*(5), 424–428.

Metzing, C., & Brennan, S. E. (2003). When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, *49*, 201–213. https://doi.org/10.1016/S0749-596X(03)00028-7

Mol, L., Krahmer, E., Maes, A., & Swerts, M. (2011). Seeing and Being Seen: The Effects

on Gesture Production. *Journal of Computer-Mediated Communication*, *17*, 77–100. https://doi.org/10.1111/j.1083-6101.2011.01558.x

Nadig, A. S., & Sedivy, J. C. (2002). Evidence of Perspective-Taking Constraints in Children's On-Line Reference Resolution. *Psychological Science*, *13*(4), 329–336. https://doi.org/10.1111/j.0956-7976.2002.00460.x

Oppenheimer, D. M. (2006). Consequences of erudite vernacular utilized irrespective of necessity: problems with using long words needlessly. *Applied Cognitive Psychology*, *20*(2), 139–156. https://doi.org/10.1002/acp.1178

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences*, *12*(6), 237–41. https://doi.org/10.1016/j.tics.2008.02.014

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic Encoding of Voice Attributes and Recognition Memory for Spoken Words. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *19*(2), 309–328.

Paraboni, I., Masthoff, J., & van Deemter, K. (2006). Overspecified reference in hierarchical domains : measuring the benefits for readers. *Proceedings of the 4th International Conference on Natural Language Generation (INLG-06), Sydney, Australia*, (July), 55–62.

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

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *The Behavioral and Brain Sciences*, *27*(2), 169-190-226. https://doi.org/10.1017/S0140525X04000056

Pickering, M. J., & Garrod, S. (2006). Alignment as the Basis for Successful Communication. *Research on Language and Computation*, *4*(2006), 203–228. https://doi.org/10.1007/s11168-006-9004-0

Rasch, B., & Born, J. (2013). About sleep's role in memory. *Physiol. Rev.*, *93*, 681–766. https://doi.org/10.1152/physrev.00032.2012

Ratcliff, R. (1978). A Theory of Memory Retrieval. *Psychological Review*, *85*(2), 59–108.

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth.

Consciousness and Cognition, *8*(3), 338–342. https://doi.org/10.1006/ccog.1999.0386

Reber, R., Winkielman, P., & Schwarz, N. (1998). Effects of perceptual fluency on affective judgements. *Psychological Science*, *9*, 45–48.

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

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory : REM-retrieving effectively from memory. *Psychonomic Bulletin and Review*, *4*(2), 145–166.

Shintel, H., & Keysar, B. (2007). You said it before and you'll say it again: expectations of consistency in communication. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *33*(2), 357–369. https://doi.org/10.1037/0278-7393.33.2.357

Simpson, I. C., Mousikou, P., Montoya, J. M., & Defior, S. (2013). A letter visual-similarity matrix for Latin-based alphabets. *Behavior Research Methods*, *45*(2), 431–9. https://doi.org/10.3758/s13428-012-0271-4

Sonnenschein, S. (1984). The Effects of Redundant Communications on Listeners : Why Different Types May Have Different Effects. *Journal of Psycholinguistic Research*, *13*(2), 147–166.

Sonnenschein, S., & Whitehurst, G. J. (1982). The Effects of Redundant Communications on the Behavior of Listeners: Does A Picture Need A Thousand Words? *Journal of Psycholinguistic Research*, *11*(2), 115–125.

Sperber, D. (1982). Comments on Clark and Carlson's paper. In N. V. Smith (Ed.), *Mutual Knowledge* (Ed, pp. 46–51). London: Academic Press.

Sperber, D., & Wilson, D. (1982). Mutual knowledge and relevance in the theories of comprehension. In N. V. Smith (Ed.), *Mutual Knowledge* (Ed). London: Academic Press.

Stickgold, R., & Walker, M. P. (2013). Sleep-dependent memory triage: evolving generalization through selective processing. *Nature Neuroscience*, *16*(2), 139–145. https://doi.org/10.1038/nn.3303

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.),

Organization of Memory (Ed, pp. 381–403). New York: Academic.

Tulving, E. (2002). Episodic Memory: From Mind to Brain. *Annu. Rev. Psychol.*, *53*, 1–25.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*(5), 352–373. https://doi.org/10.1037/h0020071

Tversky, A., & Kahneman, D. (1973). Avaliability: A Heuristic for Judging Frequency and Probability. *Cognitive Psychology*, *5*, 207–232.

Vaidya, C. J., Zhao, M., Desmond, J. E., & Gabrieli, J. D. E. (2002). Evidence for cortical encoding specificity in episodic memory: Memory-induced re-activation of picture processing areas. *Neuropsychologia*, *40*(12), 2136–2143. https://doi.org/10.1016/S0028-3932(02)00053-2

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

Walker, M. P., & Stickgold, R. (2006). Sleep, Memory, and Plasticity. *Annu. Rev. Psychol.*, *57*, 139–166. https://doi.org/10.1146/annurev.psych.56.091103.070307

Wardlow Lane, L., & Ferreira, V. S. (2008). Speaker-external versus speaker-internal forces on utterance form: Do cognitive demands override threats to referntial success? *J Exp Psychol Learn Mem Cogn*, *34*(6), 1466–1481. https://doi.org/10.1037/a0013353.Speaker-external

Wardlow Lane, L., Groisman, M., & Ferreira, V. S. (2006). Don't talk about pink elephants! : Speakers' control over leaking private information during language production. *Psychological Science*, *17*(4), 273–277. https://doi.org/10.1111/j.1467-9280.2006.01697.x.Don

Wardlow Lane, L., & Liersch, M. J. (2012). Can you keep a secret? Increasing speakers' motivation to keep information confidential yields poorer outcomes. *Language and Cognitive Processes*, *27*(3), 462–473. https://doi.org/10.1080/01690965.2011.556348

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: evidence for a dual-process model. *Journal of Experimental Psychology; Learning, Memory, and Cognition*, *20*(6), 1341–1354.

Yonelinas, A. P., Aly, M., Wang, W.-C., & Koen, J. D. (2010). Recollection and Familiarity : Examining Controversial Assumptions and New Directions. *Hippocampus*, *20*, 1178–1194. https://doi.org/10.1002/hipo.20864

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, *9*(2), 1–27.