Prediction involves two stages

Prediction involves two stages: Evidence from visual-world eye-tracking

Ruth E. Corps[1, 2], Charlotte Brooke[1], & Martin J. Pickering[1]

[1] Psychology of Language Department, Max Planck Institute for Psycholinguistics

[2] Department of Psychology, University of Edinburgh

Author note:

This research has been presented at a poster session at the July 2nd EPS Online Meeting and the 26th Architectures and Mechanisms For Language Processing conference, and in a short talk at the 34th CUNY Conference on Sentence Processing.

Data and analysis scripts are available at: https://osf.io/nkud5/

Authors' accepted manuscript – in press in *Journal of Memory and Language*

Please address correspondence to:

Ruth Elizabeth Corps

Psychology of Language Department

Max Planck Institute for Psycholinguistics

Nijmegen

The Netherlands

Ruth.Corps@mpi.nl

Word count: 13043 (excluding tables and figures)

Abstract

Comprehenders often predict what they are going to hear. But do they make the best predictions possible? We addressed this question in three visual-world eye-tracking experiments by asking when comprehenders consider perspective. Male and female participants listened to male and female speakers producing sentences (e.g., *I would like to wear the nice…*) about stereotypically masculine (target: tie; distractor: drill) and feminine (target: dress, distractor: hairdryer) objects. In all three experiments, participants rapidly predicted semantic associates of the verb. But participants also predicted consistently – that is, consistent with their beliefs about what the speaker would ultimately say. They predicted consistently from the speaker's perspective in Experiment 1, their own perspective in Experiment 2, and the character's perspective in Experiment 3. This consistent effect occurred later than the associative effect. We conclude that comprehenders consider perspective when predicting, but not from the earliest moments of prediction, consistent with a two-stage account.

Keywords: prediction, perspective-taking, gender-stereotyping, visual-world-paradigm, language comprehension

1. Introduction

Comprehenders often predict what they are going to encounter. For example, immediately after hearing a speaker say *The boy will eat…*, they tend to look at edible objects, suggesting that they predict that the speaker is about to mention such an object (e.g., Altmann & Kamide, 1999). But what exactly do comprehenders predict? And more importantly, what information do they use to make these predictions? Do they initially make the best predictions they can, or do such predictions take time and resources?

To investigate these questions, we consider when comprehenders take the perspective of the agent of the action or event described by a sentence. When a speaker utters a sentence about him or herself (using *I*), the speaker corresponds to the agent. Throughout, we use an example involving a female speaker and a male comprehender (and we assume that the comprehender perceives the speaker to be female). Let us assume that the female speaker utters *I would like to wear…*. If the male comprehender takes the perspective of the speaker, then he is likely to predict that she will refer to a stereotypically feminine article of clothing, such as a dress. Predicting this object depends on believing certain stereotypes, and also believing that the speaker will refer to objects compatible with these stereotypes. In the experiments we report in this paper, we carefully determined that both females and males from our population of participants held these stereotypes. That is, females were likely to prefer a dress and males were likely to believe that a female would refer to a dress. It would of course be useful for the comprehender to "step into the speaker's shoes" in this way, because his predictions will tend to correspond to what she actually ends up saying. Thus, prediction should ultimately tend to be *consistent* – that is, consistent with the comprehender's beliefs about what the speaker will ultimately say.

Note that our discussion refers to females and males and does not consider other gender identities (e.g., Hyde, Bigler, Joel, Tate, & van Anders, 2019). We recruited

participants and asked them to identify their gender and whether it matched the gender they were assigned at birth. They all identified as either female or male and said that their gender matched their birth gender. Therefore our discussion is in terms of (cisgender) female and male participants. We also assume that our participants have gender-binary stereotypes (and hence that their notions of femininity and masculinity are themselves stereotyped). For example, they might regard a dress as stereotypically feminine; they could also regard it as stereotypically masculine or gender-neutral, but could not regard it as stereotypically of another gender. Finally, comprehenders also make a stereotyped judgment about the gender of the speaker (based on characteristics such as voice and visual appearance) and we again assume that this judgment is binary (i.e., on a one-dimensional feminine-masculine axis). In sum, we are concerned with participants' gender stereotypes with respect to their own identity, the identity of other people, and objects and activities (such as dresses and ties), so that we can investigate the effects of perspective-taking on prediction in comprehenders.

1.1. One- and two-stage accounts of prediction

We have assumed that ultimate prediction is consistent. But is initial prediction also consistent? In other words, do comprehenders initially predict in a manner that is the same as how they would ultimately predict, and therefore consistent with the comprehender's beliefs? (By *initial prediction*, we mean predictions that occur rapidly and are not preceded by other predictions.)

We therefore contrast one- and two-stage accounts of prediction. According to a *one-stage* account, initial prediction does not differ from ultimate prediction, and so prediction is initially consistent. In our example, the male comprehender would initially predict the female speaker will refer to a stereotypically feminine article of clothing. But according to *two-stage* accounts, initial prediction is governed by different principles from ultimate prediction. We

have noted that prediction depends on perspective-taking, and we know that perspective-taking can be effortful (e.g., Lin, Keysar, & Epley, 2010). Thus, the comprehender may ignore at least some aspects of background knowledge during initial prediction, but pay attention to those aspects during ultimate prediction. If so, he might not initially predict that the speaker will refer to a stereotypically feminine article of clothing.

There are different two-stage accounts of prediction, but we identify two possibilities. On the *egocentric two-stage* account, the male comprehender initially predicts from his own perspective – that is, on the basis of what he himself would assume under the circumstances (e.g., Keysar, Barr, Balin, & Brauner, 2000). In this case, he initially predicts that the female speaker will refer to a stereotypically masculine article of clothing (e.g., a tie), compatible with his own gender stereotypes. On the *associative two-stage* account, the comprehender initially predicts on the basis of automatically generated associations (e.g., Neely, 1977; Perea & Gotor, 1997). He therefore activates semantic associates of the lexical entry for *wear* in a bottom-up manner, and uses this activation to initially predict that the speaker will refer to any wearable object. In this case, he initially predicts that the speaker could refer to either a stereotypically feminine article of clothing (e.g., a dress) or a stereotypically masculine article of clothing (e.g., a tie).

If the male comprehender predicts consistently (here, predicting a dress rather than a tie, consistent with his beliefs about the speaker's gender identity and gender stereotypes) from the earliest moments of processing, and there is no stage at which he predicts inconsistently, then a *one-stage* account of prediction would be correct. But if he initially predicts associatively (here, predicting both a dress and a tie) or egocentrically (predicting a tie rather than a dress), and predicts consistently only later, a *two-stage* account of prediction would be correct.

In our experiments, we therefore asked (1) whether comprehenders *ultimately* predict egocentrically, associatively, or consistently, and (2) whether they *initially* predict associatively, egocentrically, or consistently. We expect that the answer to (1) is that they ultimately predict consistently, but the answer to (2) is much less clear, as we discuss below. We tested among these alternatives in three experiments using the visual-world paradigm (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), which has been used to investigate both predictive processing (Altmann & Kamide, 1999) and perspective-taking (see Barr, 2016, for a review).

Below, we consider the evidence for one- and two-stage accounts of prediction. We then discuss effects of gender stereotyping during language comprehension. Finally, we describe our studies and formulate our hypotheses in more detail.

1.2. Contrasting one- and two-stage accounts of prediction

The contrast between one- and two-stage accounts of prediction echoes the distinction between interactive and modular (or encapsulated) accounts of language comprehension (see Fodor, 1983). Historically, a major focus was on parsing – how comprehenders initially select among analyses of syntactically ambiguous sentences. One-stage (or interactive) accounts of parsing assume that people can immediately draw on all potentially relevant information (MacDonald, Pearlmutter, & Seidenberg, 1994), such as background knowledge. In contrast, two-stage (or modular) accounts of parsing (Frazier, 1987) assume that initial decisions are based on some sources of information (e.g., some aspects of syntax) but not others (e.g., real-world knowledge). There is thus a distinction between initial and ultimate aspects of parsing. An extensive body of experimental work has sought to distinguish these accounts (e.g., Rayner, Carlson, & Frazier, 1983; Trueswell, Tanenhaus, & Garnsey, 1994),

with visual-world studies providing some evidence for early use of information that appears inconsistent with at least some two-stage accounts (Tanenhaus et al., 1995).

1.2.1. Evidence for a one-stage account

Research in the domain of perspective-taking has been heavily concerned with the distinction between one- and two-stage accounts. One-stage (or constraint-based) accounts propose that comprehenders integrate their own and their partner's perspectives (along with linguistic information) simultaneously from the earliest moments of processing (e.g., Brown-Schmidt, Gunlogson, & Tanenhaus, 2008; Hanna, Tanenhaus, & Trueswell, 2003; Sikos, Tomlinson, Heins, & Grodner, 2019). For example, Hanna et al. (2003, Experiment 1; see also Hanna & Tanenhaus, 2004) found that a participant following instructions from a confederate (e.g., *Now put the blue triangle on the red one*) fixated shared targets (e.g., a red triangle) that were visible to both the confederate and the participant more than targets that the participant knew were visible to just the participant. But participants also fixated the privileged target more often than an unrelated shape (e.g., a yellow square), suggesting they could not ignore their own egocentric perspective completely.

A one-stage account of prediction would also claim that comprehenders integrate perspective from the earliest moments of processing and initially predict consistently – that is, as well as they can, given everything they know. In line with this argument, Kamide, Altmann, and Haywood (2003; Experiment 2) found that participants draw on real-world knowledge rapidly when they make predictions. For example, participants fixated a picture of a motorbike when they heard *The man will ride the…*, but a picture of a carousel when they heard *The girl will ride the…*., and these fixations occurred around verb offset. The rapid use of real-world knowledge is compatible with a one-stage account, but the authors did not

investigate the time-course of consistent (agent/verb-based) versus associative (verb-based) predictions, and so there may be two effects that differed in their time course.

Heller, Grodner, and Tanenhaus (2008) presented participants with displays containing two pairs of size-contrasting objects. One pair (e.g., a big bowl and a small bowl) was visible to both participants (and the comprehender realized the speaker could see them). This was also the case for one object (e.g., a big car) from the other pair, but the other object (e.g., a small car) was visible only to the comprehender. On hearing *The big…*, comprehenders fixated the big bowl more than the big car, suggesting they took the speaker's perspective into account – the speaker could see only one car and would therefore be likely to say *The car* to refer to the big car. Such predictions are potentially consistent as they are compatible with the comprehender paying attention to the speaker's perspective. But the study does not demonstrate a one-stage account of prediction – it does not test whether associative predictions (that is, *The big* triggering looks to all big objects) precede consistent ones (that is, *The big* triggering looks to big objects that the comprehender realizes the speaker could see).

In another study, Creel (2012) found that (adult) participants who were explicitly told a character's colour preferences (e.g., a female speaker preferred pink) rapidly activated this information during comprehension: They fixated objects that matched the speaker's colour preference, even before the speaker mentioned such an object. Moreover, Borovsky and Creel (2014) familiarized (adult) participants with two talkers (e.g., a pirate and a princess) whose roles were strongly associated with particular objects. Participants predictively fixated objects consistent with the talker and the verb, for example a sword while they heard a pirate say *I would like to hold…*, well before they heard *hold*. Note also that Borovsky, Elman, and Fernald (2012) found similar results when participants heard sentence like *The pirate will hold …*

Importantly, perspective was highly salient throughout the sentences in all of these studies – in Creel (2012) and Borovsky and Creel (2014), participants even knew which objects the speaker would likely refer to before the sentence. This characteristic of the studies makes it impossible to determine whether there is an initial (encapsulated) stage of prediction that is inconsistent with ultimate prediction. Moreover, it may have obscured any effects of prediction that are not driven by perspective. For example, Borovsky and Creel's (2014) study did not determine whether the verb *hold* can lead comprehenders to predict all hold-able objects, or objects that they themselves are likely to hold, because the strong context had already led to a strong prediction of a piratical object. Thus, although research has investigated whether comprehenders integrate perspective into their predictions, these studies do not show that they do so from the earliest moments of processing. In the next section, we discuss evidence for two-stage accounts.

1.2.2. *Reasons and evidence for a two-stage account*

Instead of initially predicting consistently, comprehenders could instead initially predict either egocentrically (from their own perspective) or associatively (based on word associations). In accord with an egocentric two-stage account, there is some evidence that listeners comprehend egocentrically, and tend to ignore perspective, at least during the initial stages of bottom-up comprehension (e.g., Barr, 2008; Keysar et al., 2000; Keysar, Barr, Balin, & Paek, 1998; Kronmüller, Noveck, Rivera, Jaume-Guazzini, & Barr, 2017). For example, Keysar et al. (2000; see also Wu, Barr, Gann, & Keysar, 2013) had a confederate instruct participants to reorganise objects in a grid. Participants knew that some objects (e.g., a small candle) were visible only to them, while others (medium and large candles) were visible to both them and the participant. Even though they knew that the confederate had no knowledge of the small candle, they often considered it as a potential referent when the

confederate said *Now put the small candle above it.* In another study, Damen, van der Wijst, van Amelsvoort, and Krahmer (2020; see also Epley, Keysar, Van Boven, & Gilovich, 2004; Weingartner & Klin, 2005) found that participants expected an addressee to interpret a message as sarcastic even when participants were explicitly told that the addressee did not know the speaker's intention.

But egocentric prediction is likely to be inefficient – and specifically less efficient than consistent prediction. If comprehenders initially predict egocentrically then they will predict what they would say if they were producing the utterance themselves, rather than what the speaker would say. In some instances, the comprehender's perspective is the same as the speaker's, and so egocentric prediction will be sufficient for accurate comprehension. In other instances, however, the comprehender's perspective will differ from the speaker's and egocentric prediction will lead to errors in understanding and the need for reinterpretation.

Comprehenders could reduce error while still minimising cognitive load by instead initially predicting associatively, in accord with a two-stage associative account. On this account, the comprehender rapidly activates semantic associates of words, which makes them easier to process when they are subsequently encountered (e.g., Pickering & Gambi, 2018). For example, semantic priming studies show that *doctor* is easier to process after the participant reads *nurse* (e.g., Bentin, McCarthy, & Wood, 1985; Meyer & Schvaneveldt, 1971). We interpret the activation of such associates as reflecting prediction, just as looks to pictures corresponding to likely arguments of verbs are interpreted as reflecting prediction (e.g., Altmann & Kamide, 1999).

There is much evidence that comprehenders can predict in this way (e.g., Kukona, Cho, Magnuson, & Tabor, 2014; Kukona, Fang, Aicher, Chen, & Magnuson, 2011; Sauppe, 2016). For example, Kukona et al. (2011) found that listeners looked at both a picture of a

robber and a picture of a policeman after hearing *Bill will arrest…*, suggesting that the concept *arrest* associatively activated both *policeman* and *robber*, thus increasing fixations to these pictures, even though *policeman* is an unlikely patient of *arrest*. In our example, a male comprehender encountering a female speaker say *I would like to wear…* rapidly activates the representation for the word *wear* in a bottom-up manner, and activation then spreads to linked representations, such as those of wearable objects. Comprehenders use this spreading activation to predict that the speaker is likely to refer to associates of the verb. Out of the set of wearable associates, some of these would be stereotypically feminine (e.g., a dress), while others would be stereotypically masculine (e.g., a tie). Thus, some associative predictions will be egocentric and others will be consistent (and still others will be neither egocentric nor consistent, for example stereotypically masculine objects when a female speaker addresses a female comprehender).

One reason why comprehenders might not initially integrate perspective into their predictions is that doing so requires time and resources. For example, Lin et al. (2010; Wardlow, 2013) found that participants with lower working memory capacity comprehended egocentrically more than participants with higher capacity, perhaps because perspective-taking requires theory of mind and representation of two versions of the world (e.g., Keysar et al., 2003). There is evidence that predicting what a speaker is likely to say can be cognitively demanding: Ito, Corley, and Pickering (2018) found that participants who performed a working memory task while simultaneously comprehending showed later predictive looks than those who did not (see also Huettig & Janse, 2016).

In sum, a one-stage account claims that prediction is initially consistent. But according to two-stage accounts, initial prediction is either egocentric or associative. In our experiments, we tested these three possibilities by constructing differences in gender identity. In particular, male and female participants listened to male and female speakers producing

sentences about stereotypically male and female objects. We now discuss evidence that language processing is sensitive to gender stereotyping.

1.3. Gender stereotyping and language processing

Many studies indicate that language processing is affected by gender stereotyping (e.g., Carreiras, Garnham, Oakhill, & Cain, 1996; Osterhout, Bersick, & McLaughlin, 1997; Sturt, 2003). For example, Carreiras et al. found that participants had difficulty reading sentences containing pronouns (e.g., *He also gave an injection to avoid an infection*) that conflicted with the stereotypical gender of a previously introduced occupation (e.g., *The nurse had to suture the injury*).

Such stereotypes appear to be activated automatically. For example, Banaji and Hardin (1996) found that participants were faster to judge the gender of targets (e.g., *she*) when they were preceded by gender congruent primes (e.g., *nurse*) rather than gender incongruent primes (e.g., *doctor*). This effect occurred regardless of whether participants were aware of the relationship between the target and the prime. Similarly, Oakhill, Garnham, and Reynolds (2005; see also Garnham, Oakhill, & Reynolds, 2002; Reynolds, Garnham, & Oakhill, 2006) instructed participants to judge whether two terms (relating to occupation and roles) referred to the same person, and found that they were slower and less accurate at making these judgments when the stereotypical gender of the first term conflicted with the second.

Importantly for our purposes, Van Berkum, van de Brink, Tesink, Kos, and Hagoort (2008) found that listeners automatically make stereotype judgments from a speaker's voice. In their study, sentence content could be inconsistent with stereotypes evoked from the speaker's voice. For example, listeners heard a male speaker say *Before I leave I always check whether my make-up is still OK* or a female speaker say *I broke my ankle playing*

soccer with friends. ERPs showed that stereotype inconsistencies elicited an N400 at the relevant word (e.g., *make-up* for a male speaker; *soccer* for a female speaker), much like when participants heard a word whose meaning did not fit the context of the sentence (e.g., *The earth revolves around the* ***trouble*** *in a year*). These results demonstrate that stereotypes had an immediate effect on language comprehension. Note, however, that not all the sentences involved gender stereotypes and so the study does not demonstrate that comprehenders use the speaker's gender identity to predict what they are likely to say.

Previous studies therefore demonstrate that gender stereotypes are automatically activated during language processing. However, these studies have been largely limited to questions of whether stereotypical gender is automatically assigned to referring expressions (e.g., *the nurse*) or voice (e.g., being surprised to hear a male speaker refer to *make-up*). In contrast, we ask whether comprehenders use gender stereotypes to predict what a speaker is likely to say, rather than just to comprehend what the speaker is actually saying.

1.4. Overview of experiments

We do not know whether comprehenders consider perspective from the earliest moments of prediction. If prediction involves one stage, then comprehenders (assuming they follow the gender stereotypes that we have discussed) initially predict that the speaker will refer to an object stereotypically consistent with their gender. But taking the consistent perspective requires cognitive effort, and so comprehenders may initially predict egocentrically (from their own perspective), or associatively (based on semantic associations between words) before they predict consistently, as suggested by a two-stage account. Note that we expect that prediction will eventually be consistent.

We tested these possibilities in three experiments using the visual-world paradigm, in which we recorded participants' eye movements as they listened to sentences containing a

predictive verb (e.g., *wear*), which was associatively related to two of the four depicted objects – that is, to the target objects (e.g., a tie and a dress) but not the distractor objects (e.g., a drill and a hairdryer). We created differences in perspective by creating differences in gender identity. Thus, we manipulated the gender of the speaker (as indexed by their voice and a picture; see Van Berkum et al., 2008), the participants, and the characters in the sentences. In particular, male and female participants listened to a male or a female speaker producing sentences (e.g., *I would like to wear the nice…*) about gender-stereotyped objects displayed on-screen. These objects were rated for their (stereotypical) masculinity and femininity by a separate group of participants. One target (e.g., dress) and one distractor (e.g., hairdryer) were stereotypically feminine, while the other target (e.g., tie) and distractor (e.g., drill) were stereotypically masculine. The speaker then produced the noun compatible with their gender and the verb (here, the male speaker produced *tie* and the female speaker produced *dress* because these objects were rated as stereotypically masculine and stereotypically feminine in the pre-test). Note that all participants in all experiments identified as male or female and with the gender they were assigned at birth.

By comparing looks to the two targets when the speaker and participant had different (or mismatching) genders, we determined whether participants predicted egocentrically, associatively, or consistently, both initially (soon after encountering the verb) and ultimately (but before encountering the noun). If comprehenders predict associatively, then they should look at both targets more than both distractors, though our analyses compared looks to the target stereotypically compatible with the speaker's gender (the agent-compatible target) with the distractor stereotypically compatible with the speaker's gender (the agent-compatible distractor) in order to make them comparable with the other analyses. If comprehenders predict consistently, then they should look at the agent-compatible target more than the target stereotypically compatible with their own gender (the agent-incompatible target). If

comprehenders predict egocentrically, they should look at the agent-incompatible target more than the agent-compatible target.

The time-course of looks should be informative about initial and eventual prediction. According to a one-stage account, where initial predictions are unencapsulated, participants should predict consistently from the earliest moments of prediction. But in a two-stage account, where initial predictions are encapsulated, then such initial predictions should not be consistent: They should either be associative or egocentric, and comprehenders should then shift from making associative or egocentric predictions to making consistent predictions.

Although the sentences in Experiment 1 always used the pronoun *I*, we varied the agent in the sentences in Experiments 2 and 3. In Experiment 2, we used the pronoun *You* rather than *I,* which allowed us to separate a consistent effect from a simple effect of speaker gender. Assuming that participants treated *You* as referring to themselves, consistent predictions are now tied to their own perspective. Assuming that they follow the gender stereotypes that we have discussed, we expect them to look at the target stereotypically compatible with their own gender (the agent-compatible target) more than the agent-incompatible target. In Experiment 3, we replaced the pronouns with the name *James* (stereotypically male) or *Kate* (stereotypically female) to determine whether participants could predict consistently when the agent's name indicated their gender. If so, then we would expect participants to look at the target stereotypically compatible with the character's gender (the agent-compatible target) more than the target stereotypically compatible with their own gender (the agent-incompatible target). Table 1 gives an overview of the manipulations used in the different experiments.

Table 1. The sentences and objects used in the experiments.

Experiment	Agent	Agent Gender	Example sentence	Agent-compatible target	Agent-incompatible target	Agent-compatible distractor	Agent-incompatible distractor
1	*I*	Female	*I would like to wear the nice dress*	Dress	Tie	Hairdryer	Drill
		Male	*I would like to wear the nice tie*	Tie	Dress	Drill	Hairdryer
2	*You*	Female	*You would like to wear the nice dress*	Dress	Tie	Hairdryer	Drill
		Male	*You would like to wear the nice tie*	Tie	Dress	Drill	Hairdryer
3	*Kate*/*James*	Female	*Kate would like to wear the nice dress*	Dress	Tie	Hairdryer	Drill
		Male	*James would like to wear the nice tie*	Tie	Dress	Drill	Hairdryer

2. Experiment 1

2.1. Method

2.1.1. *Participants*

We recruited 24 native English speakers (aged between 18 and 25; *M*age = 21.29; 12 males, 12 females) from the University of Edinburgh, who participated in exchange for £5. Participants had no known speaking, reading, or hearing impairments. Our sample size was based on previous studies using the visual-world paradigm with a similar design (e.g., Altmann & Kamide, 1999). Our study involved more items than previous experiments (e.g., 28 critical sentences vs. 16 in Altmann & Kamide, 1999), and so we likely had sufficient power to detect an effect. Indeed, related studies tend to have a similar number of critical trials to our study or fewer (e.g., Kukona et al., 2011, had 640 trials; Altmann & Kamide, 1999, had 384, Borovsky & Creel, 2014, had 294, and we had 672). All experiments were approved by the University of Edinburgh ethics committee.

After the experiment, participants completed a questionnaire in which they indicated their gender, and whether they identified as the gender they were assigned at birth (see Appendix B). These questions were open-ended (i.e., gender was not assumed to be binary), and so participants could answer in any way they wished. Importantly, all participants reported being male or female and identified as the gender they were assigned at birth.

2.1.2. Materials

We created 56 pairs of sentences (as produced by the female and male speakers), each with a display of four objects (see Appendix for a full list of stimuli). The sentences contained predictable verbs (e.g., *wear*), so that two of the four depicted objects were associates (specifically, plausible patients) of the verb (i.e., targets; e.g., a tie and a dress), whereas the other two were not (i.e., distractors; e.g., a drill and a hairdryer). The sentences

began with *I,* the verb was followed by *the* and an adjective, and the sentences ended with the object that was associated with the verb and was stereotypically compatible with the speaker's gender. In this example, the sentence pair was *I would like to wear the nice dress* for the female speaker, and *I would like to wear the nice tie* for the male speaker.

We confirmed that sentences predicted the two associates using an online object selection pre-test, in which ten further participants from the same population (*M*age = 19.30, 5 males, 5 females) read sentences truncated at the final word, each accompanied by four coloured pictures. Participants were instructed to "select which of the four objects you think someone producing this sentence could refer to next (not necessarily what you would refer to next). In some cases, this will be two objects. In other cases, this will be four objects". Participants expected the speaker to refer to an average of 1.9 objects. Importantly, participants selected the two referents that were associates of the verb (e.g., the tie and the dress after reading *I would like to wear the nice…*) 96.5% of the time.

Twenty-eight of these sentences were *gendered*, meaning that two of the four pictures were stereotypically feminine (e.g., feminine target: *dress;* feminine distractor: *hairdryer*), and the other two were stereotypically masculine (e.g., masculine target: *tie;* masculine distractor: *drill*). We assessed the stereotypy of these pictures using a second online pre-test, in which 80 participants (*M*age = 19.01, 40 males and 40 females) from the same population as the main experiment were randomly assigned to one of four stimuli lists (20 per list), each containing 120 colour clipart pictures. For each picture, participants: (1) named the object, activity, or job depicted in the picture, and (2) rated the masculinity or femininity of the object, activity, or job depicted in the picture on a 1-100 scale. For half the female and half the male participants, 1 indicated that the object, activity, or job was rated as strongly masculine and 100 indicated that it was rated as strongly feminine. The scale was reversed for the rest of the participants.

On average, pictures designed to be stereotypically masculine were considered masculine (an average rating of 19.44 when 1 = masculine, and 82.06 when 100 = masculine), and pictures designed to be stereotypically feminine were considered feminine (83.27 when 100 = feminine, 17.75 when 1 = feminine). To compare the ratings of male and female participants and of stereotypically masculine and feminine pictures, we collapsed the two rating scales by calculating the difference between the maximum or minimum of the scale and the picture's average stereotypy rating across participants (see Table 2). Importantly, stereotypically masculine pictures were considered just as masculine as stereotypically feminine pictures were feminine (i.e., the difference between the maximum or minimum of the rating scale and the average stereotypy rating was similar for the stereotypically masculine and feminine pictures; $t(54) = -0.68$, $p = .50$). Additionally, ratings were unaffected by participants' own gender: The difference between the maximum or minimum of the rating scale and the average stereotypy rating was similar for the male and female participants ($F(1, 120) = 1.11$, $p = .29$) and there was no interaction between target gender and participant ($F(1, 120) = 0.42$, $p = .52$), suggesting that male and female participants did not rate stereotypically masculine and feminine targets differently. Finally, participants tended to agree on the names of the object, activity, or job depicted in the pictures, and this agreement did not differ for stereotypically masculine and feminine pictures ($t(54) = 1.06$, $p = .30$). When referring to the pictures in the eye-tracking experiment, we used the picture name that most participants used. Picture names were matched for their syllable length ($t(54) = 0.88$, $p = .38$).

Table 2.

The means (and standard deviations) of agreement on the name of the object, job, or activity depicted in the picture, the syllable length of the picture name, and the difference between the average stereotypy rating and the maximum or minimum of the rating scale for targets in the gendered and gender-neutral items. Ratings are reported collapsed across all participants, and separately for male and female participants.

		Stereotypically Masculine Picture	Stereotypically Feminine Picture	Gender-Neutral Target 1	Gender-Neutral Target 2
Picture name agreement[a]		88% (18%)	92% (12%)	93% (11%)	94% (12%)
Picture name syllable length		1.75 (0.75)	1.93 (0.77)	1.89 (0.57)	2.11 (0.88)
Distance from maximum or minimum of the rating scale[b]	Overall	18.19 (6.64)	16.74 (9.11)	48.34 (6.61)	48.61 (4.26)
	Male Participant	17.08 (7.73)	16.48 (11.03)	47.03 (8.94)	48.72 (7.92)
	Female Participant	19.21 (8.80)	16.99 (9.52)	49.29 (7.86)	50.59 (7.86)

[a] The percentage of participants who agreed on the name of the object, activity, or job depicted in the picture.

[b] Average difference between average stereotypy ratings and the maximum or minimum of the scale. For one group of participants, 1 indicated that the depicted object, activity, or job was masculine, while 100 indicate it was feminine. If these participants rated a stereotypically feminine

picture, then distance was calculated as the object's average stereotypy rating (across all participants) subtracted from 100 (the corresponding maximum of the scale). If these participants rated a stereotypically masculine picture, distance was calculated as the object's average stereotypy rating minus 1 (the corresponding minimum of the scale). For the other group of participants, the rating scale was reversed (1 = feminine, 100 = masculine).

The other 28 sentences were gender-neutral. They were designed to make our gender manipulation less obvious, while also allowing us to further test the time-course of associative prediction. They were similar in length and structure to the gendered sentences (e.g., *I would like to eat the nice…*), but the four accompanying pictures were rated as gender-neutral in the pre-test (an average stereotypy rating of 50.34 when 1 = masculine, and 50.61 when 100 = masculine). Two of the four pictures were potential targets of the verb (e.g., an apple and a banana). Participants agreed on the name of these pictures, and there was no difference in the name agreement ($t(54) = -0.29$, $p = .78$) or syllable length ($t(54) = -1.09$, $p = .28$) of the picture names.

Sentences were recorded by a native British English male speaker and a native British English female speaker, who produced the sentences at a natural, slow rate. For the gendered sentences, the speaker always referred to the target that was stereotypically compatible with their gender, as identified by participants in the stereotypy pre-test (i.e., the male speaker referred to *tie* and the female speaker referred to *dress*), so that any predictions participants made based on the speaker's gender would always be accurate. For the gender-neutral sentences, the speaker arbitrarily referred to one of these two pictures, in a manner consistent across the two speakers (e.g., if the male speaker referred to *apple*, the female speaker also referred to *apple*). Sentences were between 2221 and 4472 ms, and sentences produced by the two speakers were matched for their duration, the onset and offset of the critical verb, and the onset of the target (all *p*s > .09 in t-tests; see Table 3).

Table 3.

The means (and standard deviations) of sentence duration, critical verb onset and offset, and target onset for the sentences produced by the male speaker and the female speaker (top) in Experiment 1. The bottom panel shows the means (and standard deviations) of the difference between the stereotypy rating and the maximum or minimum of the rating for sentences produced by the male speaker and the female speaker. Differences are reported collapsed across all participants, and separately for male and female participants.

Duration descriptives				
Speaker Gender	Duration	Verb Onset	Verb Offset	Target Onset
Male	2880 (474)	1252 (397)	1579 (437)	2247 (459)
Female	2951 (272)	1339 (327)	1701 (311)	2323 (312)
Stereotypy descriptives				
Speaker Gender	Participant Gender	Mean (and standard deviation) of the distance from maximum or minimum of the rating scale		
Male	Overall	15.28 (3.10)		
	Male	17.15 (5.27)		
	Female	13.43 (3.94)		
Female	Overall	16.10 (2.37)		
	Male	19.40 (5.03)		
	Female	12.79 (4.16)		

We assessed the stereotypical masculinity and femininity of the speakers' utterances using a third online pre-test, in which 40 participants (*M*age = 19.43, 20 males, 20 females) from the same population as the main experiment were randomly assigned to one of two stimuli lists (20 per list) each containing 56 audio sentences used in the main experiment. For

each sentence, participants rated the masculinity or femininity of the speaker's voice on a 1-100 scale (with the direction reversed for half the male and half the female participants).

On average, the male speaker was considered masculine (an average stereotypy rating of 15.52 when 1 = masculine, and 83.96 when 100 = masculine), and the female speaker was considered feminine (83.12 when 100 = feminine, 16.33 when 1 = feminine). The male speaker was considered just as masculine as the female speaker was considered feminine (i.e., the difference between the maximum or minimum of the rating scale and the average stereotypy rating was similar for the male and female speaker; $F(1, 444) = 3.05$, $p = .09$; see Table 2). Male participants tended to rate the speakers as less masculine/feminine than the female participants ($F(1, 444) = 125.99$, $p < .001$), especially when rating the female speaker (interaction between participant gender and speaker gender: $F(1, 444) = 9.91$, $p = .002$). But we do not explore these differences here because it is beyond the scope of the paper and (as we shall see) we found no evidence for egocentric prediction in Experiment 1, suggesting these differences in ratings could not explain our effects. These differences are not relevant for Experiments 2 and 3 because the speaker's voice is not important for determining the consistent perspective.

2.1.3. Design

Speaker gender was manipulated within items and participants. As noted, there were two versions of each item: one produced by a male speaker (e.g., *I would like to wear the nice tie*) and one produced by a female speaker (e.g., *I would like to wear the nice dress*). Participants were assigned to one of two stimulus lists so that they heard only one version of each item, and heard: (1) 28 gendered sentences and 28 gender-neutral sentences, and (2) 14 sentences produced by a male speaker and 14 produced by a female speaker for each sentence type. In all lists, each object was shown twice: once as a target and once as a distractor.

For the gendered trials, each visual layout consisted of four pictures: (1) an agent-compatible target, which was an associate of the verb, was stereotypically compatible with the speaker's gender, and was referred to (e.g., dress when a female speaker said *I would like to wear the nice dress*); (2) an agent-incompatible target, which was an associate of the verb, but was incompatible with the speaker's gender (e.g., tie); (3) an agent-compatible distractor, which was not an associate of the verb, but was compatible with the speaker's gender (e.g., hairdryer); and (4) an agent-incompatible distractor, which was not an associate of the verb, and was incompatible with the speaker's gender (e.g., drill). For the gender-neutral trials, participants saw two targets and two distractors, which were gender-neutral.

Twenty layout combinations (e.g., agent-compatible target top left, agent-incompatible target top right, agent-compatible distractor bottom left, agent-incompatible distractor bottom right) were used once, and four randomly selected layouts were used twice. Note that the layout for the visual scenes changed depending on speaker gender. For example, if the dress (agent-compatible target) appeared in the top left when produced by the female speaker, then the tie also appeared in the top left when produced by the male speaker.

2.1.4. Procedure

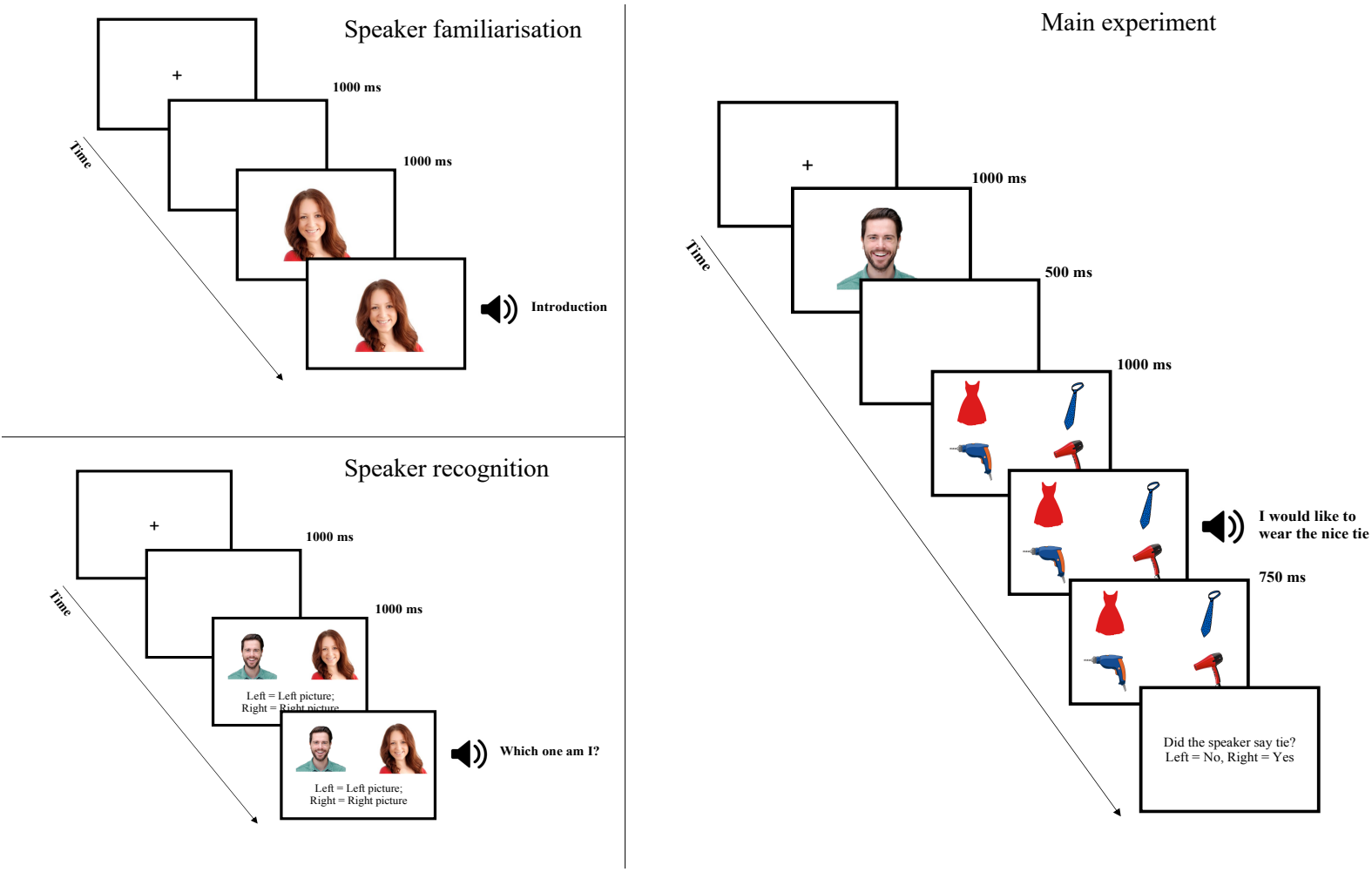
Participants were seated in front of a 1024 x 768 pixel monitor and were instructed to listen to the sentences and look at the accompanying pictures. Their eye movements were recorded using an EyeLink 1000 Tower mount eye-tracker sampling at 1000 Hz from the right eye. After reading the instructions, participants placed their head on the chin rest and the eye-tracker was calibrated using a nine-point calibration grid.

Before beginning the experiment, participants were familiarized with the two speakers (see Figure 1). They were told that one speaker was female, while the other speaker was male. Participants heard each speaker introduce themselves once (with order counterbalanced

across participants) by saying “Hi, I am Sarah/Andrew and you are going to hear me describe some objects. Please listen carefully and look at the objects on-screen”. (*Sarah* and *Andrew* are common names of clear stereotypical gender and of similar length). While listening to this introduction, participants saw a picture of the speaker (at a size of 300 x 300 pixels) displayed on the screen. This picture was displayed from 1000 ms before audio onset until audio offset.

Figure 1.

Schematic representation of the procedure for each phase in Experiment 1.



After familiarizing participants with each speaker, we ensured they could easily recognise the speakers by asking them to identify the picture of the speaker from their voice.

Participants heard each speaker ask “Which one am I?” once (with order counterbalanced) and saw both speakers’ pictures displayed in the center of the screen (one on the left and one on the right, counterbalanced across participants). Participants indicated via button-press response (left button for the speaker on the left; right button for the speaker on the right) which picture corresponded to the heard speaker. Participants always correctly identified the speaker from their voice.

In the main experiment, each trial started with a drift correct, followed by a 300 x 300 pixel picture of the speaker displayed in the center of the screen for 1000 ms. A blank screen was then displayed for 500 ms and the four pictures were presented in each of the four corners of the screen. Sentence playback began 1000 ms later (e.g., Ito, Corley, & Pickering, 2018), and the pictures remained on-screen for 750 ms after sentence end. Participants then answered a comprehension question, which asked if the speaker referred to a particular object (e.g., *Did the speaker say hairdryer*? *Left = No, Right = Yes*). Half of the time, the comprehension question mentioned an object the speaker had referred to; the other half of the time, the question referred to one of the other three unmentioned objects. Participants pressed the left button on the response box to answer yes, and right to answer no, and the next trial then began immediately (without feedback). Participants completed four practice trials and were given the opportunity to take a break after 28 experimental trials.

2.2. Data Analysis

We analysed the eye-tracking data in RStudio (version 1.2.5042). The fixations on the four pictures were coded binomially (fixated = 1; not fixated = 0; e.g., Ito, 2019) for each 50 ms bin from 1000 ms before to 1500 ms after verb onset. Fixations were regarded as falling on a picture if they fell in the area of 300 x 300 pixels around the picture. Blinks and fixations outside the interest areas were coded as 0 (i.e., no fixation on any of the objects) and

were included in the data. We analysed the gendered trials to determine: (1) whether participants predicted associatively (e.g., looking at wearable objects after hearing the verb *wear*); (2) whether they predicted consistently (or egocentrically), fixating the target stereotypically compatible with the speaker's gender over the target stereotypically consistent with their own gender (or vice versa); and (3) whether associative prediction occurred before consistent prediction (in accord with a two-stage account) or not (in accord with a one-stage account).

There are a number of ways we could analyse our data. One possibility is to compare the fixations to each object at each timepoint. In particular, we fitted Bayesian generalized linear mixed effects models (GLMM), in which fixations were predicted by Image Type, to every 50 ms bin from 1000 ms before to 1500 ms after verb onset. Image Type was dummy-coded, with agent-compatible target as the reference level. Thus, we could determine whether participants predicted associatively (agent-compatible target vs. agent-compatible distractor) and consistently (agent-compatible target vs. agent-incompatible target). We first fitted generalized linear mixed effects models (Baayen, Davidson, & Bates, 2008) using the *glmer* function of the *lme4* package (version 1.1-21; Bates, Maechler, Bolker, & Walker, 2019) with a binomial family, but these models produced singular fit errors even when we used the simplest random effects structure. As a result, we instead fitted Bayesian generalized linear mixed effects models using the *bglmer* function of the *blme* package (version 1.0-4; Chung, Rabe-Hsketh, Dorie, Gelman, & Liu, 2013), with a binomial family, the default priors, and the nlminbwrap optimizer. To summarise, this analysis showed that participants predicted associatively from 450 ms after verb onset and consistently from 600 ms after verb onset. Note that there was no indication of predictive looks before verb onset.

One problem with this analysis, however, is that fixations are non-independent: Both the target and distractor are on-screen, and the participant cannot simultaneously fixate both

objects at the same time. We could address this issue by transforming fixation proportions and calculating the ratio between the log odds of looking towards the agent-compatible target and the log odds of looking to the agent-compatible distractor. By fitting one-sample t-tests to every 50 ms time bin, we could compare the log ratios to 0, with a ratio greater than 0 indicating bias towards looking at the agent-compatible target over the agent-compatible distractor. However, both this approach and the binning analysis involves fitting as many models as there are timepoints (51 in this case), which increases the chance of Type 1 error (Hochberg & Tamhane, 1987). There is also the issue of autocorrelation: The eye-tracker records a fixation once per millisecond, but fixations tend to last for hundreds of milliseconds (e.g., Rayner, 1998). As a result, neighbouring bins are highly correlated.

For these reasons, we focus our interpretation on a bootstrapping analysis, which identifies the time point at which looks to one object (e.g., the agent-compatible target) diverged from looks to another (e.g., the agent-compatible distractor; Stone, Lago, & Schad, 2020). Our analysis procedure is identical to that used by Stone et al., but we summarise it here for clarity. The analysis involves three steps. First, we apply a one-sample *t*-test to fixation proportions at each timepoint, aggregating over items. Average fixation proportions are compared to .50, with a significant *p* value indicating that the object attracted more than half of the fixations. Second, a divergence point is identified by determining the first significant timepoint in a run of at least ten consecutive significant timepoints. Third, new datasets are generated 2000 times using a non-parametric bootstrap, which resamples data from the original data set using the categories participant, timepoint, and image type (e.g., agent compatible-target vs. agent-compatible distractor). A new divergence point is estimated after each resample, and the mean is calculated. Confidence intervals (CIs) indicate variability around the average divergence point. Note that the strong autocorrelation structure

present in the data is preserved during resampling, because resampling occurs within timepoints rather than between them.

In our first analysis, we compared fixation proportions to the agent-compatible target versus fixations to the agent-compatible distractor to determine whether participants predicted associatively. We then determined whether participants predicted consistently by comparing fixation proportions to the agent-compatible target versus fixations to the agent-incompatible target. Figure 2 suggests participants did not prefer one object over any of the others before verb onset, and the binning analysis did not show any significant effects before verb onset, and so we ran the divergence point bootstrapping analysis from verb onset (0 ms) to 1500 ms after verb onset.

To preview our results, our analysis showed that participants fixated the agent-compatible target more than the agent-incompatible target, thus suggesting they predicted consistently. But this analysis was based on all the gendered trials – that is, those in which the participant and the speaker had the same gender (the gender-match trials) and those in which they had different genders (the gender-mismatch trials). If participants simply predicted egocentrically (from their own perspective), then we would have expected no difference in looks to the agent-compatible and agent-incompatible targets in our collapsed analysis because egocentric prediction would be correct half of the time (the gender-match trials), but incorrect the other half of the time (the gender-mismatch trials). We did not observe this pattern of effects, and it is clear from Figure 2 that there was no stage at which participants fixated the agent-incompatible target more than the agent-compatible target. However, it is possible that participants were initially egocentric in their predictions, but the egocentricity effect was drowned out by the larger consistency effect. We tested this possibility in a third analysis, focusing on the gender-mismatch trials. In particular, we compared looks to the agent-compatible target versus the agent-incompatible target to determine whether there was

any stage at which participants predicted egocentrically. Comparing these trials to the match trials is not necessary for testing our predictions, given that the match trials do not allow us to isolate egocentric and consistent effects. But for the sake of completeness, we conducted an identical analysis for the gender-match trials.

In our final analysis, our goal was to determine whether associative prediction occurred before consistent prediction (in accord with a two-stage account) or not (in accord with a one-stage account). Associative predictions are indexed by the difference (at any relevant time point) between fixations to the agent-compatible target and fixations to the agent-compatible distractors. Consistent predictions are indexed by the difference between fixations to the agent-compatible target and fixations to the agent-incompatible target. To compare between groups, we bootstrapped the difference between their divergence points. In particular, we subtracted the onset of the associative effect from the onset of the consistent effect, following the same procedure as Stone et al. (2020). Our gender-mismatch analysis suggested that there was no point at which participants predicted egocentrically, and so we calculated a difference for all of the gendered trials, regardless of whether the speaker and participant had same or different genders.

These analyses are concerned with the gendered trials, since these are critical for testing consistent and egocentric prediction. But we also analysed the gender-neutral trials to look for further evidence for associative prediction. In particular, we compared looks to the two targets to looks to the two distractors. We focus our interpretation on the divergence point analysis; results from the GLMM analysis and log-ratio *t*-tests are reported in footnotes for the interested reader. But not that we could not run comparable difference analyses using GLMMS or log ratio *t*-tests because this analysis rests on calculating the difference between bootstrap distributions. Raw data and scripts for all analyses are available on Open Science Framework at: https://osf.io/nkud5/

2.3. Results

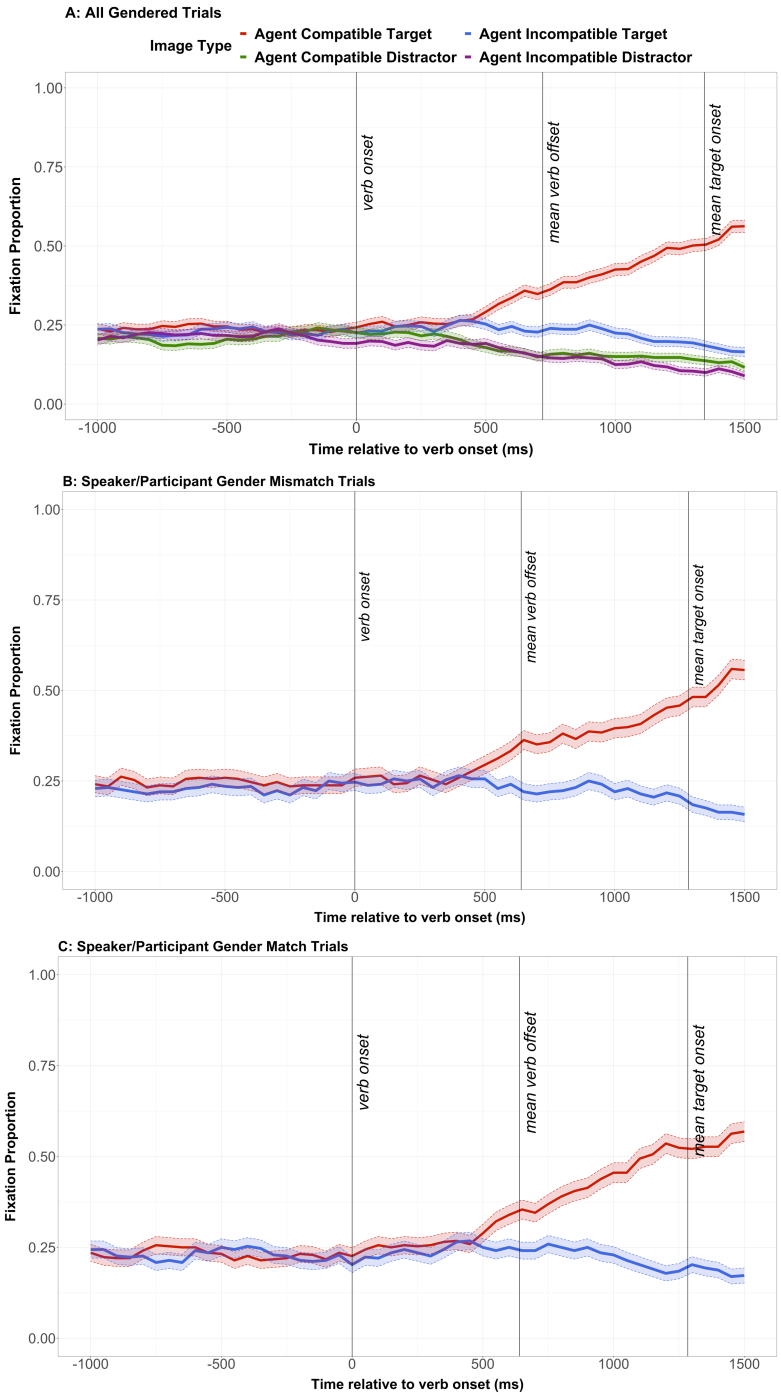
2.3.1. Comprehension question accuracy

The mean accuracy for the comprehension questions in all trials was 98%.

2.3.2. Eye-tracking data

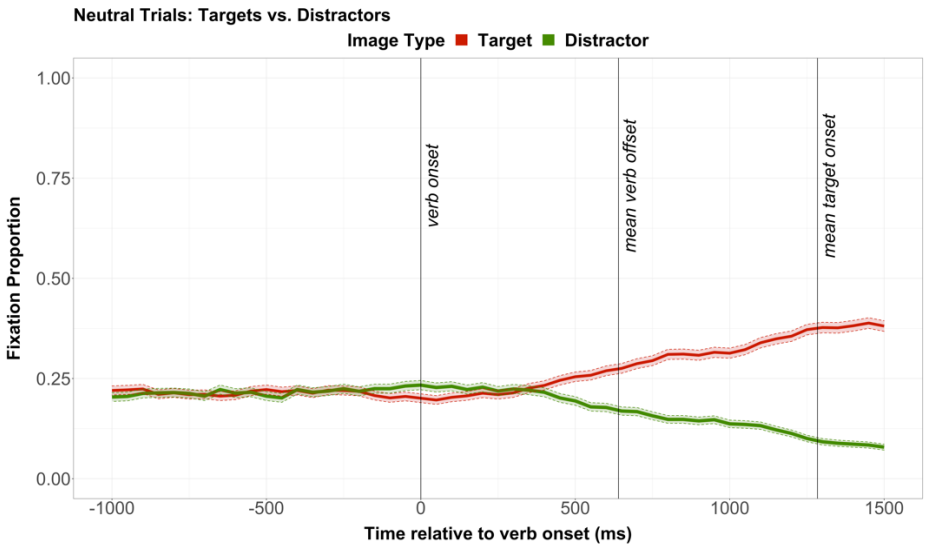
Figure 2 shows the mean fixation proportions on the four pictures for the gendered sentences (panel A), which is then divided into the mean fixation proportions on agent-compatible targets and agent-incompatible targets for the gender-match (panel B) and gender-mismatch (panel C) trials. Time was synchronized to verb onset, and the graph shows the time window from 1000 ms before to 1500 ms after verb onset.

Figure 2. Eye-tracking results for the gendered trials in Experiment 1. Panel A shows the mean fixation proportions on the four pictures for all gendered trials. Panels B and C show the mean fixation proportions on agent-compatible and agent-incompatible targets for the gender-match trials (speaker and participant have different gender; panel B) and the gender-mismatch trials (speaker and participant have same gender; panel C). Transparent thick lines are error bars representing standard errors.



The bootstrapping analysis showed that participants fixated the agent-compatible target more than the agent-compatible distractor from 519 ms after verb onset (CI[500, 650]; red vs. green in Figure 2A). The CI does not contain zero, and thus supports a reliable difference between the two objects.[1] We observed a similar pattern for the gender-neutral trials: Participants fixated the two targets more than the two distractors from 523 ms (CI[500, 650]; Figure 3). Note that this analysis compares looks to the two targets to the two distractors, while the analysis of the gendered trials compares looks to one target to one distractor. Nevertheless, these results suggest that participants predicted associatively.[2]

Figure 3. Eye-tracking results for the gender-neutral trials in Experiment 1. Transparent thick lines are error bars representing standard errors.



[1] The GLMM analysis showed that participants fixated the agent-compatible target more than the agent-compatible distractor from 450 ms after verb onset. The log-ratio *t*-tests showed a difference from 400 ms.

[2] The GLMM analysis showed that participants fixated targets over distractors from 550 ms after verb onset; the log-ratio *t*-tests showed a difference from 450 ms.

Participants also predicted consistently, fixating the agent-compatible target (which the speaker actually referred to and was compatible with the speaker's gender) more than the agent-incompatible target (which the speaker did not refer to) from 641 ms after verb onset (CI[600, 950]; red vs. blue in Figure 2A). Our analysis of the gender-mismatch trials (Figure 2B) confirmed that there was no point at which participants predicted egocentrically: They fixated the agent-compatible target more than the agent-incompatible (and egocentric) target from 651 ms (CI[450, 850]), and there was no point at which they fixated the agent-incompatible target more than the agent-compatible target. Moreover, these findings were essentially replicated in the gender-match trials (Figure 2C): Participants fixated the agent-compatible target more than the agent-incompatible target from 628 ms (CI[500, 850]), and there was no point at which they fixated the agent-incompatible target more than the agent-compatible target. Thus, participants predicted consistently from the speaker's perspective, looking at the stereotypically feminine target (e.g., the dress) when they heard a female speaker say *I would like to wear the nice…,* but at the stereotypically masculine target (e.g., the tie) when they heard the same sentence produced by a male speaker.[3]

Figure 2A suggests that consistent predictions tended to occur later than the associative predictions. We tested this difference significantly by subtracting the onset of the associative effect from the onset of the consistent effect. This analysis showed that the mean difference in divergence points between the associative and consistent effects was 122 ms (CI[0, 350]), suggesting that the consistent effect occurred later than the associative effect.

[3] The GLMM analysis showed that participants fixated agent-compatible targets more than agent-incompatible targets from 600 ms for all trials, from 650 ms for the mismatch trials, and from 800 ms for the match trials. The log-ratio analysis showed a difference from 550 ms for all trials and for the match and mismatch trials separately.

Note, however, that the lower boundary of the confidence interval is zero, and therefore we regard the difference as marginal.

2.4. Discussion

In Experiment 1, we investigated whose perspective comprehenders predict from, and whether this process of perspective-taking requires time. We found that participants predicted associatively, rapidly showing increased looks to pictures semantically associated with critical verbs. For example, if participants heard *wear*, then they fixated pictures of a tie and a dress more than a drill and a hairdryer. These associative predictions showed a similar time-course for both the gendered (519 ms after verb onset) and gender-neutral trials (523 ms after verb onset). We also found that participants predicted consistently from 641 ms after verb onset: They fixated the agent-compatible target, which was stereotypically compatible with the speaker's gender, more than the agent-incompatible target, which was stereotypically compatible with their own gender. For example, participants who heard a female speaker say *wear* would fixate dress more than tie.

This consistent effect occurred marginally later than the associative effect. Our results therefore provide some evidence that listeners take perspective into account when predicting, but not from the earliest moments of prediction (though further evidence for this conclusion is necessary). This timing difference could occur because perspective-taking is cognitively demanding (e.g., Lin et al., 2010), and it takes participants time to integrate perspective into their predictions. We discuss the theoretical implications of this finding in more detail in the General Discussion.

In sum, Experiment 1 suggests that comprehenders predict consistently. We have assumed that they do so by taking the perspective of the agent of the sentence. The sentences used *I*, and so the agent corresponded to the speaker. However, it is also possible that

participants' looks were simply driven by the speaker's voice (or the associated face). In other words, they determined the speaker's gender and simply looked at pictures that were stereotypically consistent with it, regardless of whether or not they were plausible referents of the verb. One argument against this explanation is that listeners did not look at objects that were stereotypically compatible with the speaker's voice from the start of the sentence (e.g., the dress and the hairdryer when listening to the female speaker).

But to be more confident that participants' looks were not simply driven by the speaker's voice, we conducted Experiment 2 which tested whether comprehenders predict consistently when the consistent perspective is their own, rather than the speaker's. Experiment 2 thus separates effects of perspective-taking from effects of speaker gender. At the same time, it provides a further test of whether associative prediction takes place before consistent prediction.

3. Experiment 2

Experiment 2 was identical to Experiment 1, with the exception that the pronoun *I* was replaced with *You* and the speaker referred to the target that was stereotypically consistent with the participant's gender (see Table 1). Assuming that the participant interprets *You* as referring to him or herself, the consistent perspective is now the comprehender's. If participants predict consistently, and the effect in Experiment 1 is not simply an effect of speaker voice (and face), then we expect them to look at the target that is stereotypically consistent with their own gender (the agent-compatible target) more than the target stereotypically consistent with the speaker's gender (the agent-incompatible target). For example, a female participant should look at a dress more than a tie when she hears a male speaker say *You would like to wear the nice…*. We again considered the time-course of associative prediction and the time-course of consistent prediction. All participants identified

as either male or female (and the gender they were assigned at birth). Thus, participants who identified themselves as female (or male) should consider stereotypically feminine (or masculine) objects as compatible with their gender identity. Note that the experiment is not informative about egocentric versus consistent prediction because the participant and the sentence agent are the same, but we found no evidence for egocentricity in Experiment 1.

3.1. Method

3.1.1. Participants

Thirty-two further native English speakers (*M*age = 20.63, 16 males, 16 females, who all identified as the gender they were assigned at birth) at the University of Edinburgh participated on the same terms as Experiment 1. We initially recruited 24 participants, but our sample included more females than males. Rather than throwing out data to balance gender, we recruited more male participants.

3.1.2. Materials, design, and procedure

Experiment 2 used the same stimuli as Experiment 1, except that sentences began with *You* rather than *I*. There were thus four versions of each sentence: one produced by a male speaker who referred to a stereotypically feminine object (e.g., *You would like to wear the nice dress*), one produced by a male speaker who referred to a stereotypically masculine object (e.g., *You would like to wear the nice tie*), one produced by a female speaker who referred to a stereotypically masculine object, and one produced by a female speaker who referred to a stereotypically feminine object. Participants heard one version of each sentence, and heard the version that ended with the target stereotypically compatible with their gender (rather than the target stereotypically compatible with the speaker's gender, as in Experiment 1). Thus, only two of the four versions were relevant for the female participants (one

produced by a male speaker; one produced by a female speaker), and only two of the four versions were relevant for the male participants. Sentences were recorded by the same two speakers from Experiment 1 and were between 2048 and 3750 ms in duration. Sentences produced by the two speakers were matched for their duration, the onset and offset of the critical verb, and the onset of the targets (all *p*s > .19 in t-tests; see Table 4). The rest of the design and procedure was the same as Experiment 1.

Table 4.

The means (and standard deviations) of sentence duration, critical verb onset and offset, and target onset for the sentences produced by male and female speakers in Experiment 2.

Speaker Gender	Duration	Verb Onset	Verb Offset	Target Onset
Male	2784 (364)	1250 (347)	1542 (361)	2156 (386)
Female	2866 (315)	1293 (334)	1651 (320)	2262 (308)

3.2. Results

3.2.1. Comprehension question accuracy

The mean accuracy for the comprehension questions in all trials was 95%.

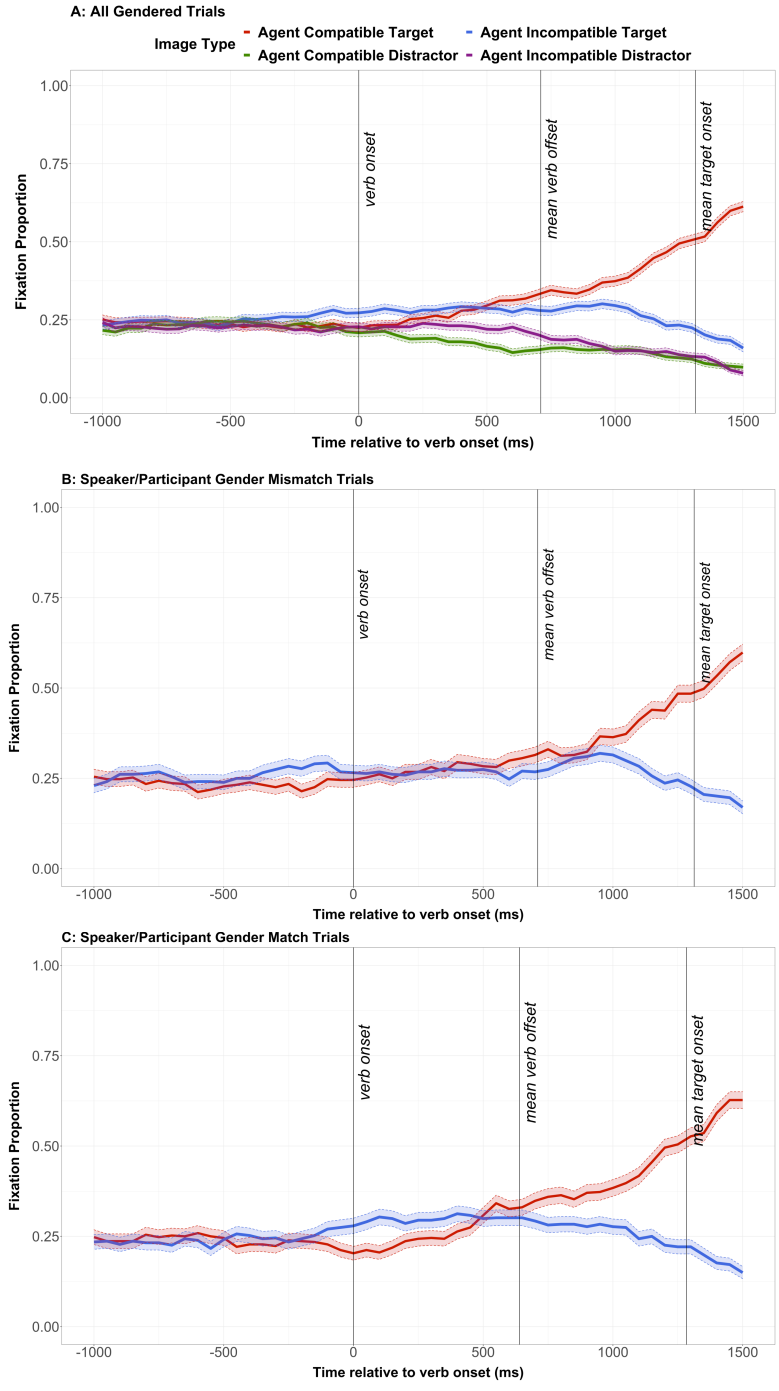
3.2.2. Eye-tracking data

The data were analysed as in Experiment 1, but we defined agent-compatible targets or distractors as those that were stereotypically compatible with the participant's gender, and agent-incompatible targets or distractors as those that were stereotypically incompatible with the participant's gender (and therefore stereotypically compatible with the speaker's gender for the gender-mismatch trials, but stereotypically incompatible with the speaker's gender for

the gender-match trials). In Experiment 1, we split the gendered trials into those where the speaker and participant had different genders (gender-mismatch trials) and those where they had the same gender (gender-match trials) to determine whether there was any point at which participants predicted egocentrically, from their own perspective. Here, however, the consistent perspective is the participant's, and so the split analysis does not allow us to test for egocentricity. But this split analysis does allow us to test whether the consistent effect in Experiment 1 occurred simply because participants fixated pictures stereotypically consistent with the speaker's gender. If so, on the gender-mismatch trials, participants should look at the agent-incompatible target (i.e., the speaker-compatible target) more than the agent-compatible target (i.e., the speaker-incompatible target).

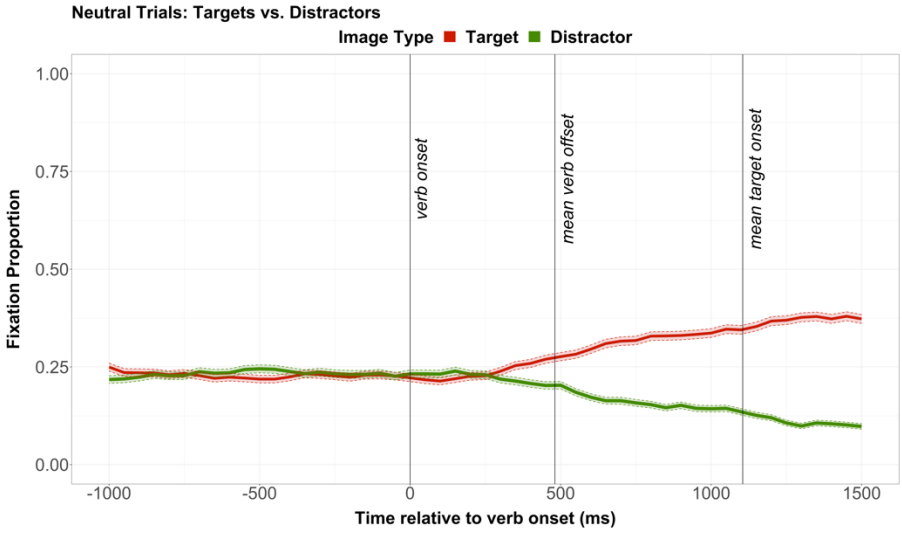
Figure 4 shows the mean fixation proportions on the four pictures for the gendered sentences (panel A), which is then divided into mean fixation proportions on agent-compatible targets and agent-incompatible targets for the gender mismatch (panel B) and gender match (panel C) trials.

Figure 4. Eye-tracking results for the gendered trials in Experiment 2. Panel A shows the mean fixation proportions on the four pictures for all gendered trials. Panels B and C show the mean fixation proportions on agent-compatible and agent-incompatible targets for the gender-match trials (speaker and participant have different gender; panel B) and the gender-mismatch trials (speaker and participant have same gender; panel C). Transparent thick lines are error bars representing standard errors.



As in with Experiment 1, the bootstrapping analysis for Experiment 2 suggests that participants used the verb to rapidly predict associatively. In particular, participants fixated the agent-compatible target more than the agent-compatible distractor from 329 ms after verb onset (CI[250, 500]; red vs. green in Figure 4A).[4] We observed a similar pattern for the gender-neutral trials: Participants fixated the two targets more than the two distractors from 447 ms (CI[400, 600]; Figure 5).[5]

Figure 5. Eye-tracking results for the gender-neutral trials in Experiment 2. Transparent thick lines are error bars representing standard errors.



Participants also predicted consistently, fixating the agent-compatible target (which the speaker actually referred to and was stereotypically compatible with the participant's

[4] The GLMM analysis showed that participants fixated the agent-compatible target more than the agent-compatible distractor from 300 ms after verb onset. The log-ratio *t*-tests showed a difference from 200 ms.

[5] The GLMM analysis showed that participants fixated targets over distractors from 550 ms after verb onset; the log-ratio *t*-tests showed a difference from 350 ms.

gender) more than the agent-incompatible target (which the speaker did not refer to) from 939 ms after verb onset (CI[800, 1050]; red vs. blue in Figure 4A). Our analysis of the gender-mismatch trials (Figure 4B) confirmed that there was no point at which participants predicted inconsistently: They fixated the agent-compatible (and egocentric) target more than the agent-incompatible target from 995 ms (CI[1000, 1050]), and there was no point at which they fixated the agent-incompatible target more than the agent-compatible target. Thus, participants did not simply hear the speaker's voice and look at pictures stereotypically compatible with the speaker's gender. Instead, they predicted consistently from their own perspective. For example, a female participant looked at the dress when she heard the speaker say *You would like to wear the nice…*, regardless of the speaker's gender. Our analysis of the gender-match trials (Figure 4C) essentially replicated these findings: Participants fixated the agent-compatible target more than the agent-incompatible target from 958 ms (CI[900, 1050]).[6]

The mean difference in divergence points between the associative and consistent effect was 611 ms (CI[400, 800]). Note that the confidence interval does not contain zero, and so provides strong evidence that the consistent effect occurred later than the associative effect.

[6] The GLMM analysis showed that participants fixated agent-compatible targets more than agent-incompatible targets from 1000 ms for all trials, from 1100 ms for the mismatch trials, and from 1000 ms for the match trials. The log-ratio analysis showed a difference from 900 ms for all trials and for the match and mismatch trials separately.

3.3. Discussion

In Experiment 2, we used sentences beginning with *You* rather than *I* (Experiment 1) to investigate whether comprehenders predict consistently even when these consistent predictions are tied to their own perspective. As in Experiment 1, participants rapidly predicted associatively, looking at targets semantically associated with the verb before the target was named. These associative predictions showed a similar time-course in both the gendered and gender-neutral sentences (329 ms after verb onset for the gendered sentences, and 447 ms for the gender-neutral sentences).

We also found that participants predicted consistently: They fixated the agent-compatible target, which was stereotypically compatible with their own gender, more than the agent-incompatible target, which was stereotypically compatible with the speaker's gender. For example, female participants who heard the sentence *You would like to wear the nice…* would fixate dress more than tie, regardless of the speaker's gender. Importantly, this finding suggests that the consistent predictions in Experiment 1 did not occur simply because participants heard the speaker's voice and fixated objects stereotypically compatible with the speaker's gender. Instead, these results suggest that participants consider the agent's perspective, thus predicting consistently.

The consistent effect occurred later than the associative effect, suggesting that predictions are initially driven by associations, and only subsequently are based on the agent's perspective. These findings are more conclusive than those of Experiment 1, in which the lower bound of the confidence interval comparing the time-course of the associative and consistent effect was zero. We return to this issue in the General Discussion, and discuss the potential differences between Experiments 1 and 2. Nevertheless, Experiment 2 provides further evidence that comprehenders initially predict associatively and subsequently predict consistently.

4. Experiment 3

In Experiment 3 we used the same stimuli as Experiment 1, except that the pronoun *I* was replaced with the name *James* or *Kate* (i.e., two clearly gendered and highly frequent names that would likely be very familiar to our participants). The speaker always referred to the target that was stereotypically compatible with the character's gender, and so participants heard sentences such as *Kate would like to wear the nice dress* or *James would like to wear the nice tie*. Thus, we could test whether comprehenders adopt the perspective of a third person and further separate consistent prediction from effects of speaker and comprehender gender. If listeners predict consistently, then we would expect them to look at the target that is stereotypically compatible with the character's gender (the agent-compatible target) more than the target compatible with their own gender (the agent-incompatible target). For example, participants should look at a picture of a dress more than a picture of a tie when they hear a speaker say *Kate would like to wear the nice…*.

This experiment has some similarity to Kamide et al.'s (2003) Experiment 2 (see *Evidence for a one-stage account*). They found that participants immediately (while hearing the verb) predicted consistently, fixating a picture of a motorbike more after hearing the sentence *The man will ride…* than after hearing *The girl will ride…*. Associative effects emerged later (while hearing the word following the verb): Participants fixated the motorbike more after hearing *The girl will ride…* than after hearing *The girl will taste…*. If we replicate their pattern of findings, we would expect participants to initially predict consistently, and ultimately predict associatively. However, this is not the pattern of results that we have observed in Experiments 1-2, where participants initially predicted associatively before ultimately predicting consistently. Thus, we compared the time-course of associative and consistent prediction, and also investigated whether egocentric prediction occurred.

4.1. Method

4.1.1. Participants

Thirty-two further native English speakers (*M*age = 22.69, 16 males, 16 females, who all identified as the gender they were assigned at birth) at the University of Edinburgh participated on the same terms as Experiment 1.

4.1.2. Materials, design, and procedure

Experiment 3 used the same stimuli as Experiment 1, except that sentences began with either *James* or *Kate* rather than *I*, and ended with the target that was associated with the verb and was stereotypically compatible with the character's gender (rather than the speaker's gender or the participant's gender; see Table 1). Sentences were recorded by the same two speakers from Experiment 1 and were between 1901 and 3774 ms in duration. Verb offsets were later for sentences produced by the female speaker than those produced by the male speaker ($F(1, 220) = 5.11$, $p = .02$; see Table 5). However, this difference is unlikely to affect the time-course of participants' predictions, because the prediction can start at verb onset. Furthermore, sentences produced by the two speakers were matched for their duration, the onset of the critical verb, and the onset of the targets (all *p*s > .05 in ANOVAs).

Table 5.

The means (and standard deviations) of sentence duration, critical verb onset and offset, and target onset for the experimental sentences in Experiment 3.

Speaker Gender	Character	Duration	Verb Onset	Verb Offset	Target Onset
Male	James	2782 (368)	1385 (362)	1645 (377)	2220 (382)
	Kate	2694 (347)	1325 (356)	1570 (362)	2121 (354)
Female	James	2824 (353)	1439 (335)	1725 (345)	2285 (377)
	Kate	2798 (350)	1427 (311)	1702 (318)	2247 (354)

Each speaker produced sentences involving both James and Kate, and so there were four versions of each item. Participants were randomly assigned to one of four stimulus lists, so they heard one condition per item, and heard: (1) 28 gendered sentences and 28 gender-neutral sentences, (2) 14 sentences about James produced by a male speaker and 14 produced by a female speaker for each sentence type, and (3) 14 sentences about Kate produced by a male speaker and 14 produced by a female speaker for each sentence type. In other respects, the procedure was as in Experiment 1.

4.2. Results

4.2.1. Comprehension question accuracy

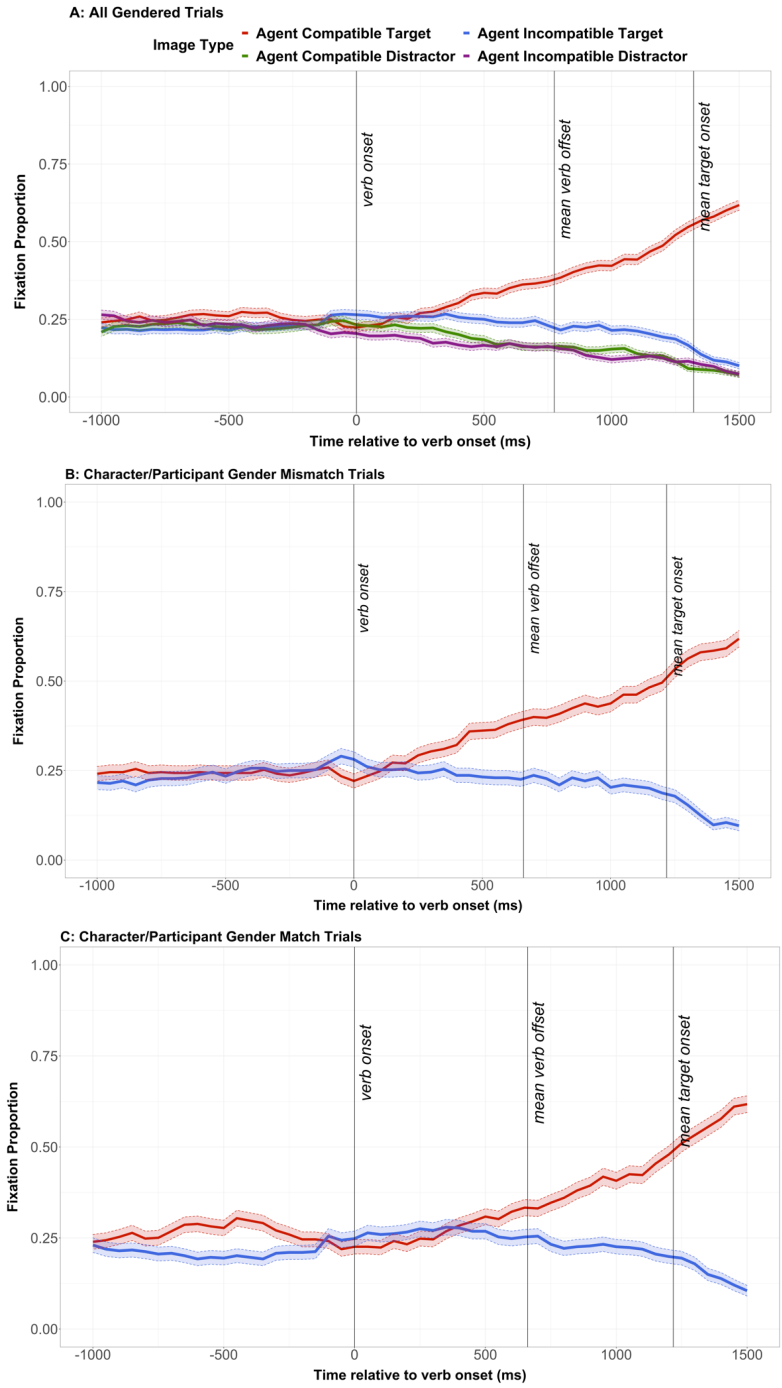
The mean accuracy for comprehension questions in all trials was 97%.

4.2.2. Eye-tracking data

The data were analysed as in Experiment 1, but agent-compatible targets or distractors were defined as those that were stereotypically compatible with the character's gender while agent-incompatible targets or distractors were those that were stereotypically incompatible

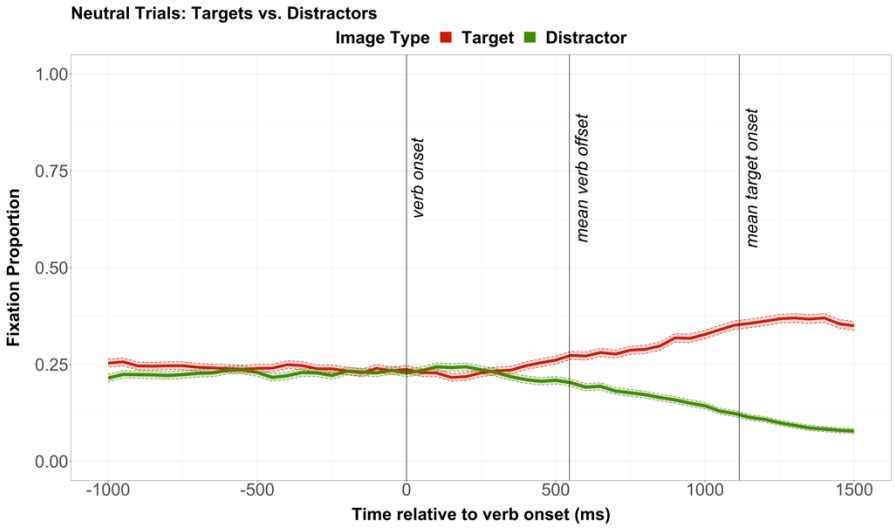
with the character's gender (and compatible with the participant's gender when analysing the gender-mismatch trials but incompatible with the participant's gender when analysing the gender-match trials). In addition, gender-mismatch trials were those where the participant and character had different genders, and gender-match trials were those where they had the same gender. Figure 6 shows the mean fixation proportions on the four pictures for the gendered sentences (panel A), which is then divided into mean fixation proportions on agent-compatible targets and agent-incompatible targets for the gender-mismatch (panel B) and gender-match (panel C) trials.

Figure 6. Eye-tracking results for the gendered trials in Experiment 3. Panel A shows the mean fixation proportions on the four pictures for all gendered trials. Panels B and C show the mean fixation proportions on agent-compatible and agent-incompatible targets for the gender-mismatch trials (character and participant have different gender; panel B) and the gender-match trials (character and participant have same gender; panel C). Transparent thick lines are error bars representing standard errors.



The bootstrapping analysis showed that participants fixated agent-compatible targets more than agent-compatible distractors from 384 ms after verb onset (CI[350, 500]; red vs. green in Figure 6A).[7] Participants also fixated targets more than distractors from 537 ms onwards for the gender-neutral trials (CI[450, 750]; Figure 7).[8] Together, these results suggest that participants predicted associatively.

Figure 7. Eye-tracking results for the gender-neutral trials in Experiment 3. Transparent thick lines are error bars representing standard errors.



Participants also predicted consistently: They fixated the agent-compatible target (which was stereotypically compatible with the character's gender) more than the agent-incompatible target (which was not) from 636 ms after verb onset (CI[500, 850]; red vs. blue

[7] The GLMM analysis showed that participants fixated the agent-compatible target more than the agent-compatible distractor from 350 ms after verb onset. The log-ratio *t*-tests showed a difference from 250 ms.

[8] The GLMM analysis showed that participants fixated targets over distractors from 550 ms after verb onset; the log-ratio *t*-tests showed a difference from 400 ms.

in Figure 6A). Our separate analysis of the gender-mismatch and gender-match trials confirmed that there was no point at which participants predicted egocentrically: They fixated the agent-compatible target more than the agent-incompatible (and egocentric) target from 583 ms for the gender mismatch trials (CI[450, 1000]; Figure 6B) and from 848 ms onwards for the gender-match trials (CI[800, 1050]; Figure 6C). Thus, participants predicted consistently from the character's perspective, for example fixating the dress when they heard the sentence *Kate would like to wear the nice…* but fixating the tie when they heard the same sentence beginning with *James*.[9]

As in Experiments 1 and 2, this consistent effect occurred later than the associative effect. In particular, the mean difference in divergence points between the associative and consistent effect was 252 ms (CI[500, 850]).

5. General Discussion

In three experiments, we used an eye-tracking task to contrast different accounts of prediction: a one-stage account, in which comprehenders predict consistently from the earliest moments of processing; and two two-stage accounts, in which comprehenders initially predict egocentrically or associatively, and predict consistently only at a later stage. To do so, we had male and female participants listen to male and female speakers producing sentences about stereotypically masculine and feminine objects that were displayed on-screen.

[9] The GLMM analysis showed that participants fixated agent-compatible targets more than agent-incompatible targets from 450 ms for all trials, from 450 ms for the mismatch trials, and from 750 ms for the match trials. The log-ratio analysis showed a difference from 450 ms for all trials, 600 ms for the match trials, and 400 ms for the mismatch trials separately.

In all three experiments, participants rapidly fixated objects semantically associated with critical verbs (e.g., hearing *wear* and fixating wearable objects), thus replicating previous research (e.g., Altmann & Kamide, 1999) and suggesting that comprehenders predicted associatively. Participants also predicted consistently, from the speaker's perspective in Experiment 1, from their own perspective in Experiment 2, and from a character's perspective in Experiment 3. In all three experiments, these consistent predictions occurred later than associative predictions (although this difference was marginal in Experiment 1). There was no evidence of egocentric prediction.

Our findings are incompatible with a one-stage account of prediction, which claims that comprehenders initially "step into the speaker's shoes" and make the best predictions they can from the earliest moments of processing. They are also incompatible with a two-stage account in which the first stage involves egocentric prediction. Instead, our results suggest that comprehenders initially predict associatively before subsequently predicting consistently.

These findings are compatible with the claim that perspective is one of many sources of information used to constrain processing (Heller et al., 2016). But this account claims that perspective can be integrated from the earliest moments of processing, which is not what we found. Instead, our findings support accounts of prediction that postulate multiple mechanisms (e.g., Huettig, 2015; Hintz, Meyer, & Huettig, 2017; Pickering & Gambi, 2018), in which comprehenders predict using multiple different sources of information over different time-courses. For example, Pickering and Gambi (2018) claimed that comprehenders predict associatively, which is characterised by very quick spreading activation between related concepts (e.g., Neely, 1977; Perea & Gotor, 1997). But these associative predictions tend to be error-prone (e.g., Kukona et al., 2011). As a result, comprehenders also make predictions with the processes they use to produce language (e.g., Levelt, 1989). In particular, they

covertly imitate what they have heard and derive the speaker's intention. They consider both linguistic and non-linguistic information (e.g., perspective) to adjust for differences between the speaker and themselves, which allows them to predict consistently. Comprehenders then run the derived intention through their own production system, retrieving at least some of the representations of the speaker's upcoming utterance, but stop short of actually speaking. Thus, consistent prediction is relatively slow (at least in comparison to prediction-by-association), because language production itself is relatively slow.

We noted that there is much evidence for immediate use of contextual information during language comprehension (e.g., Tanenhaus et al., 1994; Trueswell et al., 1995), including information relating to perspective (e.g., Hanna et al., 2003), and indeed relating to stereotypes associated with the speaker's voice (Van Berkum et al., 2008). These findings have been used to support a one-stage account of comprehension. But our findings suggest that their conclusions relate only to bottom-up aspects of comprehension and not to prediction. Specifically, we argue that comprehenders initially predict by association (in our case, by fixating objects associated with the verb), and then subsequently draw on other relevant information (here, relating to gender) at a second stage. In other words, perspective constrains top-down prediction but not bottom-up aspects of comprehension. In accord with this argument, Barr (2008) found that participants initially looked at objects visible to both themselves and their partner more than hidden objects, suggesting they predicted consistently. But Barr's participants also showed phonological interference effects from the competitor, regardless of whether it was visible or hidden, suggesting that perspective did not constrain bottom-up lexical processing.

Our findings are also incompatible with work showing that comprehenders initially predict consistently (e.g., Borovsky & Creel, 2014; Creel, 2012; Heller et al., 2009, Kamide et al. 2003). However, some of these studies used strong manipulations of perspective (such

as occluding objects) or familiarised participants with the speaker's preferences before comprehension, which greatly emphasised the importance of perspective and may have drowned out effects of associative (or, in theory, egocentric) prediction – though note that Kamide et al. did find a late evidence for associative prediction. In particular Borovsky and Creel found that participants initially fixated objects consistent with the speaker's identity; for example, if they heard the pirate speaking then they fixated the sword and ship more than the wand and the carriage. Once they heard the verb, they subsequently focused on the target object (the sword). We did not find this pattern of results in our experiments; in fact, we found no evidence that participants initially (i.e., before the verb) fixated objects just because they were stereotypically compatible with the speaker's gender. We propose that the explicit identification of the pirate as the speaker meant that participants rapidly fixated piratical objects, as it would be implausible for a pirate to refer to a wand or a carriage. But in our experiments, a female speaker is likely to refer to a tie or a dress.

Could the consistent effect have emerged later than the associative effect in our experiments because it takes longer to access gender-stereotyped information at the verb than it takes to access selectional restrictions (e.g., items that are wearable)? This explanation fits with a two-stage account of prediction, but the second stage would be limited in the information it uses (i.e., gender stereotypy). However, this explanation does not accord with the evidence that gender stereotyping occurs rapidly and automatically (e.g., Banaji & Hardin, 1996; Reynolds, Garnham, & Oakhill, 2006). Another theoretical possibility is that the consistent effect emerged later than the associative effects because participants may not have strongly believed in gender stereotypes, and so did not have a strong preference for objects stereotypically consistent with the speaker's, their own, or the character's gender. But our items were strongly stereotyped for the population of participants used in the experiment, and therefore that our participants should have adopted these stereotypes.

In our experiments, we investigated different types of consistent prediction. In particular, consistent prediction was (1) based on the speaker's perspective (Experiment 1); the participant's perspective (Experiment 2); or (3) a third character's perspective (Experiment 3). Previous studies of perspective-taking have simply looked at what happens when the perspectives of the speaker and the comprehender are in conflict, such as when the comprehender knows about objects that the speaker cannot see (e.g., Keysar et al., 2000). In contrast, our experiments demonstrate that comprehenders can weigh multiple perspectives (e.g., their own, the speaker's, and a third character's) and predict using whichever perspective is consistent. In other words, comprehenders are flexible in their perspective-taking and adopt the perspective that is most likely to lead to accurate comprehension.

It is worth noting that the difference in the time-course of associative and consistent prediction varied considerably across Experiment 1 (122 ms), Experiment 2 (611 ms) and Experiment 3 (252 ms). This small difference in Experiment 1 may have occurred because these associative predictions (519 ms) occurred later than those in Experiments 2 (329 ms) and 3 (252 ms). In principle, the participants in Experiment 1 may activated associates of the verb comparatively slowly. However, there is no reason to believe that these participants were different from those in Experiments 2 and 3.

Additionally, participants predicted consistently within 641 ms and 636 ms of the critical verb in Experiments 1 and 3, somewhat more quickly than participants in Experiment 2 (939 ms), though note we did not conduct any cross-experiment comparisons. In principle, the participants in Experiment 2 may not have strongly believed in gender stereotypes or not strongly identified as male or female. However, there is no reason to believe that these participants were different from those in Experiments 1 and 3. It is more likely that participants in Experiment 2 did not always or initially interpret the pronoun *You* as referring to themselves (e.g., Brunyé, Ditman, Mahoney, Augustyn, & Taylor, 2009) – they may have

sometimes interpreted *You* generically (meaning "one") or may not have regarded the recorded voice as addressing them (e.g., they might have believed they were hearing an utterance to another addressee).

Note also that some of our items were definitionally, rather than stereotypically, related to gender (e.g., *I really wanted to become a good princess/king*). It might be easier to predict consistently for the definitional items than for the stereotypical items. We did not test this prediction: This analysis would likely be underpowered because only seven of the 56 objects were definitionally feminine (five) or masculine (two). Moreover, definitional items are often related to stereotypy judgments – for example, it is possible for a female to say *I would like to become a good king*.

Our experiments also provide insight into the role of gender stereotyping during language processing. Previous research has demonstrated that participants consider stereotypes from a variety of domains, including gender, when comprehending what a speaker is saying (e.g., Van Berkum et al., 2008). Our experiments extend this research by demonstrating that comprehenders take stereotypes into account when predicting what a speaker is likely to say. Even though the consistent effects, which were based on gender-stereotyping, emerged later than associative effects, they still occurred before target onset, suggesting gender stereotypy had a rapid effect on prediction. Together, these findings suggest that comprehenders can rely on non-linguistic information to accurately predict and comprehend language.

In conclusion, we used the visual-world paradigm to demonstrate that comprehenders take perspective into account when predicting language, but do not do so from the earliest moments of prediction. In particular, we found that participants rapidly predict associatively, looking at semantic associates of a verb (e.g., *wear*) irrespective of whether they were consistent what the speaker is likely to say or not. Participants were slower to predict

consistently, from the speaker's perspective (Experiment 1), their own perspective (Experiment 2), and from a third character's perspective (Experiment 3); there was no evidence that they made inconsistent egocentric predictions. We conclude that prediction takes place in two stages during comprehension.

Acknowledgements

Ruth Corps was supported by the Economic and Social Research Council [grant number ES/J500136/1] and a Leverhulme Research Project Grant [RPG-2018-259] awarded to Martin Pickering. We thank Kate Corps and Nigel Corps for patiently recording the stimuli audio. We also thank Aine Ito for suggesting the bootstrapping analysis, and Kate Stone for suggesting how to calculate differences between groups.

References

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*, 247-264.

Bates D, Mächler M, Bolker B, Walker S (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67, 1–48.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, *59*, 390-412.

Banaji, M. R., & Hardin, C. D. (1996). Automatic stereotyping. *Psychological science*, *7*, 136-141.

Barr, D. J. (2008). Pragmatic expectations and linguistic evidence: Listeners anticipate but do not integrate common ground. *Cognition*, *109*, 18-40.

Barr, D. J. (2016). Visual world studies of conversational perspective taking. In P. Knoeferle, P. Pyykkönen-Klauck, & M. W. Crocker (Eds), *Visually situated language comprehension* (pp. 261-289). John Benjamins Publishing Company.

Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and clinical Neurophysiology, 60,* 343-355.

Borovsky, A. & Creel, S. C. (2014). Children and adults integrate talker and verb information in online processing. *Developmental Psychology, 50,* 1600-1613.

Borovsky, A., Elman, J. L., & Fernald, A. (2012). Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology*, *112*, 417-436.

Brown-Schmidt, S., Gunlogson, C., & Tanenhaus, M. K. (2008). Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition*, *107*, 1122-1134.

Brunyé, T. T., Ditman, T., Mahoney, C. R., Augustyn, J. S., & Taylor, H. A. (2009). When you and I share perspectives: Pronouns modulate perspective taking during narrative comprehension. *Psychological Science*, *20*, 27-32.

Carreiras, M., Garnham, A., Oakhill, J., & Cain, K. (1996). The use of stereotypical gender information in constructing a mental model: Evidence from English and Spanish. *The Quarterly Journal of Experimental Psychology Section A*, *49*, 639-663.

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika*, *78*, 685-709.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology, 6,* 84-107.

Corps, R. (2021, February 10). Prediction involves two stages: Evidence from visual-world eye-tracking. https://doi.org/10.17605/OSF.IO/NKUD5

Creel, S. C. (2012). Preschoolers' use of talker information in on-line comprehension. *Child Development, 83,* 2042-2056.

Damen, D., van Amelsvoort, M., van der Wijst, P., Pollmann, M., & Krahmer, E. (2021). Lifting the curse of knowledge: How feedback improves perspective-taking. *Quarterly Journal of Experimental Psychology, 74,* 1054-1069.

Damen, D., van der Wijst, P., van Amelsvoort, M., & Krahmer, E. (2020). Can the curse of knowledge be lifted? The influence of explicit perspective-focus instructions on readers' perspective-taking. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46,* 1407-1423.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology, 87,* 327-339.

Fodor, J. A. (1983). *The modularity of mind*. MIT press.

Frazier, L. (1987). Sentence processing: A tutorial review. In M. Coltheart (Ed.), *Attention and performance 12: The psychology of reading* (p. 559–586). Lawrence Erlbaum Associates, Inc.

Garnham, A., Oakhill, J., & Reynolds, D. (2002). Are inferences from stereotyped role names to characters' gender made elaboratively?. *Memory & Cognition*, *30*, 439-446.

Hanna, J. E., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cognitive Science*, *28*, 105-115.

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*, 43-61.

Heller, D., Grodner, D., & Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition, 108,* 831-836.

Heller, D., Parisien, C., & Stevenson, S. (2016). Perspective-taking behavior as the probabilistic weighing of multiple domains. *Cognition, 149,* 104-120.

Hintz, F., Meyer, A. S., & Huettig, F. (2017). Predictors of verb-mediated anticipatory eye movements in the visual world. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 1352-1374.

Hochberg, Y. & Tamhane, A. C. (1987). *Multiple comparison procedures.* John Wiley & Sons.

Huettig, F. (2015). Four central questions about prediction in language processing. *Brain Research*, *1626*, 118-135.

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

Hyde, J. S., Bigler, R. S., Joel, D., Tate, C. C., & van Anders, S. M. (2019). The future of sex and gender in psychology: Five challenges to the gender binary. *American Psychologist, 74,* 171-193.

Ito, A., Corley, M., & Pickering, M. J. (2018). A cognitive load delays predictive eye movements similarly during L1 and L2 comprehension. *Bilingualism*, *21*, 251-264.

Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and language*, *49*, 133-156.

Keysar, B. (1994). The illusory transparency of intention: Linguistic perspective taking in text. *Cognitive Psychology*, *26*, 165-208.

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*, 32-38.

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*, 1-20.

Kronmüller, E., Noveck, I., Rivera, N., Jaume-Guazzini, F., & Barr, D. (2017). The positive side of a negative reference: the delay between linguistic processing and common ground. *Royal Society Open Science*, *4*, https://doi.org/10.1098/rsos.160827.

Kukona, A., Cho, P. W., Magnuson, J. S., & Tabor, W. (2014). Lexical interference effects in sentence processing: Evidence from the visual world paradigm and self-organizing models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 326-347.

Kukona, A., Fang, S., Aicher, K. A., Chen, H., & Magnuson, J. S. (2011). The time course of anticipatory constraint integration. *Cognition*, *119*, 23-42.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

Lin, S., Keysar, B., & Epley, N. (2010). Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, *46*, 551-556.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological review*, *101*(4), 676-703.

Matuschek, H., Kliegel, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type 1 error and power in linear mixed models. *Journal of Memory and Language, 94,* 305-315.

Meyer, D. E. & Schvanevelt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90,* 227-234.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of experimental psychology: general*, *106*, 226-254.

Oakhill, J., Garnham, A., & Reynolds, D. (2005). Immediate activation of stereotypical gender information. *Memory & cognition*, *33*, 972-983.

Osterhout, L., Bersick, M., & McLaughlin, J. (1997). Brain potentials reflect violations of gender stereotypes. *Memory & Cognition*, *25*, 273-285.

Perea, M., & Gotor, A. (1997). Associative and semantic priming effects occur at very short stimulus-onset asynchronies in lexical decision and naming. *Cognition*, *62*, 223-240.

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*, *144*, 1002-1044.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124,* 372-422.

Rayner, K., Carlson, M., & Frazier, L. (1983). The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *Journal of verbal learning and verbal behavior*, *22*, 358-374.

Reynolds, D., Garnham, A., & Oakhill, J. (2006). Evidence of immediate activation of gender information from a social role name. *Quarterly Journal of Experimental Psychology*, *59*, 886-903.

Ryskin, R., Ng, S., Mimnaugh, K., Brown-Schmidt, S., & Federmeier, K. D. (2020). Talker-specific predictions during language processing. *Language, Cognition and Neuroscience, 35,* 797-812.

Sikos, L., Tomlinson, S. B., Heins, C., & Grodner, D. J. (2019). What do you know? ERP evidence for immediate use of common ground during online reference resolution. *Cognition, 182,* 275-285.

Stone, K., Lago, S., & Schad, D. J. (2020). Divergence point analyses of visual world data: applications to bilingual research. *Bilingualism: Language and Cognition,* https://doi.org/10.1017/S1366728920000607.

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, *48*, 542-562.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632-1634.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of memory and language*, *33*, 285-318.

Van Berkum, J. J., Van den Brink, D., Tesink, C. M., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of cognitive neuroscience*, *20*, 580-591.

Wardlow, L. (2013). Individual differences in speakers' perspective taking:The roles of executive control and working memory. Psychonomic Bulletin & Review, 20, 766–772.

Weingartner, K. M., & Klin, C. M. (2005). Perspective taking during reading: An on-line investigation of the illusory transparency of intention. *Memory & cognition*, *33*, 48-58.

Wu, S., Barr, D. J., Gann, T. M., & Keysar, B. (2013). How culture influences perspective taking: differences in correction, not integration. *Frontiers in human neuroscience*, *7*. https://dx.doi.org/10.3389/fnhum.2013.00822.

Appendices

Appendix A: Gendered and gender-neutral sentence fragments and target picture names used in Experiment 1. Predictable verbs are highlighted in **bold**.

Table A1: Gendered sentences used in Experiment 1. Experiments 2 and 3 used the same sentences and pictures, but *I* was replaced with *You* in Experiment 2 and with the name *James* or *Kate* in Experiment 3. The speaker always referred to the target stereotypically compatible with the agent's gender.

Sentence	Masculine Target	Feminine Target	Masculine Distractor	Feminine Distractor
I went to dinner last night and **wore** a nice	Shirt	Corset	Builder	Mermaid
I decided not to **wear** the nice	Turban	Makeup	Truck	Doll
I really wanted to **become** a good	King	Princess	Tie	Dress
I would really like to **buy** the nice	Barbeque	Roses	Mechanic	Cheerleader
I have decided to **buy** a nice	Wallet	Necklace	Firefighter	Ballerina
I have decided to **wear** the new	Belt	Perfume	Chainsaw	Tweezers
I once dreamed about **becoming** a nice	Knight	Nun	Waistcoat	Cardigan
Today, I will **wear** the new	Vest	Skirt	Hammer	Hairbrush
Later on, I will **use** a great	Drill	Hairdryer	Beer	Cocktail
Tonight, I will **wear** the nice	Cufflinks	Earrings	Digger	Pram
Later on today, I will **purchase** a nice	Kilt	Ring	Pirate	Witch
Later, I will go out and **buy** the great	Gun	Diamond	Plumber	Nurse
Tonight, it is likely I will **wear** a great	Tie	Dress	Drill	Hairdryer
I would really like to **drink** the nice	Beer	Cocktail	Turban	Makeup
Later, I am going to **use** the new	Urinal	Tampon	King	Princess

In the evening, I will **play** some good	Golf	Volleyball	Cufflinks	Earrings
I used to dream about **becoming** a great	Pirate	Witch	Wallet	Necklace
I had a dream about **becoming** a great	Builder	Mermaid	Vest	Skirt
When I go out, I will **carry** a nice	Briefcase	Handbag	Shirt	Corset
I have decided to **become** a good	Mechanic	Cheerleader	Kilt	Ring
I used to dream of **becoming** a great	Plumber	Nurse	Briefcase	Handbag
I would not like to **wear** the nice	Tuxedo	Earmuffs	Barbeque	Roses
I will go out and **buy** the nice	Hammer	Hairbrush	Knight	Nun
When I was younger, I liked to **push** the new	Digger	Pram	Urinal	Tampon
I used to enjoy **playing with** the nice	Truck	Doll	Belt	Perfume
I will go out and **help** the nice	Firefighter	Ballerina	Tuxedo	Earmuffs
Today, I would like to **wear** the nice	Waistcoat	Cardigan	Gun	Diamond
I have decided to **use** the nice	Chainsaw	Tweezers	Golf	Volleyball

Table A2: Gender-neutral sentences used in Experiment 1. Experiments 2 and 3 used the same sentences and pictures, but *I* was replaced with *You* in Experiment 2 and the name *James* or *Kate* in Experiment 3. The speaker randomly referred to one of the two targets, but this target was the same for a male and female speaker.

Sentence	Target 1	Target 2	Distractor 1	Distractor 2
Later on, I will **eat** the nice	Apple	Banana	Water	Milk
I am going to **eat** the nice	Cookie	Donut	Hoodie	Socks
I have decided that I will **wear** the great	Trainers	Wellies	Cake	Mushroom
I have decided to **eat** the nice	Kiwi	Carrot	Hat	Glasses
Later, it is likely that I will **eat** the nice	Bread	Pie	Bed	Toaster
I once thought about **becoming** a good	Dentist	Optician	Toothbrush	Pencil

I would like to **become** a great	Chef	Vet	Coffee	Tea
I have decided to **eat** some nice	Chocolate	Spaghetti	Tennis	Badminton
I would like to **eat** some good	Popcorn	Cereal	Headphones	Gloves
I am going to **feed** the nice	Parrot	Zebra	Poncho	Dungarees
I would like to **eat** a great	Pumpkin	Tomato	Jumper	Suitcase
I thought about **becoming** a great	Doctor	Photographer	Computer	Piano
Tomorrow, I will **visit** the nice	Pyramids	Volcano	Bread	Pie
I would like to **wear** the nice	Headphones	Gloves	Cookie	Donut
Today, I will **wear** the new	Hat	Glasses	Kiwi	Carrot
I would like to **drink** some great	Water	Milk	Chocolate	Spaghetti
This afternoon, I will **drink** a great	Coffee	Tea	Monkey	Tiger
I will go out later and **wear** the nice	Hoodie	Socks	Pumpkin	Tomato
I would like to **play** some great	Tennis	Badminton	Popcorn	Cereal
Later today, I will go out and **buy** a new	Bed	Toaster	Chef	Vet
I need to go out and **buy** a new	Jumper	Suitcase	Dentist	Optician
Later, I will **buy** a new	Computer	Piano	Doctor	Photographer
Tomorrow, I will **wear** the new	Poncho	Dungarees	Pancakes	Cheese
Tomorrow, it is likely that I will **eat** a nice	Cake	Mushroom	Parrot	Zebra
I have decided that I will **feed** the nice	Monkey	Tiger	Earplugs	Medal
I would like to **use** the nice	Toothbrush	Pencil	Pyramids	Volcano
I have decided to **wear** the nice	Medal	Earplugs	Apple	Banana
Later, I will **eat** the new	Pancakes	Cheese	Trainers	Wellies

Appendix B: Questionnaire used to collect gender information from participants.

Question	Answer
Age	
Gender	
Is your gender identity the same as you were assigned at birth?	
Native language (first language you learned to speak)	
Are you wearing glasses or contact lenses?	
Are you left or right handed?	

Please list any other languages you can speak or understand, and rate your ability in each language on a scale of 1 to 7 (7=high/4=moderate/1=low)

Language	Ability (write a number)

What do you think this experiment was about?

Have you heard about this experiment from anyone else? (If yes, please give details of what you've been told)
