GLOSSA psycholinguistics

Kandel, M., Wyatt, C. R., & Phillips, C. 2022. Agreement attraction error and timing profiles in continuous speech. *Glossa Psycholinguistics* 1(1): 4, pp. 1–46. DOI: https://doi.org/10.5070/G601157

eScholarship University of California

Agreement attraction error and timing profiles in continuous speech

Margaret Kandel, Harvard University, Cambridge, Massachusetts, 02138, USA, mkandel@g.harvard.edu

Cassidy R. Wyatt, University of Maryland, College Park, Maryland, 20742, USA, crwyatt@umd.edu

Colin Phillips, University of Maryland, College Park, Maryland, 20742, USA, colin@umd.edu

Studies of agreement attraction in language production have shown that speakers systematically produce verb agreement errors in the presence of a local noun whose features differ from that of the agreement controller. However, in attraction experiments, these errors only ever occur in a subset of trials. In the present study, we applied a naturalistic scene-description paradigm to investigate how attraction affects the distribution of errors and the time-course of correctly inflected verbs. We conducted our experiment both in the lab and in an unsupervised web-based setting. The results were strikingly similar across the experimental settings for both the error and timing analyses, demonstrating that it is possible to conduct production experiments via the internet with a high level of similarity to those done in the lab. The experiments replicated the basic number attraction effect, though they elicited comparable interference from both singular and plural local nouns, challenging common assumptions about a strong plural markedness effect in attraction. We observed slowdowns before correct verbs that paralleled the distribution of agreement errors, suggesting that the process resulting in attraction can be active even when no error is produced. Our results are easily captured by a model of agreement attraction in which errors arise at the point of computing agreement, rather than reflecting earlier errors made during initial encoding of the subject number.

1. Introduction

Systematic speech errors, such as agreement attraction errors, provide a window into the processes that underlie language production. Agreement attraction occurs when the process of agreement is disrupted by a nearby noun, as in (1) in which the verb agrees with the plural attractor *cabinets*, rather than the singular subject head *key*.

(1) *The key to the cabinets are on the table. (Bock & Miller, 1991)

Speech errors like that in (1) have been documented across languages in both experimental contexts and natural speech (e.g. Bock & Cutting, 1992; Bock & Eberhard, 1993; Bock & Miller, 1991; Den Dikken, 2001; Francis, 1986; Franck et al., 2002; Hartsuiker et al., 2003; Haskell et al., 2010; Pfau, 2009; Slioussar, 2018; Vigliocco et al., 1996), and grammatical facilitation occurs for attraction errors in comprehension tasks (e.g. Clifton et al., 1999; Dillon et al., 2013; Kaan, 2002; Lago et al., 2015; Pearlmutter et al., 1999; Shen et al., 2013; Tanner et al., 2014; Wagers et al., 2009).

There has been much attention in the literature on the factors that govern attraction effects and in what contexts errors are most likely to occur. Many studies have focused on number attraction errors (as in 1), though attraction effects have also been elicited for other linguistic features such as grammatical gender (e.g. Acuña-Fariña et al., 2014; Badecker & Kuminiak, 2007; Paspali & Marinis, 2020; Slioussar & Malko, 2016; Vigliocco & Franck, 1999). Number attraction is indexed by the presence of more agreement errors when the attractor has a different number (singular or plural) from the subject head (mismatch environments) than when they match in number (match environments).

A common finding in early number attraction studies was that agreement attraction errors occurred almost exclusively when the attractor noun was plural (e.g. Bock & Miller, 1991; Bock et al., 2004; Bock et al. 1999; Eberhard, 1993, 1997; but cf. Franck et al., 2002), a phenomenon known as the markedness effect. The contrast between plural and singular interference has been demonstrated across languages (e.g. Hartsuiker et al., 2003; Vigliocco et al., 1996; inter-alia) and displays analogous effects in comprehension (e.g. Almeida & Tucker, 2017; Dillon et al., 2017; Wagers et al., 2009). The markedness effect is taken to be so reliable that studies of agreement attraction have often only included conditions with singular heads (e.g. Bock et al., 2001; Brehm & Bock, 2013; Eberhard, 1999; Gillespie & Pearlmutter, 2011; Haskell & MacDonald, 2003; Solomon & Pearlmutter, 2004; Veenstra, Acheson, et al., 2014; inter alia in production; Franck et al., 2010; Lago et al., 2015; Schlueter et al., 2019; Tanner et al., 2014; inter alia in comprehension).

While prior research has led to fruitful discussion about the processes involved in subject–verb agreement formation, in the majority of studies, the inferences drawn are based on errors that only ever occur on a small subset of trials. In order to probe agreement formation more generally, not just in the cases when the process leads to error, the present study analyzes

both traditional speech error profiles as well as the production time-course of correct verbs in continuous speech. This analysis allows us to investigate whether the processes leading to attraction interference are active during verb planning even when no error is produced. The presence of a verb timing effect could indicate that number attraction results from pressures active during the agreement computation itself (i.e. the process of matching the subject and verb features) as opposed to earlier in sentence planning (e.g. specification of the subject features). We elicited responses using a scene-description task, a more naturalistic alternative to the traditional preamble elicitation paradigm (e.g. Bock & Miller, 1991). Given the speed and scalability of web-based research coupled with limits on in-person testing during the COVID-19 pandemic, we were additionally interested in whether suitable data for our analyses could be collected online. We therefore conducted both in-lab and web-based versions of our experiment. By analyzing errors and timing in concert within both experiments, we were able to observe a relationship between verb timing and accuracy, finding evidence of a speed-accuracy trade-off in attraction.

1.1 Models of agreement attraction

There are two model frameworks typically used to describe the processes underlying agreement attraction effects: representational accounts and retrieval accounts. These accounts have been applied to describe attraction effects in both production and comprehension.

Representational accounts attribute attraction effects to a faulty or ambiguous representation of the subject phrase number. This representation causes the incorrect number to be used when it is accessed by the generator to compute verb agreement (resulting in an error) or by the parser to check verb agreement (leading to grammatical illusion). Different accounts attribute the faulty or ambiguous subject number representation to different sources. Some accounts propose that an attractor's number feature interferes with the subject number representation through a featural encoding error caused by simultaneous activation of the attractor and subject head (Gillespie & Pearlmutter, 2011) or via upward percolation of the attractor's number feature through the syntactic structure (Bock & Eberhard, 1993; Franck et al., 2002; Nicol et al., 1997; Vigliocco & Nicol, 1998). The marking and morphing model (Eberhard et al., 2005) suggests that plurality lies on a continuum and that the plurality of a subject phrase is dependent on the number features of all the nouns it contains (in addition to notional number factors such as collectivity). Under this model, attraction effects arise probabilistically when the subject head and attractor differ in number, causing the plural value for the subject phrase to be more ambiguous than when they have the same notional and grammatical number.

Retrieval models of attraction (e.g. Badecker & Kuminiak, 2007; Dillon et al., 2013; Wagers et al., 2009; inter alia) attribute effects to faulty retrieval of the subject during feature matching, rather than to an issue with the subject number representation itself. In a retrieval framework, the subject is retrieved from content-addressable memory (McElree, 2000; McElree et al., 2003) when

computing verb agreement (in production) or when checking verb agreement (in comprehension). The subject representation is accessed via cue-based memory retrieval, which picks out the item in memory that best matches a set of retrieval cues. Retrieval accounts propose that this process can go awry when an attractor partially matches the set of cues used for retrieval, causing the attractor to become activated during retrieval and, in some cases, erroneously accessed and used for the agreement computation in place of the subject head.

Although retrieval has been used to describe attraction effects in both comprehension and production, the cues available to pick out the subject differ in each task (see Slioussar & Malko, 2016 for a similar observation). In comprehension, the morphosyntactic features of the verb are available to be used as retrieval cues. In fact, verb number is typically assumed to be a retrieval cue for agreement checking (e.g. Wagers, 2008; Wagers et al., 2009). In production, on the other hand, where the speaker uses retrieval to generate a verb form rather than to check its features, there is no verb number feature available to be used as a retrieval cue. The process must instead rely on other cues to pick out the agreement controller, such as category, case, or position-based features.

While many models of attraction fall into one of these two categories, others have proposed hybrid frameworks combining elements of representational and retrieval accounts. For example, self-organizing sentence processing (SOSP) models attribute attraction effects to structural encoding errors as the wrong NP is linked to the subject position attached to the verb through a linking process driven by feature matching (e.g. Smith et al., 2018; Smith et al., 2021; Villata et al., 2018). The competition model (Nozari & Omaki, 2022) models agreement production as a lexical selection process, in which the morphosyntactic features of preceding nouns (including number) activate candidate verb forms. The activation of candidate forms is proportional to the activation of the features they match (with more recent features being more active). When the subject head and attractor differ in number, both singular and plural verb forms become activated, increasing the likelihood of errors.

In the present study, we do not attempt to arbitrate between these different model frameworks but rather focus on a key dimension that distinguishes accounts of attraction: the locus of the attraction effect. In many representational accounts, agreement errors reflect a problem in the encoding of the subject number. In these accounts, the issue leading to attraction errors is independent from the agreement computation (i.e. the matching of the subject and verb features): verb agreement merely provides an opportunity to observe the presence of an error earlier in sentence processing. In other approaches (such as retrieval-based accounts, a lexical competition model, and some instantiations of marking and morphing), the error arises from the agreement computation itself – that is, the error arises when the generator references the subject to match its features. This distinction is orthogonal to other differences between representational and retrieval accounts, such as whether agreement attraction is always mediated by the

subject representation or whether attractors influence the computation of agreement directly. Investigating the relationship between verb errors and timing can provide insight into when in sentence planning the process leading to attraction occurs.

If the pressure leading to errors is active during the agreement computation (e.g. resolving retrieval/lexical competition or deciding how to interpret an ambiguous subject number phrase), we may expect to see this reflected in the production time-course of the verb, potentially resulting in pre-verbal slowdowns in the same environments where errors are elicited, even when the correct verb is produced. On the other hand, if agreement errors are a transparent reflection of a process that went wrong at an earlier point of encoding, then the pressure caused by a mismatching attractor should not influence verb agreement directly, meaning that we are less likely to see timing effects at the verb caused by the agreement computation itself. We may, however, still expect to see slowdowns resulting from other processes such as self-monitoring; we discuss this possibility in the *General Discussion*.

1.2 Looking for attraction effects in continuous speech

Our study used a scene-description task to elicit subject–verb agreement in continuous speech, following a recent trend to use more naturalistic tasks to elicit attraction errors (Nozari & Omaki, 2022; Veenstra, Acheson, & Meyer, 2014). Many previous agreement attraction experiments have used preamble paradigms to elicit errors in participant speech (e.g. Bock & Cutting, 1992; Bock & Eberhard, 1993; Bock & Miller, 1991). These tasks present participants with a sentence fragment that they need to repeat and complete as a full sentence, thereby combining both comprehension and production, as participants need to parse the preamble fragment before planning their response.

In our task, on the other hand, participants saw simple scenes and were asked to describe them. The participants thus produced sentences that we could analyze for attraction effects without being provided any pre-packaged linguistic material for their responses or needing to interpret, remember, and repeat a part of a sentence they had heard or read. This task involves a clear mapping from the message (the events of the scene) to a sentence, which allowed us to probe the sentence planning process in a more naturalistic setting than preamble paradigms, in which speakers may have weaker access to the message-level representations of the sentences they produce.

While other studies have shown the efficacy of description tasks to elicit agreement attraction errors (Nozari & Omaki, 2022; Veenstra, Acheson, & Meyer, 2014, Experiment 1), ours is the first to apply this type of paradigm to study timing effects in the generation of correct agreement. A number of prior studies have found a relationship between response time and number attraction errors (e.g. Brehm & Bock, 2013; Haskell & MacDonald, 2003; Staub, 2009, 2010; Veenstra, Acheson, et al., 2014; Veenstra, Acheson, & Meyer, 2014, Experiment 2), however these studies

used tasks that rely on comprehension of a preamble as opposed to more naturalistic production. In our study, we applied a forced-aligner (McAuliffe et al., 2017) to look for pauses in articulation directly before production of correct verbs in continuous scene descriptions. The speech time-course observed in our task should primarily reflect production planning processes (in addition to visual parsing of the scenes prior to message-level planning) instead of also including time to comprehend and recall a provided preamble. Our study further differs from prior research by investigating timing effects utterance-medially within complete declarative sentences, as opposed to measuring button press response times (cf. Staub, 2009, 2010; Veenstra, Acheson, et al., 2014, Experiment 2; Veenstra, Acheson, & Meyer, 2014, Experiment 2) or the onset of sentence fragments (cf. Brehm & Bock, 2013; Veenstra, Acheson, et al., 2014, Experiment 1) or interrogative sentences (cf. Haskell & MacDonald, 2003). Our study also included conditions with both singular and plural heads (cf. Brehm & Bock, 2013; Haskell & MacDonald, 2003; Veenstra, Acheson, et al., 2014), enabling us to look for markedness asymmetry in speech timing effects.

1.3 Markedness asymmetry in agreement attraction

Several foundational preamble studies have elicited a binary markedness asymmetry in agreement attraction error effects, observing increased verb errors in mismatch environments with plural attractors but almost no attraction errors in those with singular attractors (e.g. Bock & Miller, 1991; Bock et al., 2004; Bock et al., 1999; Eberhard, 1993, 1997; but cf. Franck et al., 2002). Comparable effects have been observed for comprehension, with grammatical illusions or signs of processing facilitation virtually occurring only in the presence of plural attractors (Almeida & Tucker, 2017; Dillon et al., 2017; Wagers et al., 2009).

In the past, both representational and retrieval accounts of attraction have captured the absence of interference from singular attractors by assuming a privative number feature system, with singular number being the unmarked default. In representational models, a singular attractor cannot influence the representation of the subject phrase number because it does not have a number feature. Retrieval models have accounted for the markedness asymmetry by proposing that singular number cannot be used as a retrieval cue if it is the unmarked default and/or that the retrieval processes is biased to access marked representations (e.g. Badecker & Kuminiak, 2007; Wagers et al., 2009), making erroneous retrieval of an plural attractor more likely than erroneous retrieval of an singular attractor.

Although there have been demonstrations of clear contrasts between singular and plural attractors, recent studies using more naturalistic sentence elicitation paradigms have observed somewhat reduced markedness effects, eliciting errors in sentences with plural heads and singular attractors (the PS condition) as well as the singular heads and plural attractors (the SP condition) (Nozari & Omaki, 2022; Veenstra, Acheson, & Meyer, 2014). While their results still show evidence of a markedness effect, manifesting as stronger interference from plural attractors than

singular ones, the reduced asymmetry between singular and plural attraction could suggest that the markedness effect is graded instead of binary in more naturalistic speech. Response time data from Staub (2009, 2010) and Veenstra, Acheson, and Meyer (2014, Experiment 2) (who elicited the same target sentences from their picture-description experiment in a forced-choice task) provide further evidence for interference from singular attractors, eliciting slowdowns for both the SP and PS conditions compared to their match counterparts (SS and PP); in fact, Veenstra, Acheson, and Meyer (2014) observed no markedness effect in response times despite observing asymmetries for errors in both their forced-choice and picture-description experiments.

Our study can be used to further assess the reliability of interference from singular attractors in more naturalistic production, looking for attraction effects in both errors and speech timing. If it is possible to elicit reliable singular attraction, models of agreement attraction must adapt the way that singular number is represented and/or used such that it is able to cause interference. In particular, the privative number feature system often assumed within these frameworks may not be necessary to capture agreement attraction patterns.

1.4 Testing the feasibility of web-based production research

While several studies have shown the viability of online research for language science (e.g. Gibson & Federenko, 2013; Gibson et al., 2012; Sprouse, 2011), few have focused on production data in particular, especially open-ended speech. Most internet-based research has used tasks like linguistic judgments, self-paced reading, and typed sentence completions that involve straightforward measures such as button presses, survey questions, or typed responses (e.g. Corley & Scheepers, 2002; Linnman et al., 2006; Enochson & Culbertson, 2015; Skitka & Sargis, 2006; Vesker et al., 2019). Nevertheless, prior web-based production studies (e.g. Fairs & Strijkers, 2021; Vogt et al., 2021; Ziegler & Snedeker, 2018) have shown that it is also possible to collect speech recordings online. The present study adds to this literature, further testing the feasibility of collecting open-ended production data via the internet by conducting a side-by-side comparison of in-lab and web-based versions of our agreement attraction experiment. By analyzing both agreement errors as well as the timing of correct responses, our study serves as an appropriate test of whether web-collected data is suitable to analyze measures of the kind probed in production research (speech errors and articulation time-course) and to detect the types of effects typically studied by language production researchers.

There are many advantages of internet-based testing that would be beneficial to extend to language production research. Web-based experimentation creates more flexibility and efficiency in the data collection process, since data can be collected outside of lab hours, without specialized lab equipment, and without an experimenter present. These properties can make web-based testing methods more accessible and scalable than in-lab testing, providing the opportunity for faster data collection from more diverse participant samples. By making it more feasible to reach

populations further from researchers' home institutions, web-based testing has the potential to overcome limitations faced by in-lab testing, such as the strong pressure to use convenience samples of undergraduate students dominated by speakers of the language(s) spoken close to the researchers' institution (such samples are typically younger and more homogeneous than the broader population, and their behavior is not always representative; Henrich et al., 2010). Crucially for language researchers, web-based experimentation allows us to reach native speakers of additional languages and to increase the linguistic diversity of our samples. Moreover, the COVID-19 pandemic has recently limited in-person testing, increasing the value of internet-based alternatives.

Despite the potential benefits of transitioning to web-based research, there are also potential concerns about the accuracy and consistency of production data collection online. In particular, speech recording in web-based experimentation may be especially susceptible to limitations of web-based research, with previous literature identifying concerns about the precision of data recording in web-based experiments (Reips, 2002; Skitka & Sargis, 2006). These concerns can arise due to variations in participants' software, hardware, internet connections, and environment, as well as the lack of experimenter supervision. Speech recordings may be more variable or noisier when elicited and recorded outside of a controlled lab environment.

Subject–verb agreement attraction provides a useful test case for examining web-collected speech production data because it is a well-documented phenomenon in language production shown to present a mix of robust and more subtle effects. Given the prevalence of verb attraction errors in production experiments and the ease of their identification, we can easily test for the basic attraction error effect (the presence of more errors in mismatch environments) as well as whether it is possible to replicate the more specific shape of the effect, such as the size of the markedness asymmetry. Attempting to replicate speech timing effects in our online experiment provides a more rigorous test of the quality of web-collected data than the error analysis, requiring higher quality speech recordings and representing a more subtle measure to try to detect.

Our chosen paradigm and planned analyses provide an interesting test of web-based production beyond those of prior web-based production experiments we have encountered. Our scene-description task elicits more open-ended speech than the experiments conducted by Fairs and Strijkers (2021) and Vogt et al. (2021), which elicited picture names. Our study requires full sentence responses from participants in a consistent format, despite the lack of an experimenter present to explain the task, answer questions, and provide clarifications. Ziegler and Snedeker (2018) similarly elicited responses in a scene-description task, though our analyses additionally required audio data of sufficient quality to be submitted to a forced-aligner for fine-grained analyses of utterance-medial timing effects. This type of analysis may require higher quality recordings than are necessary for response transcriptions alone or for speech onset identification in picture naming.

The side-by-side comparison of our two experiments furthermore allowed us to observe the effect of different time pressures on agreement accuracy and pre-verbal pauses, as the response time window in the web-based experiment was lengthened to adapt to the online setting.

2 Methods

This study comprises two experiments investigating subject–verb agreement attraction in production. Experiment 1 was originally conducted in the lab as a part of a larger study executed prior to the onset of the COVID-19 pandemic (and the ensuing halts on in-person research) comparing subject–verb agreement with anaphor–antecedent agreement (Kandel & Phillips, 2022). This comparison was performed between participants and thus should have no bearing on the present results. Experiment 2 was a replication of Experiment 1, conducted in an unsupervised web-based setting. By utilizing an already-conducted in-lab experiment and moving it to the internet, we were able to compare two data collection methods at a time when only one was possible. Supplementary Materials are available from https://osf.io/jwnsz/.

2.1 Experiment 1 (in-lab experiment)

2.1.1 Participants

The participants for Experiment 1 were 45 native American English speakers (34F, 11M) from the University of Maryland community with an average age of 21.1 years (SD = 4.5y). We ran an additional four omitted participants. Two were omitted for not passing our native speaker test, one was omitted for not following task directions by speaking in incomplete sentences, and one was omitted because over 1/3 of their trials were excluded (see *Analysis* for trial exclusion criteria). This experiment was part of an hour-long session with an unrelated language comprehension study. For the full session, the participants were given course credit or monetary compensation ($12.00).

2.1.2 Materials

In the task, the participants saw scenes of animated aliens called greenies, blueys, and pinkies (**Figure 1a**). Participants were told that the aliens have an ability called *mimming*, during which the aliens' antennae light up. Each scene involved 1s of preview followed by 3s of mimming by a subset of aliens in the scene (**Figure 1b**). For each trial, participants were asked to describe the events of the scene using the verb *mim*. Our paradigm thus differs from previous description paradigms by asking participants to describe an action rather than a property of the subject referent (both Nozari & Omaki, 2022 and Veenstra, Acheson, & Meyer, 2014 elicited descriptions of the subject referent's color). The decision to elicit sentences with novel words was motivated by the planned comparison to reflexive attraction in Kandel and Phillips (2022), as it allowed us to flexibly alter the argument structure of the verb *mim* to elicit different linguistic constructions while holding the verb constant.

To create an environment suitable to induce attraction, the elicited descriptions contained a subject head (N1), a verb that marked agreement (*is/are*), and an intervening attractor (N2). Participants were trained to use the sentence frame *the + N1 + preposition + N2 + is/are + mimming* in their responses (e.g. *the pinky above the greenies is mimming*). The task used contrast in the scenes to elicit the prepositional phrases containing N2. Each scene contained two groups of aliens, which prompted participants to use prepositional modifiers to disambiguate which alien performed the action (e.g. instead of simply saying *the pinky is mimming*).

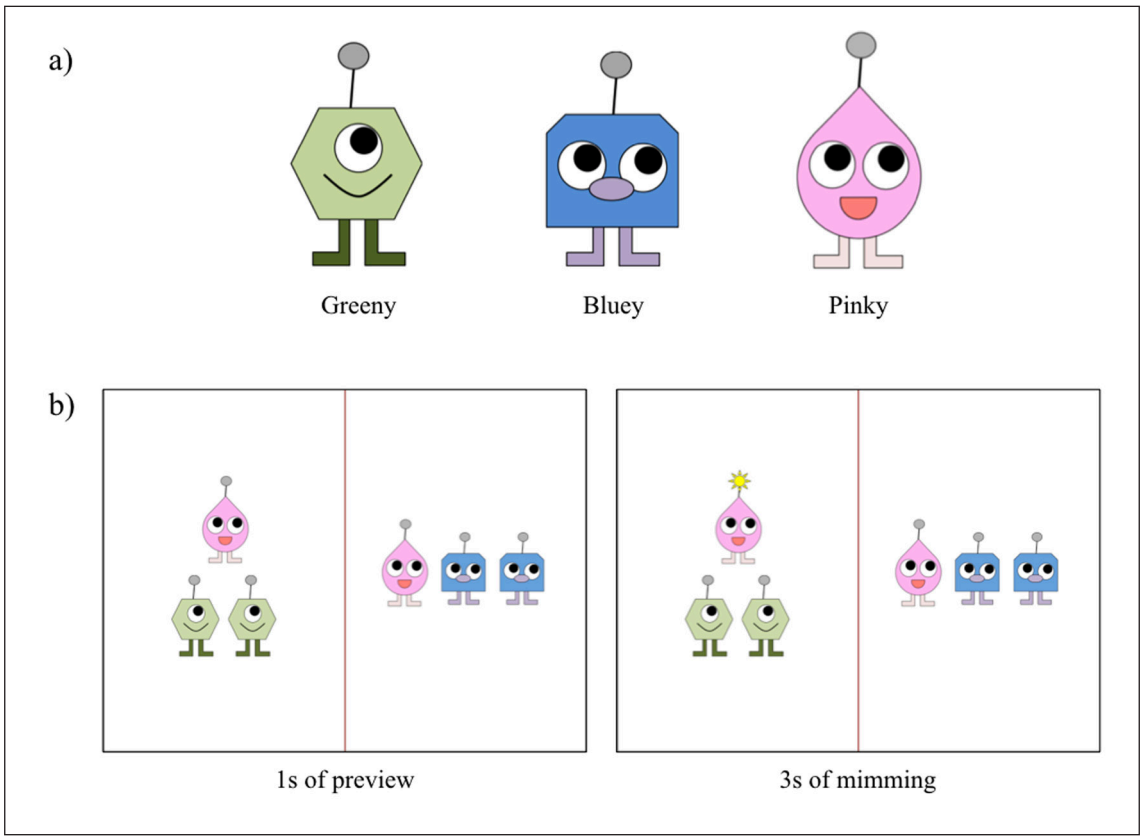

Figure 1: a) The aliens included in the stimuli. **b)** Example scene eliciting the target sentence *the pinky above the greenies is mimming*.

The scenes were constructed such that N1 and N2 in the target description either matched or mismatched in number. N1 and N2 could be either singular or plural, leading to four experimental conditions (**Table 1**). There were 24 target sentences per condition (96 total).

The target sentences were pseudo-randomized in four presentation lists (each eliciting the full set of 96 sentences) such that no more than two consecutive trials were the same condition, had the same match, elicited the same preposition, or involved the same pairing of aliens (blueys + pinkies, pinkies + greenies, greenies + blueys) and no more than three consecutive trials had the same N1 number, N1 alien type, or N2 alien type.

Match	Condition	Example
match	SS	The pinky above the greeny is mimming.
match	PP	The pinkies above the greenies are mimming.
mismatch	SP	The pinky above the greenies is mimming.
mismatch	PS	The pinkies above the greeny are mimming.

Table 1: Example target sentences for each condition.

Scenes were created for each list. In each list, the side of the screen where the mimming action occurred (left vs. right) was balanced and pseudorandomized such that the action did not occur on the same side on more than three consecutive trials. For each scene, the group in which the action occurred was pseudorandomly paired with another group of aliens such that the same overall scene configuration did not occur in two consecutive trials and each possible alien group arrangement (e.g. two blueys above one greeny) appeared an equal number of times on either side of the screen.

A subset of eight target sentences were also elicited during the experiment's practice sessions. In each practice session, each preposition type was used once, each alien type appeared as both N1 and N2, and the mimming action occurred an even number of times on either side of the screen. Each experiment list appeared with the same set of practice scenes. The practice scenes were created using the same criteria as the experimental scenes, with the added constraint that the practice scene configurations did not appear in experimental trials in any of the four presentation lists.

2.1.3 Procedure

Participants were distributed across the four presentation lists: 11 participants saw list 1, 10 participants saw list 2, 13 participants saw list 3, and 11 participants saw list 4. Experiment 1 was presented to the participants in our lab on a 2013 15" or 13" Macbook Pro using PsychoPy v1.85.3 (Peirce et al., 2019). The experiment session was recorded with a Rode NT1 microphone with a Blue Icicle USB interface in Audacity v2.2.1 at 44100 Hz. Reference sounds at the onset and offset of each trial allowed the trials to be identified within the continuous session recording. An experimenter was present in the room for the entire duration of the experiment.

The experiment started with an introductory sequence introducing the task and training the response formula. The sequence was led by the experimenter, though all instructions were also displayed onscreen. The introduction first showed the three types of aliens and demonstrated the mimming action. To familiarize participants with this new vocabulary, participants practiced a simple version of the task with only one group of aliens, eliciting the response *the blueys are mimming*. After this example, participants saw a similar scene with two groups of aliens, and the

experimenter explained that in this context the response *the blueys are mimming* does not provide enough information to identify exactly which aliens are performing the action. Participants were told that they could use words like *above*, *below*, *to the right of*, and *to the left of* to make their descriptions more precise. Participants then saw a display of aliens in the four possible spatial relations that elicited these prepositions and practiced describing the mimming actions using the prepositions (e.g. *the pinky to the right of the bluey is mimming*). Participants then completed eight practice trials, divided into two practice sessions (four trials each).

The first practice session was untimed. Participants pushed a button on the keyboard to initiate each trial. After participants gave their responses, they pressed again to reveal the target sentence below the group where the action occurred. Participants pressed the button a third time to end the trial. The second practice session followed the same format as the experimental trials. During the second practice, the scenes automatically ended after 3s of mimming. This time pressure was intended to encourage participants to speak quickly and decrease potential revision time during sentence planning, thereby increasing the likelihood of production errors. Participants were not shown the target sentences for the trials in the second practice. The experimenter encouraged the correct response format in the practice sessions, though verbal feedback given during the practice never referenced number agreement or whether participants produced correct or incorrect verb agreement in their responses. After completing the practice sessions, participants proceeded to the experimental trials.

2.2 Experiment 2 (web-based experiment)

2.2.1 Participants

The participants for Experiment 2 were 39 native American English speakers (27F, 12M) from the Amazon Mechanical Turk participant pool, with an average age of 41.7 years (SD = 9.3y; age data missing for one participant). We ran an additional six omitted participants. One was omitted for not following task directions by not using the correct sentence structure (instead producing sentences in the form *the pinky is mimming above the greenies*), two participants were omitted because all three researchers judged their reported language background information to be misleading, and the other three were omitted because over 1/3 of their trials were excluded (see *Analysis* for trial exclusion criteria). In order to qualify to participate in our task on Amazon Mechanical Turk, individuals had to pass our native speaker test. Participants typically completed this experiment in about 20 minutes and were given monetary compensation ($4.00).

2.2.2 Materials

Experiment 2 elicited the same 96 target sentences from Experiment 1. The experiment used the introductory sequence, practice trials, and experiment scenes from presentation list 1 in Experiment 1. The only difference in the scenes from Experiment 1 was the time frame participants

had to respond. In a short pilot of the task (N = 5), we noticed that the pilot participants had difficulty completing their responses in the 3s time frame given in the Experiment 1. In Experiment 2, each trial was recorded individually as opposed to recording the entire experiment session in a single recording (individual trial recordings were more efficient for analysis, less intrusive to participant privacy outside of a lab setting, and resulted in a smaller data file to upload to our data collection server at the end of the task). Since recordings ended automatically at the offset of each trial, we added an additional 1s of mimming to the scenes (for a 4s response window) so that participants' full responses would be more likely to be captured by the trial recordings.

2.2.3 Procedure

Experiment 2 was presented using the PennController for IBEX (PCIbex) (Zehr & Schwartz, 2018). Participants were asked to complete the experiment on a computer using the Google Chrome web browser (a commonly-used browser that permits collection of audio data via PCIbex). Participant responses were recorded through the microphone connected to their computer. Recordings started and stopped automatically at the beginning and end of each trial (removing the need for the reference sounds played in Experiment 1).

To minimize the influence of internet connection variability on stimulus presentation, all experiment materials were preloaded into the browser cache prior to starting the experiment. Recordings were stored in the browser cache throughout the task and sent to our data collection server in a .zip file at the end of the experiment. Using these features, we were able to avoid pauses mid-experiment to load stimuli or send recordings, instead moving these delays to the beginning and end of the experiment. To ensure that participants waited until their responses had been sent to the server before closing out of the experiment, participants were required to submit a completion code on Amazon Mechanical Turk that was given only after their responses successfully uploaded to the server (or failed to upload). If for any reason a participant's responses failed to send to the server, they were given the opportunity to download a .zip archive of their recordings to submit via email.

Experiment 2 followed the same general procedure as Experiment 1. Given Experiment 2's web-based setting, we were concerned about the attention participants would give to the instructions and about the clarity of the recorded audio. To address these concerns, we made two modifications to the procedure. To ensure that participants would attend to the introductory sequence without an experimenter present to explain it, we added audio instructions to the sequence. The audio instructions gave the same information presented onscreen in a slightly different format, just as the experimenter had presented the information in Experiment 1. By having the audio not simply repeat verbatim the information on the screen, we hoped to encourage participants' attention. Presenting the task instructions in multiple formats decreased

the likelihood that participants would miss important information due to inattention. Crucially, participants could not advance between screens in the introduction until the audio for the current screen had finished playing, meaning that participants could not skip over any of the information.

To address the concern of recording clarity, the instructions asked participants to move to a room that was quiet and away from distractions. Participants initially received this instruction at the beginning of the experiment and were reminded of it before starting the experimental trials. Prior to the introductory sequence, we required participants to record themselves saying a sample sentence and play the audio back to themselves. The experiment would not proceed until participants had recorded a response. Participants could re-record their sample sentence as many times as they wanted. This recording check allowed participants not only to test whether their microphone was working (so that they did not submit recordings containing no sound) but also whether their recordings were clear and free from background noise (and to make adjustments as necessary). The recording test additionally ensured that participants had their computer sound on prior to the introductory sequence so that they could hear the audio instructions.

3 Analysis

The goals of our planned analyses were to assess i) whether verb agreement errors were more likely in the mismatch conditions than the match conditions and ii) whether participants were more likely to pause in these same conditions before producing the correct verb form in utterances without errors. The data for Experiment 1 and Experiment 2 were analyzed separately (post-hoc analyses combining the data from both experiments are available in the Supplementary Materials). All statistical analyses were performed in R v 4.1.0 (R Core Team, 2021).

We fit Bayesian generalized linear mixed effects models to analyze the speech error and timing data. The models were fit with a Markov Chain Monte Carlo (MCMC) approach using the package {rstanarm} v2.21.1 (Goodrich et al., 2020). All models in our study were computed in four sampling chains, with 10,000 iterations each (5,000 of which were used for warm-up/burn-in in each chain, leaving a total of 20,000 sampling iterations in the analysis). We utilized Bayesian estimation for our analyses because it can handle complex model structures that can be difficult to estimate using frequentist methods (such as maximum likelihood estimation), particularly when datasets contain few or no observations of an outcome in one or more cells of the analysis, as is likely to occur in agreement attraction experiments when errors are less common in the match conditions.

Bayesian models estimate posterior distributions of probable values for the model parameters. We report the posterior medians and 95% credible intervals (CrIs) for the relevant parameter coefficients in our analyses. Highest density intervals were used as CrIs. CrIs indicate the range of values in which the regression coefficient for the parameter is 95% likely to lie, providing a more

intuitive measure of probability than frequentist confidence intervals. For hypothesis testing, we check whether zero is included in the CrI for a parameter; if zero is not in the CrI, we can be 95% certain that the parameter had a non-zero effect.

For each experiment, participant responses were transcribed and coded for agreement errors. A response was coded as containing an agreement error when the verb form produced by the participant did not match the number of the subject head (we included both revised and unrevised agreement errors in our analyses). During the transcription process, we also made notes of other errors and disfluencies in the responses. Responses were omitted from the analysis if the verb form was unidentifiable, if the response contained an error that changed the meaning of the sentence such that it did not match the scene it described (e.g. producing incorrect number marking on one or both of the nouns, saying the wrong alien name, and/or using an incorrect preposition), or if the response did not match the target formula (this category included incomplete responses). In Experiment 2, trial recordings that did not contain the complete target sentence were included in the analysis if the agreeing verb (*is* or *are*) was articulated before recording offset.[1] Using a preposition that did not match the target but expressed the same meaning (e.g. *under* in place of *below*), saying *a* in place of *the*, or producing an alternative pronunciation of mimming (e.g. *meeming*) were not considered errors. If the participant revised an error that would result in trial exclusion in a single revision (e.g. *the bluey… blueys above the greenies are mimming*), the response was not omitted and instead was coded as containing a disfluency error. Other disfluency errors included omitting a determiner, repeating a word or the beginning of a word, false starts to a word (e.g. *the gr- blueys*), word revisions (excluding revisions of agreement errors), and saying the color of an alien instead of its name (e.g. *the greens*).

We analyzed the likelihood of producing an agreement error in the different experiment conditions using logistic mixed effects analyses. We fit Bayesian generalized linear mixed effects models with a binomial distribution and logit link. The models had fixed effects of match and N1 number (with an interaction), random intercepts for target sentence and participant, and a random slope for match by participant. Since our models contained interactions, we used effects contrast coding to estimate the overall effects of match and N1 number, which allowed us to compare each fixed effect to the grand mean (analogous to main effects). The models were fit with a Student-*t* prior centered at 0 with 7 degrees of freedom and a scale of 2.5, a weakly-informative prior family commonly used in logistic regression analyses (e.g. Gelman et al., 2008); we used this same prior for both the regression coefficients and the intercept. For all other priors, we used the defaults from the {rstanarm} package v2.21.1. For each experiment, we analyzed both

[1] We included these trials in the Experiment 2 analysis to more closely parallel our analysis for Experiment 1. In the Experiment 1 analysis, we permitted responses that extended beyond the end of the trial because they were captured in the continuous recording of the experiment session. Participants' experience completing the task was the same in both experiments; in both tasks, participants were instructed to finish their response before the offset of the trial.

the full set of responses as well as a more restrictive dataset containing no disfluency errors. For both Experiment 1 and Experiment 2, we observed the same pattern of results in both datasets. We report the results from the full sets of responses because they contained a larger number of observations.

In addition to the error distribution analyses, we also conducted time-course analyses on responses that did not contain any errors (agreement or disfluency) to see if there were any slowdowns in speech caused by processing the agreement. We forced-aligned the audio recordings to their transcriptions using the Montreal Forced Aligner v1.0.0 (McAuliffe et al., 2017) to identify the onset and offset of each word in the responses. We used the forced-aligned responses to analyze the likelihood of pausing before articulation of the verb in the different sentence conditions. While it is possible that attraction effects could also manifest as lengthening of the subject phrase preceding the verb, differences between our conditions of interest could not be confidently interpreted as attraction effects because the match manipulation is confounded with word length, particularly in the SS vs. SP comparison (the plural nouns in our experiments take longer to articulate). Given the rarity of detectable pauses between words in forced-aligned continuous speech, we decided to perform a likelihood analysis.[2] Investigating the durational difference between the offset of N2 and the onset of the verb would result in a large number of zero values in the analysis, which can present difficulties for statistical modeling without transforming the data. Consequently, we analyzed the likelihood of pre-verbal gaps in the responses, defined as non-zero differences between the offset of N2 and the onset of the verb.

For the pre-verbal gap distribution analyses, we fit Bayesian logistic mixed effects models with the same effects structures and priors used in the error distribution analyses. As an exploratory analysis, we also investigated the duration of the pre-verbal gaps when they occurred. For these exploratory analyses, we fit Bayesian generalized linear mixed effects models with a gamma distribution and log link. The exploratory models were fit using the recommended default weakly-informative priors from the {rstanarm} package v2.21.1 for the gamma distribution family (without auto-scaling): for both the coefficients and intercept, the prior distribution was a normal distribution centered at 0 with a scale of 2.5. For all other priors, we used the defaults from the {rstanarm} package v2.21.1. The exploratory models had the same effects structure as the error and gap likelihood analyses.

In our results, we report the parameter estimates on the scales used in the analyses (log odds for the likelihood analyses and log milliseconds for the duration analyses), which are commonly used

[2] In our experience, the forced-aligner we used (McAuliffe et al., 2017) detects mid-word pauses of 30 ms or greater. While this threshold may be unable to detect more subtle timing effects, 30 ms is well below average reaction time estimates (e.g. Bańkosz et al., 2013; Wilkinson & Allison, 1989), and the mismatch timing effects for correct responses observed in prior attraction studies (e.g. Brehm & Bock, 2013; Haskell & MacDonald, 2003; Staub, 2009, 2010; Veenstra, Acheson, et al., 2014; Veenstra, Acheson, & Meyer, 2014) tend to be larger than 30 ms.

for hypothesis testing. To assist with interpretation, for each condition, we provide probability estimates (for likelihood analyses) or duration estimates from our models. Posterior probability distribution plots, effect plots, and fixed effect Probability of Direction (PD), Effective Sample Size (ESS), and R-hat values for our analyses are available in the Supplementary Materials.

4 Results

For Experiment 1 (the in-lab experiment), we collected a total of 4320 responses, 304 of which were omitted from the analysis (7%). Of the remaining 4016 responses, 249 contained disfluency errors (6%). Experiment 2 (the web-based experiment) elicited a total of 3744 utterances, 252 of which were omitted from the analysis (7%). Of the remaining 3492 responses, 275 contained disfluency errors (8%).

In both experiments, there were higher counts of errors resulting in trial omission in the mismatch conditions (tables showing the distribution of non-agreement errors in the responses are available in the Supplementary Materials). In Experiment 1, there were 129 such errors in the match conditions and 190 in the mismatch conditions; in Experiment 2, there were 129 such errors in the match conditions and 145 in the mismatch conditions. The error type with the greatest difference between the match and mismatch conditions was incorrect number marking on one or both nouns in the sentence. This was the most common error type in Experiment 1 and the second most common in Experiment 2; the most common error type in Experiment 2 was incomplete errors.[3] Across both experiments, number marking errors were more common for plural nouns than singular ones: in Experiment 1, 136 out of 176 total number marking errors (77%) were on plural nouns, and in Experiment 2, 46 out of 67 number marking errors (69%) were on plural nouns.

4.1 Agreement error distribution analyses

4.1.1 Experiment 1

Of the 4016 responses in the Experiment 1 analysis, 489 responses contained agreement errors (48 errors were in sentences that also contained at least one disfluency error). The distribution of agreement errors across conditions is given in **Table 2**. We elicited greater counts and higher percentages of errors in the mismatch conditions than the match conditions. **Figure 2** shows the participant error rates in each condition.

[3] The increased number of incomplete errors in Experiment 2 relative to Experiment 1 may be in part due to a coding difference between the two experiments. As noted above, in the Experiment 1 analysis, we permitted responses that extended beyond the 3s response window because they were captured in the continuous recording; thus, incomplete errors in Experiment 1 comprise the cases when participants stopped talking before completing the full target sentence structure. In Experiment 2, on the other hand, incomplete errors include cases when the agreeing verb was not articulated before recording offset.

Match	Condition	Error Count	Response Count	Error Percentage
match	SS	4	1046	0.4%
match	PP	30	994	3.0%
mismatch	SP	246	990	24.8%
mismatch	PS	209	986	21.2%

Table 2: Agreement error counts and rates in Experiment 1 (in-lab experiment).

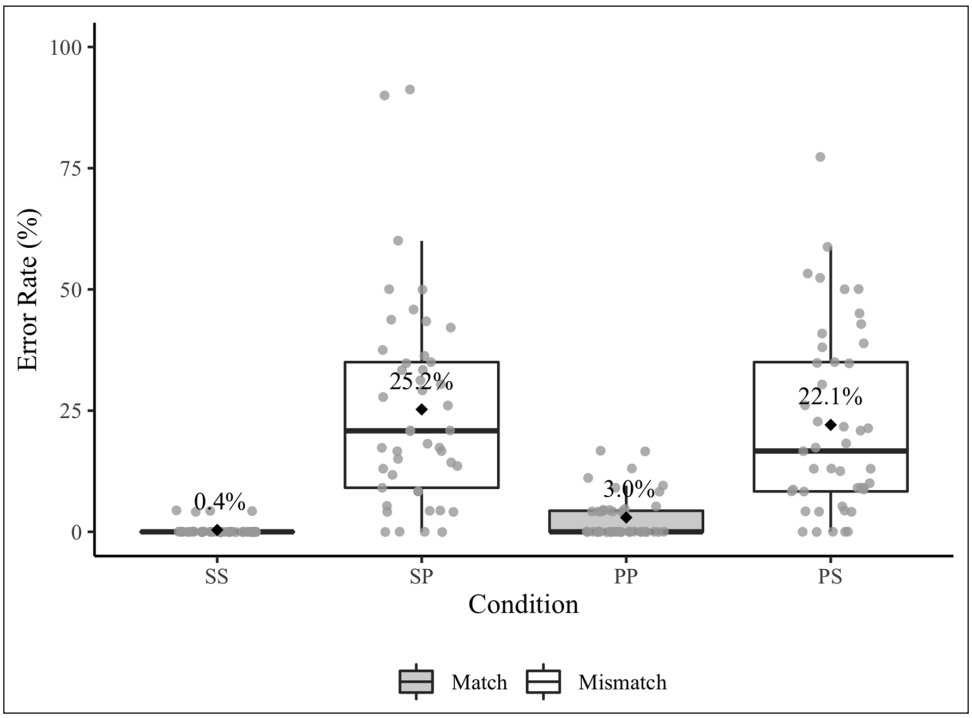

Figure 2: Boxplot of Experiment 1 participant agreement error rates. Mean error rates for each condition are labelled and identified by the black diamonds. The grey points represent participant rates.

The posterior median for the overall effect of match was 1.78 (95% CrI [1.44, 2.15]), suggesting that errors were more likely in the mismatch conditions (SP, PS). The posterior median for the overall effect of N1 number was 0.47 (95% CrI [0.21, 0.77]), suggesting that errors were more likely in the plural head conditions (PP, PS). The posterior median for the interaction was –0.60 (95% CrI [–0.89, –0.33]), implying a difference in the match effect between the singular and plural head conditions (see Jaccard, 2001 for discussion of how to interpret interaction coefficients in logistic regression). The estimated probabilities of agreement errors were 0.2% (95% CrI [0.1, 0.6]) in the SS condition, 20.5% (95% CrI [14.5, 27.8]) in the SP condition, 1.9% (95% CrI [0.9, 3.2]) in the PP condition, and 16.8% (95% CrI [11.7, 23.2]) in the PS condition. Follow-up analyses fitting Bayesian models of the same structure with dummy-coded variables

estimated the match effect to have a posterior median of 4.32 in the singular head conditions (95% CrI [3.42, 5.33]) and 2.25 in the plural head conditions (95% CrI [1.67, 2.82]), suggesting that errors were more likely in the mismatch conditions for both the SS–SP and PP–PS comparisons.

4.1.2 Experiment 2

Of the 3492 responses in the Experiment 2 analysis, 320 responses contained agreement errors (36 errors were in sentences that also contained at least one disfluency error). The distribution of the errors across conditions for Experiment 2 is given in **Table 3**. As in Experiment 1, we elicited more errors in the mismatch conditions than the match conditions. **Figure 3** shows the participant error rates in each condition.

Match	Condition	Error Count	Response Count	Error Percentage
match	SS	0	892	0%
match	PP	21	861	2.4%
mismatch	SP	130	859	15.1%
mismatch	PS	169	880	19.2%

Table 3: Agreement error counts and rates in Experiment 2 (web-based experiment).

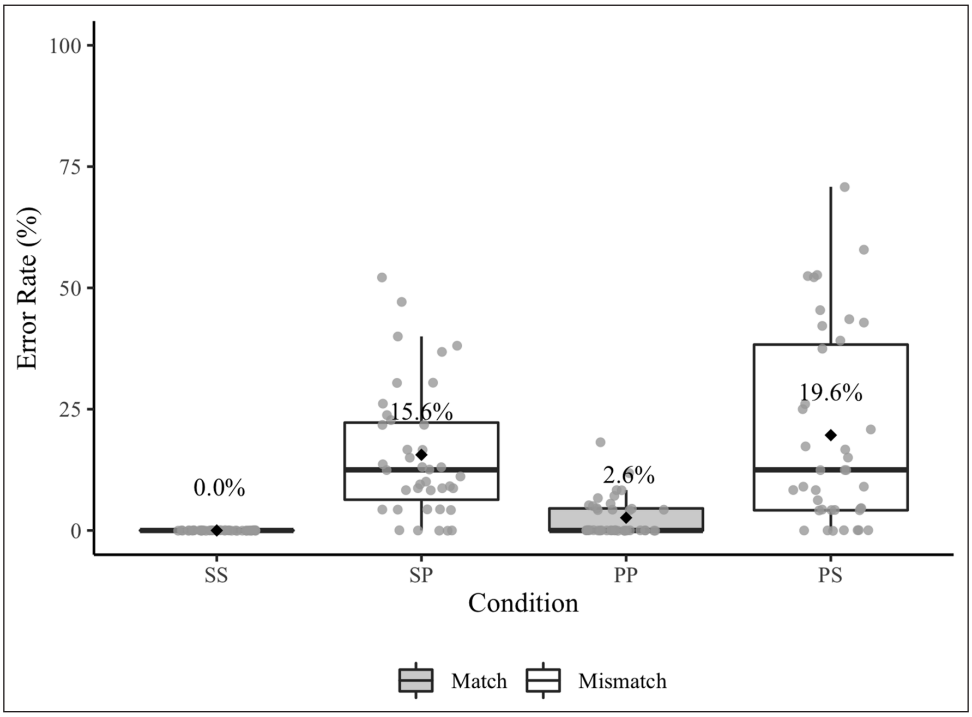

Figure 3: Boxplot of Experiment 2 participant agreement error rates. Mean error rates for each condition are labelled and identified by the black diamonds. The grey points represent participant rates.

The posterior median for the overall effect of match was 2.50 (95% CrI [1.60, 3.86]), suggesting that errors were more likely in the mismatch conditions (SP, PS). The posterior median for the overall effect of N1 number was 1.39 (95% CrI [0.55, 2.72]), implying that errors were more likely in the plural head conditions (PP, PS). The posterior median for the interaction was –1.23 (95% CrI [–2.52, –0.35]), suggesting that there was a difference in the match effect between the singular and plural head conditions. The estimated probabilities of agreement errors were 0.0% (95% CrI [0.0, 0.1]) in the SS condition, 9.8% (95% CrI [5.6, 16.0]) in the SP condition, 1.2% (95% CrI [0.5, 2.5]) in the PP condition, and 12.9% (95% CrI [7.5, 20.3]) in the PS condition. Follow-up analyses with dummy-coding estimated the match effect to have a posterior median of 4.92 in the singular head conditions (95% CrI [3.34, 6.82]) and 2.39 in the plural head conditions (95% CrI [1.57, 3.23]), suggesting that errors were more likely in the mismatch conditions in both the SS–SP and PP–PS comparisons.

4.2 Time-course analyses

4.2.1 Experiment 1

The time-course analysis for Experiment 1 was performed on 3267 responses containing no agreement or disfluency errors (one participant was omitted from this analysis because their responses could not be forced-aligned). Among these responses, 326 contained gaps (non-zero pauses) between the end of N2 and the beginning of the verb. The distribution of gaps across conditions is given in **Table 4**.

Match	Condition	Gap Count	Response Count	Gap Percentage
match	SS	45	972	4.6%
match	PP	55	898	6.1%
mismatch	SP	100	686	14.6%
mismatch	PS	126	711	17.7%

Table 4: Gap counts and rates in Experiment 1 (in-lab experiment).

In the gap likelihood analysis, the median posterior estimate of the overall match effect was 0.85 (95% CrI [0.60, 1.10]), suggesting that pre-verbal gaps were more likely in the mismatch conditions (SP, PS). The median posterior of the overall effect of N1 number was 0.19 (95% CrI [0.04, 0.33]), suggesting that gaps were more likely before *are* than *is*. The median posterior for the interaction between match and N1 number was –0.02 (95% CrI [–0.16, 0.12]), implying that there was not a reliable difference in the match effect between the singular and plural head conditions. The estimated probability of pre-verbal gaps was 1.9% (95% CrI [0.9, 3.5]) in the SS

condition, 9.9% (95% CrI [6.2, 15.0]) in the SP condition, 2.9% (95% CrI [1.5, 5.1]) in the PP condition, and 13.4% (95% CrI [8.7, 19.5]) in the PS condition.

In the exploratory gap duration analysis, the posterior median for the overall effect of match was 0.08 (95% CrI [–0.10, 0.24]). The posterior median for the overall effect of N1 number was 0.28 (95% CrI [0.14, 0.41]), suggesting that gaps were longer before *are* than *is*. The posterior median for the interaction between match and N1 number was –0.06 (95% CrI [–0.19, 0.08]). The estimated gap durations were 56 ms (95% CrI [36, 85]) in the SS condition, 74 ms (95% CrI [56, 96]) in the SP condition, 109 ms (95% CrI [73, 165]) in the PP condition, and 114 ms (95% CrI [88, 146]) in the PS condition.

4.2.2 Experiment 2

The time-course analysis for Experiment 2 included 2929 responses containing no agreement or disfluency errors (one participant's responses and an additional three trials were omitted from the analysis because they could not be forced aligned or contained an *uh* mid-utterance). Among these responses, 656 contained pre-verbal gaps. The distribution of gaps is presented in **Table 5**.

Match	Condition	Gap Count	Response Count	Gap Percentage
match	SS	106	819	12.9%
match	PP	103	772	13.3%
mismatch	SP	214	680	31.5%
mismatch	PS	232	658	35.3%

Table 5: Gap counts and rates in the Experiment 2 (web-based experiment).

In the gap likelihood analysis, the median posterior estimate for the overall effect of match was 0.83 (95% CrI [0.64, 1.01]), suggesting that pre-verbal gaps were more likely in the mismatch conditions (SP, PS). The posterior median for the overall effect of N1 number was 0.08 (95% CrI [–0.05, 0.21]). The posterior median for the interaction between match and N1 number was 0.04 (95% CrI [–0.08, 0.18]). The estimated probability of pre-verbal gaps was 8.0% (95% CrI [4.5, 13.0]) in the SS condition, 29.3% (95% CrI [19.5, 41.2]) in the SP condition, 8.5% (95% CrI [4.8, 13.9]) in the PP condition, and 34.7% (95% CrI [23.8, 47.3]) in the PS condition.

In the exploratory gap duration analysis, the median posterior estimate for the overall effect of match was 0.21 (95% CrI [0.11, 0.30]), suggesting that gaps were longer in the mismatch conditions. The posterior median for the overall effect of N1 number was 0.08 (95% CrI [0.00, 0.16]). The posterior median for the interaction between match and N1 number was 0.06 (95% CrI [–0.03, 0.14]). The estimated gap durations were 58 ms (95% CrI [46, 73]) in the SS condition,

78 ms (95% CrI [64, 96]) in the SP condition, 61 ms (95% CrI [49, 76]) in the PP condition, and 103 ms (95% CrI [84, 127]) in the PS condition.

4.2.3 Pre-verbal slowdowns before agreement errors (post-hoc analysis)

To investigate whether the pressures leading to timing effects in correct utterances are also present when participants produce errors, we conducted an exploratory post-hoc analysis investigating pre-verbal slowdowns in responses with unrevised agreement errors and no disfluency errors. We restricted our investigation to fluent responses with no verb agreement revisions or non-agreement speech errors for maximum comparability to the correct responses included in the time-course analyses.[4] There were 384 such responses across both experiments (237 from Experiment 1, 147 from Experiment 2). Within the 383 responses included in the present investigation (one response from Experiment 2 was omitted because it could not be forced-aligned), 43 contained pre-verbal gaps. The distribution of responses and gaps is presented in **Table 6**; the table counts collapse across experiments, though the counts for Experiment 1 are given in parentheses. Given the small sample of responses in the match conditions, we did not analyze gap likelihood across conditions.

Match	Condition	Gap Count	Response Count	Gap Percentage
match	SS	0 (0)	1 (1)	0%
match	PP	6 (3)	23 (14)	26.1%
mismatch	SP	15 (7)	180 (126)	8.3%
mismatch	PS	22 (7)	179 (96)	12.3%

Table 6: Gap counts and rates in responses with unrevised agreement errors. Counts from Experiment 1 are given in parentheses.

To investigate the relationship between timing and accuracy in attraction-inducing environments, we compared the likelihood of gaps before correct verbs and agreement errors in the mismatch conditions. Our analysis included 2735 correct responses and 359 error responses.

[4] The responses elicited in the experiments had a relatively high rate of agreement error revision: 217 out of 489 agreement errors in Experiment 1 were revised (44%), and 125 out of the 279 errors in fully-articulated Experiment 2 responses were revised (45%) (the rate of revision in the full set of Experiment 2 responses (43%) was similar but likely underestimates the true revision rate, as it includes recordings that cut off responses early, for which it is unknown whether a revision occurred after recording end). The majority of revisions occurred directly after articulation of the initial agreeing verb form (prior to articulating the verb *mimming*), thus we restricted Experiment 2 responses in the present investigation to those that include at least the onset of the verb *mimming* within the recording, thereby reducing the likelihood that the responses included a revision not captured in the recording. 10 responses from Experiment 2 were consequently excluded from the investigation (1 match, 9 mismatch).

We fit a Bayesian generalized linear mixed effect model with a binomial distribution and a logit link using the same priors as the planned error and gap likelihood analyses. The model had fixed effects of response type (correct vs. error) and experiment with an interaction as well as random intercepts for participant and target sentence. We used effects contrast coding for the fixed effects.

The median posterior estimate for the effect of response type was 0.64 (95% CrI [0.43, 0.85]), suggesting that pre-verbal gaps were more likely for correct responses. The posterior median for experiment was –0.45 (95% CrI [–0.79, –0.07]), indicating that gaps were more likely in Experiment 2. The posterior median for the interaction between experiment and response type was –0.23 (95% CrI [–0.44, –0.01]), suggesting that the difference between correct and error sentences was different in Experiment 1 and Experiment 2. The estimated gap probabilities in Experiment 1 were 10.8% (95% CrI [7.0, 15.9]) for correct responses and 5.1% (95% CrI [2.4, 9.7]) for error responses; the estimated gap probabilities in Experiment 2 were 32.0% (95% CrI [22.7, 42.5]) for correct responses and 7.7% (95% CrI [3.9, 14.1]) for error responses. Follow-up analyses fitting Bayesian models of the same structure with dummy coded variables estimated the response type effect to have a posterior median of 0.80 (95% CrI [0.21, 1.43]) in Experiment 1 and 1.70 (95% CrI [1.16, 2.28]) in Experiment 2, suggesting that pre-verbal gaps were more likely for correct responses in both experiments. The follow-up analyses suggested that the gap likelihood in correct responses was greater in Experiment 2 than Experiment 1 (posterior median 1.33, 95% CrI [0.69, 1.98]) but that gap likelihood in error responses was similar in both experiments (posterior median 0.44, 95% CrI [–0.52, 1.44]).

5 General Discussion

In this study, we investigated subject–verb agreement attraction effects in continuous speech elicited in both a supervised in-lab experiment (Experiment 1) and an unsupervised web-based experiment (Experiment 2). For both experiments, we analyzed the likelihood of producing verb agreement errors in the presence of a mismatching attractor and looked for production time-course effects in the articulation of correct responses, allowing us to assess the influence of attractor interference during verb agreement even when no overt error was produced. We observed remarkably similar results between our two experiments in both measures (speech errors and articulation time-course), demonstrating that it is feasible to gather reliable data for production experiments via the internet, despite potential concerns about eliciting and collecting speech recordings online (see also Fairs & Strijkers, 2021; Vogt et al., 2021; Ziegler & Snedeker, 2018).

Three key findings shed further light on the computations underlying subject–verb agreement: 1) our more naturalistic scene-description paradigm elicited reliable interference from singular attractors (both in-lab and online), contradicting the traditionally-assumed stark markedness contrast, 2) attractors influence verb processing even when the correct verb is produced (as

reflected in articulation time-course), suggesting that error rates underestimate the proportion of trials on which attraction pressure occurs, and 3) there is evidence for a trade-off in error and timing manifestations of attraction effects. The relationship between verb accuracy and timing in our results suggests that the pressures leading to attraction are active during or shortly before verb planning, as would be expected if attraction effects arise within the verb agreement computation itself (i.e. as the generator references the subject features to determine the verb features). We first discuss the methodological comparison before turning to the agreement attraction results and consequences for models of subject–verb agreement formation.

5.1 Feasibility of web-based production experiments

Our study demonstrates that it is both feasible and convenient to collect language production data via the internet, even in a task that involves more open-ended responses such as a scene-description task (see also Ziegler & Snedeker, 2018). Despite potential concerns that the collection and quality of production data might be more susceptible to limitations of web-based research than other forms of linguistic data (Reips, 2002; Skitka & Sargis, 2006), we were able to collect recordings of consistent responses with high enough audio quality to be both transcribed and forced-aligned. In fact, issues that can plague data collection in more traditional web-based tasks, such as concerns about non-qualified participants and bots (including humans using automation), appear to be largely moot in production experiments like ours. Native-sounding spoken descriptions are not easily faked nor are they a form of response that contemporary automation can easily simulate. This not only facilitates screening measures but also appears to have discouraged participation from non-qualified individuals and bots who may have passed our native speaker test. We had no obvious bots complete the task on Amazon Mechanical Turk, and only two participants who completed the task had non-native accented speech (these two participants were not included in the analysis). Furthermore, we observed the same agreement attraction effects in our web-based experiment as in our in-lab experiment, suggesting that the web-collected data is sensitive to the types of effects of interest in production studies.

We attribute our success in part to specific measures we took to mitigate potential drawbacks of web-based data collection, particularly variations in home set-ups (e.g. hardware, software, background environment) and internet connections as well as the lack of experimenter supervision (these measures are described in detail in the Experiment 2 *Procedure*). These steps helped to reduce trial and participant loss as well as to avoid potential effects of varying internet connection stability that could result in stimulus presentation delays. While there was a somewhat higher participant exclusion rate in the web-based experiment, with approximately 8% of participants run in the task omitted in the in-lab experiment and 13% omitted in the web-based experiment, slight increases in participant loss rate in web-based settings are easily offset by the more efficient

use of experimenter time: it took three months to complete data collection for Experiment 1, whereas data collection for Experiment 2 was completed within only nine days.

Online data collection also allowed us to recruit a more diverse participant sample. While Experiment 1 included only participants close to the University of Maryland (primarily undergraduates), Experiment 2 included participants from 23 U.S. states, with ages ranging from 25–61 years. Despite this variability, the response patterns of participants in the web-based experiment closely resembled those in the in-lab experiment, as reflected in similar trial error rates and parallel results. Our comparison demonstrates that it is possible to run a relatively complex production experiment online with similar response quality to an in-lab experiment, even without an experimenter present to answer questions and provide feedback: in both Experiment 1 and Experiment 2, only one participant consistently did not follow task instructions to produce the correct form of response, and only 1% of responses elicited from other participants[5] failed to follow the target format.

Nonetheless, there were some notable differences between the in-lab and web-based data. Although the recordings collected in Experiment 2 had sufficient audio quality to perform detailed analyses of participant responses, there was an apparent influence of hardware variations on the forced-aligner's ability to align the recordings to their transcribed content. For some individuals, the recordings captured the release of the button press used to start the trial, perhaps because the participant was using a mechanical keyboard, a more sensitive microphone, or a microphone closer to their keyboard set-up (such as a laptop's built-in microphone). In these trials when the button press was particularly loud, the forced aligner identified the keyboard sound as the first word in the response. This misalignment did not influence our time-course analyses, as we were interested in pre-verbal pauses later in the sentence, but it could influence other types of analyses common in language production research. We have found in subsequent experiments that this type of noise in the recordings can be avoided by adding a short buffer time (e.g. 200 ms) after trial onset before the recording starts.

We also observed slightly different response behavior between the two experiments. We found from piloting that an additional second was needed in Experiment 2 to allow participants to complete their responses within the trial recordings. Participants appeared to use this additional time to plan and articulate their responses, taking longer to start speaking and articulating more slowly (measuring from NP1 onset to avoid inaccuracies in the forced-aligner's identification of the first word). For responses included in the timing analysis, the average NP1 onset in Experiment 2 was 1.96s (SD = 0.23s), compared to 1.77s (SD = 0.21s) in Experiment 1; onsets were slower in Experiment 2 in all four experiment conditions. The average response

[5] This estimate includes participants omitted from analysis for having too many excluded trials but excludes those who failed the native speaker test or had misleading language background information.

durations (measured from NP1 onset to sentence offset) were 2.27s (SD = 0.38s) in Experiment 2 (excluding trials that were not fully articulated within the recording) and 1.96s (SD = 0.38s) in Experiment 1; durations were longer in Experiment 2 in all four conditions. The slower responses in Experiment 2 may be due to the older participant sample or because participants prefer to use additional time to plan and give responses when that time is available. Even with these response timing differences, however, we found that the pattern of the results was remarkably similar between our two experiments.

5.2 Agreement error results

In both experiments, the error analysis replicated the basic attraction effect for subject–verb agreement: verb agreement errors were more plentiful and more likely in sentences containing an attractor that mismatched the subject head in number. The in-lab experiment elicited errors in approximately 23% of mismatch trials and 2% of match trials, and the web-based experiment elicited errors in approximately 17% of mismatch trials and 1% of match trials. In a post-hoc analysis combining the data from both experiments (see Supplementary Materials), we did not observe an interaction between match condition and experiment, suggesting similar attraction effects in both experiments despite their different experimental settings. The overall effect of experiment was reliable in the post-hoc analysis, with errors more likely in Experiment 1 than Experiment 2. The reduced likelihood and number of agreement errors in the web-based experiment may be the result of the additional time participants in the online version had to plan and articulate their responses (we discuss this hypothesis further in the *Timing results*).

In addition to the attraction effect, we also observed a reliable effect of subject head number in our experiments such that errors were more likely in the plural head conditions (PP, PS). A trend towards more errors in trials with plural subject heads has been reported elsewhere in the literature (e.g. Bock & Cutting, 1992; Franck et al., 2002; Nozari & Omaki, 2022; Staub 2009, 2010; Thornton & MacDonald, 2003; Veenstra, Acheson, & Meyer, 2014; Vigliocco & Nicol, 1998; Vigliocco et al., 1995). We observed parallel effects of subject head number in the timing analyses: participants were more likely to pause and paused for longer before producing *are* than *is* in correct sentences (see Supplementary Materials for post-hoc gap likelihood and duration analyses). Prior studies have explained effects of plural heads as reflecting increased processing complexity for plural nouns (e.g. Eberhard, 1997; Franck et al., 2002). This hypothesis is consistent with the fact that we observed more number marking errors on plural nouns than singular ones in our experiments. We may also observe more erroneous productions of *is* than *are* due to the form's higher frequency: *is* occurs more frequently than *are* in corpora (e.g. SUBTLEX-us; Brysbaert & New, 2009), and singular nouns are more frequent than plural ones (Greenberg, 1966), meaning that singular verb agreement is also likely more frequent. In fact, there is evidence for an overall bias towards singular verb forms in both spontaneous and elicited speech (Duffield, 2013).

5.2.1 The reduced markedness effect

A striking finding in our study is that our experiments elicited relatively high error rates in both the SP and PS conditions, replicating the reduced markedness asymmetry observed in other description tasks (Nozari & Omaki, 2022; Veenstra, Acheson, & Meyer, 2014). This finding contrasts with traditional demonstrations of the markedness effect that elicited almost no interference from singular attractors in the PS condition (e.g. Bock & Miller, 1991; Bock et al., 2004; Bock et al., 1999; Eberhard, 1993, 1997; inter-alia). Although we observed differences in attraction for conditions with singular and plural heads (the interaction between match condition and head noun number was reliable in the experiment analyses as well as the post-hoc gap likelihood analysis), our experiments demonstrated reliable attraction effects from both plural and singular intervening nouns. Our post-hoc analysis did not reveal a three-way interaction between match condition, head noun number, and experiment, suggesting that the shape of the markedness effect was similar across both experiments.

While the large number of PS errors in our experiments may at first glance seem like an outlier, an investigation of the production literature reveals that the singular attraction effects we elicited are within the range observed in prior studies. In fact, interference from singular attractors appears to be fairly common in experiments investigating the agreeing verb *to be*. We searched for agreement attraction studies eliciting the verb *to be* and found nine experiments from seven studies that tested both singular and plural head conditions (**Table 7**). These experiments involve a range of paradigms (description, forced-choice, and preamble tasks) and languages (English, Dutch, and French). For the majority of the studies, the conditions tested were SS, SP, PP, and PS. The one exception is Franck et al. (2002), which involved two interfering noun phrases intervening between the subject head and the verb, thereby resulting in SSS, SSP, SPS, SPP, PPP, PPS, PSP, and PSS conditions. When investigating the markedness effect, Franck et al. (2002) primarily focused on the SSS vs. SPS and PPP vs. PSP comparisons.

With the exception of Franck et al. (2002, Experiment 1), all experiments in **Table 7** report a markedness effect with stronger interference from plural attractors. Nevertheless, in all but two studies (Thornton & MacDonald, 2003; Vigliocco & Nicol, 1998), the PS error rates are higher than the PP error rates (in the case of Franck et al., 2002, the PSP error rates are higher than the PPP rates). Both Franck et al. (2002) and Veenstra, Acheson, and Meyer (2014) explicitly tested for attraction effects in the plural head conditions: Veenstra, Acheson, and Meyer (2014) report a significant difference between the PS and PP conditions in Experiment 1 but not Experiment 2, and Franck et al. (2002, Experiment 1) report a significant difference between the PPP and PSP conditions. Our study is thus not alone in demonstrating interference from singular attractors.

The mean PS error rates in our experiments (although on the high end of the scale) are within the range of PS error rates elicited in other experiments using the same agreeing

Study	Experiment/ Condition	Language	Paradigm	Additional Manipulation	SS Error Rate	SP Error Rate	PP Error Rate	PS Error Rate	SP-SS (Pl attraction)	PS-PP (Sg attraction)	Pl vs. Sg Difference
Present	1	English	Description	NA	<1%	25%	3%	22%	25%	19%	6%
	2	English	Description	NA	0%	16%	3%	20%	16%	17%	–1%
Staub (2009)	2	English	Forced-choice	NA	3%	27%	9%	22%	24%	13%	11%
Staub (2010)	Intervening attractors	English	Forced-choice	NA	6%	33%	14%	21%	27%	7%	20%
	Non-intervening attractors	English	Forced-choice	NA	5%	25%	10%	16%	20%	6%	14%
Veenstra, Acheson, & Meyer (2014)	1	Dutch	Description	Preposition type (with vs. next to)	1% #	9–11% #	2–5% #	8–9% #	8–10%	3–7%	1–7%
	2	Dutch	Forced-choice	Preposition type (with vs. next to)	1–2% #	6% #	2–3% #	4–5% #	4–5%	1–3%	1–4%
Nozari & Omaki (2022)	NA	English	Description	Cue flash vs. Target flash	<1% #	3–5% #	1% #	3% #	3–5%	2%	1–3%
Franck et al. (2002)	1	French	Preamble	Two intervening NPs	1% (SSS) 0% (SSP)	10% (SPS) 5% (SPP)	1% (PPP) 3% (PPS)	9% (PSP) 6% (PSS)	9% (SPS-SSS) 5% (SPP-SSP)	8% (PSP-PPP) 3% (PSS-PPS)	1% 2%
Thornton & MacDonald (2003) †	1	English	Preamble	Plausibility of N2 as a subject	2%	18%	6%	5%	16%	–1%	17%
Vigliocco & Nicol (1998) †	1	English	Preamble	NA	0%	16%	6%	6%	16%	0%	16%
	2	English	Preamble	Elicited questions	3%	16%	5%	4%	13%	–1%	14%

Table 7: Mean participant error rates from a sample of agreement attraction studies eliciting the verb *to be*.

Note. Rates marked with a # were estimated from plots (exact values not reported in the text). For studies marked with a †, error rates were calculated as the percentage of agreement errors out of the sum of correct responses and those with errors (collapsing across other manipulations, when relevant). In experiments with additional manipulations, error rates are presented as a range when the mean error rate differed between the two manipulation conditions.

verb, and the degree of difference between plural and singular attraction (i.e. the size of the markedness asymmetry) generally falls within previously-observed ranges (though is smaller for Experiment 2). Our results are roughly consistent with the general trend in **Table 7** that accuracy across conditions follows the order SS > PP > PS > SP, a pattern that Staub (2009) additionally argues is consistent with previous literature (though the PS–SP contrast was not reliable in our experiments[6]). Participants in our study may have been especially prone to interference due to the presence of a time limit, the semantic similarity of the subject head and attractor referents,[7] the visual salience of the attractor, the use of a lexically-reduced item set with novel words and concepts, and/or the minimal syntactic and semantic variability between sentences.

The range of markedness asymmetries observed in the production literature suggests that the markedness effect can have substantial variation in strength. The strength of the markedness effect appears to vary with production paradigm, with generally higher rates of singular attraction in more naturalistic experiments (e.g. Experiment 1, Experiment 2; Nozari & Omaki, 2022; Veenstra, Acheson, & Meyer, 2014, Experiment 1) and forced-choice paradigms (e.g. Staub, 2009, 2010; Veenstra, Acheson, & Meyer, 2014, Experiment 2) compared to more traditional preamble completion paradigms (e.g. Bock et al., 1999; Eberhard 1993, 1997; Thornton & MacDonald, 2003; Vigliocco & Nicol, 1998; but cf. Franck et al., 2002). In fact, when we elicited the target sentences from our scene-description paradigm using a preamble elicitation task (Kandel & Phillips, 2022), the markedness asymmetry widened: we observed errors in 3% of SS trials and 16% of SP trials (a 13% point difference) compared to 4% of PP trials and 9% of PS trials (a 5% point difference). Consequently, it may be the case that the markedness asymmetry in natural production is more graded than in tasks like preamble paradigms that rely on comprehension and repetition of a sentence fragment to guide production planning.[8]

---

[6] The PS–SP comparison was not reliable in Experiment 1 (posterior median 0.24, 95% CrI [–0.07, 0.54]) or Experiment 2 (posterior median –0.33, 95% CrI [–0.93, 0.25]), though the SS–PP comparison was reliable in both Experiment 1 (posterior median 2.84, 95% CrI [1.37, 4.83]) and Experiment 2 (posterior median 1.85, 95% CrI [0.93, 2.81]). As previously mentioned, we also observed interactions between subject head number and match condition in the experiment analyses as well as the post-hoc error likelihood analysis. Similar PS and SP error rates could reflect differential attraction strength from singular and plural attractors coupled with different accuracy baselines for singular and plural heads in the match conditions (suggested by the SS–PP contrasts and the overall effects of head noun number in our statistical analyses).

[7] Shared features between subject heads and attractors (such as animacy) have been shown to increase attraction rates (Barker et al., 2001). Shared features such as gender, animacy, and semantic category may impair one's ability to retrieve information about similar items from memory (Gordon et al., 2001) and contribute to similarity-based interference in encoding (Villata et al., 2018).

[8] Corpus studies suggest that singular attraction can arise in natural production, though current studies do not provide a clear estimate of the size of the markedness effect in spontaneous speech. While there is evidence that spontaneous productions are consistent with a relatively sharp markedness effect (e.g. Haskell et al., 2010; Pfau, 2009), this pattern needs to be treated with a great deal of caution, as it could reflect the types of things individuals talk about rather than a contrast in underlying mechanisms due to number markedness, and there is potentially conflicting evidence that singular agreement errors are fairly common in natural speech (Duffield, 2013).

One hypothesis proposed by Ted Gibson (p.c.) to explain the variability in the markedness effect across task types is that a more binary asymmetry may arise in preamble paradigms due to misinterpretations of the preamble that are more likely to occur in the SP condition than the PS condition. This hypothesis assumes a noisy-channel framework (e.g. Gibson et al., 2013; Levy, 2008) in which comprehension can be swayed by interlocutors' expectations about the intended message of an utterance and potential noise in the signal. Within this framework, a participant may misinterpret a SP preamble as a PP preamble because they infer the subject head to have had intended but omitted plural number marking, thereby leading them to produce a plural verb in place of a singular one (e.g. Bergen & Gibson, 2012; Ryskin et al., 2021). The SP–SS contrast may arise because participants are less likely to misinterpret SS preambles due the high prevalence of singular-singular NP-Prepositional Phrase sequences in spoken English (Ryskin et al., 2021). A strong markedness effect may arise because insertions of the English plural suffix *-s* are less common than deletions (Stemberger, 1985), meaning that misinterpretations of PS preambles as intended SS preambles are unlikely, thereby leading to few agreement errors in the PS condition (Ryskin et al., 2021).

Another hypothesis to explain the stronger markedness effect in preamble experiments is that preamble tasks encourage different verb planning strategies than naturalistic speech. In a preamble task, participants may have more time between when they know the identity of the subject (at the beginning of the preamble) and when they are prompted to produce a verb (after hearing and repeating the full preamble). It is possible that participants leverage this additional time to help them produce fluent responses with few mid-articulation pauses. Under this hypothesis, participants in preamble tasks may at times adopt a generate-and-test strategy, generating a best-guess verb form upon hearing the subject and then later checking this candidate verb form against the full preamble. This checking process may operate similarly to feature-checking in comprehension, resulting in less susceptibility from singular attractors, similar to the patterns observed in comprehension studies of attraction (e.g. Wagers et al., 2009). This strategy may be more likely to be used in preamble paradigms with free verb choice (e.g. Bock et al., 1999; Eberhard 1993, 1997), in which participants may start to pick out verbs related to the subject head after encountering the subject at the beginning of the preamble. The strategy may also be common in preamble paradigms in which participants are provided with the sentence predicate before the preamble, such as in the two studies that showed strong markedness asymmetries in **Table 7** (Thornton & MacDonald, 2003 provided a verb to be used in a passive construction; Vigliocco & Nicol, 1998 provided a predicate adjective to describe the subject). This strategy may be less common in preamble tasks that use the same verb in each trial (e.g. Franck et al., 2002, Experiment 1) or when verb forms are presented for selection after preamble (e.g. Staub, 2009, 2010), eliminating the need to decide upon a predicate. This strategy may additionally be less common in forced-choice paradigms, as participants do not produce a response, thus removing the incentive to plan verb forms early in order to avoid pauses mid-articulation.

5.2.2 Accounting for markedness variability in models of agreement attraction

As mentioned in the *Introduction,* the markedness effect has commonly been accounted for in models of agreement attraction by assuming that plural nouns have a marked plural feature, thereby allowing them to interfere with agreement processes, whereas singular nouns are unmarked for number (e.g. Bock & Eberhard, 1993; Eberhard, 1997). This assumption allows a binary markedness effect to fall naturally out of an agreement system in which attraction effects result from ambiguity or inaccuracy in the subject number representation, as only plural attractors can influence with this representation. The presence of reliable singular attraction is more difficult to capture within such a framework, as it is unclear how a singular attractor could influence the subject number representation if it has no number feature to cause interference.

Dropping the assumption of a privative number feature system can make it easier to explain interference from singular attractors within a representational framework. For example, a binary feature system (e.g. [+/–Plural]) could allow the number representation of the subject phrase to be swayed by an attractor in either the plural ([+Plural]) or singular ([–Plural]) direction. Gradation in the strength of plural and singular attraction could arise if different weights are given to [+Plural] and [–Plural] features in a marking and morphing-style account or if negative features are less likely to percolate through the syntactic structure or to compete during encoding processes.

While representational accounts must move away from a privative number feature system to explain more graded markedness effects, we propose that similar influence from singular and plural attractors in production arises naturally within a retrieval framework of attraction, independent of whether the number feature system is privative or binary. This is due to the nature of the retrieval cues available to pick out the agreement controller in production. Unlike in comprehension, where the parser uses retrieval to check a verb's features against the agreement controller, the task of the generator is to pick out the agreement controller to inspect its features for agreement. Since the generator does not start with a verb form, it cannot use the morphosyntactic features of the verb (such as number) as retrieval cues to pick out the controller. Instead, the generator must use other retrieval cues such as category, case, or position-based features (see Hartsuiker et al., 2003 for evidence from German that case plays a role during number agreement formulation). Errors arise when an attractor matches these retrieval cues, which can at times lead to its erroneous retrieval (Badecker & Kuminiak, 2007). For example, an attractor embedded in a subject phrase shares category and case features with the subject (as part of the subject phrase, the attractor is nominative in the sentence planning process) (Wagers et al., 2009). Plural and singular attractors are equally likely as each other to match category, case, or position-based retrieval cues used to pick out the agreement controller, thus both types of attractor should be available to interfere with the agreement process and result in attraction errors. The gradation of the markedness effect could be explained if markedness influences a representation's likelihood of retrieval by making it more visible or prioritized in the

retrieval process, meaning plural representations would be more likely than singular ones to be erroneously retrieved, even if verb number is not used as a feature cue.[9]

In comprehension, on the other hand, the morphosyntactic features of the verb are available to be used as retrieval cues to check verb agreement. For example, when comprehending the sentence **the key to the cabinets are on the table*, upon encountering the plural verb *are*, the parser may use cues such as [Number: Plural] and [Case: Nominative] in addition to other cues to check for a plural subject in memory that satisfies the verb agreement (Wagers, 2008). In this example, *cabinets* is activated by the number cue, causing it in some cases to out-compete the subject head *key* (which does not match the number cue) for retrieval, thereby satisfying the verb agreement check and leading to a grammatical illusion. A markedness asymmetry arises from this framework when assuming a privative feature system, as singular number cannot be used as a retrieval cue, leading to less attractor activation when checking the verb form *is* than the verb form *are* (Wagers, 2008).[10] A similar effect can be achieved in a binary feature system if negative features cannot be targeted as retrieval cues or are weighted differently than positive ones during retrieval.

Although the number markedness asymmetry is less well documented in comprehension than in production, in the few studies that have tested for it, the asymmetry appears as a strong contrast, manifesting in a binary fashion with little or no attraction from singular distractors (e.g. Almeida & Tucker, 2017; Dillon et al., 2017; Wagers et al., 2009). The apparent comprehension–production contrast in the shape of the number markedness effect can arise naturally by assuming different retrieval cues are used in each domain (see Slioussar & Malko, 2016 for a similar account of gender attraction differences in production and comprehension); since number features can serve as a retrieval cues in comprehension, attractor number will have a more direct influence on the retrieval process. On the other hand, accounting for the apparent comprehension–production contrast within a representational account requires that the number representation of the subject be computed or processed differently when listening versus speaking. If the same representation were used in the same manner in both production and comprehension, we would expect to observe similar interference from singular attractors in both domains. Nevertheless, few studies have directly compared markedness effects in language production and comprehension using the same materials (but cf. Villata & Franck, 2020), and as mentioned in the *Introduction*, many comprehension studies do not include conditions with singular heads (e.g. Franck et al., 2010; Lago et al., 2015; Patson & Husband, 2015; Schlueter et al., 2019; Tanner et al., 2014; inter alia).

---

[9] Such prioritization could arise through the application of ranked constraints similar to Optimality Theory (Bresnan, 2000; McCarthy, 2002) (Badecker & Kuminiak, 2007) or the use of a special cue looking for a marked representation (Wagers et al., 2009).

[10] A potential challenge that arises for retrieval accounts assuming a privative number system is why partial activation of the attractor should yield more retrieval errors in production, where by our hypothesis number cues are not used in retrieval, than in comprehension trials with singular attractors, where singular number is assumed to be unmarked and hence not used. This distinction could arise if the attractor matches more retrieval cues in production, leading to more attractor activation than in comprehension and thus greater likelihood of interference.

It is therefore entirely possible that the number markedness effect in comprehension may not always manifest in a binary fashion but rather vary in strength similar to production. Additional research is necessary to assess the shape of the markedness effect in comprehension and the extent to which it parallels or differs from that in production.

5.3 Timing results

Our timing analysis revealed an effect of attraction-inducing environments on verb production even when no agreement errors were produced. In both the in-lab and web-based experiments, pre-verbal pauses in correct utterances were more likely in the same environments that elicited agreement errors: pre-verbal pauses were more likely in the mismatch conditions, and such slowdowns were similar in the presence of both plural and singular attractors, parallelling the reduced markedness asymmetry in the error analysis. The in-lab experiment elicited pre-verbal pauses in approximately 16% of mismatch trials and 5% of match trials, and the web-based experiment elicited pause in approximately 33% of mismatch trials and 13% of match trials.

In a post-hoc analysis combining the gap likelihood data from both experiments (available in the Supplementary Materials), we did not observe an interaction between match condition and experiment or a three-way interaction between match condition, N1 number, and experiment, suggesting that the size of the attraction effect and lack of markedness asymmetry were similar in both experiments. The overall effect of experiment was reliable such that gaps were more likely in the web-based experiment. A post-hoc analysis investigating gap durations in both experiments (see Supplementary Materials) found that pre-verbal pauses were not only more common but also were longer in the mismatch conditions (though this effect was not reliable for the Experiment 1 data when analyzed separately).

These results reinforce and extend previous findings that timing information can be used to index agreement attraction effects (e.g. Brehm & Bock, 2013; Haskell & MacDonald, 2003; Staub, 2009, 2010; Veenstra, Acheson, et al., 2014; Veenstra, Acheson, & Meyer, 2014) by showing that timing effects can be found utterance-medially in complete sentences elicited through naturalistic scene description. The presence of a timing effect on the production of correct verbs indicates that error rates underestimate the proportion of trials on which attraction pressure occurs. We believe that the relationship we observe between verb errors and timing is most easily (though not exclusively) captured by models of agreement attraction in which the attraction effect arises at the verb as opposed to earlier in sentence planning (e.g. in initial encoding of the subject phrase number).

5.3.1 Interpreting the attraction effect in the timing of correct verb agreement

The slowdowns before correct verbs in the mismatch conditions point to difficulty in subject–verb agreement in the presence of potential attractors, even when the correct form is ultimately produced. Interestingly, pre-verbal pauses occurred in only a subset of mismatch trials, suggesting that the process resulting in slowdown was not uniformly engaged in the presence of a mismatching

attractor. It is possible that there exists a smaller, more uniform timing effect engendering pre-verbal slowdowns that are too small to be detected by the forced-aligner, but at a minimum we can conclude that there is a subset of mismatch trials that show a stronger timing effect.

We identify two possible scenarios to explain the slowdown effects observed in our study. While both scenarios can be realized within representational and retrieval frameworks of attraction, we believe that they are most easily captured in models that assume that the process leading to attraction effects is active at the point of the agreement computation, as opposed to earlier in sentence planning. This is a natural prediction of accounts that attribute attraction effects to the agreement computation itself (e.g. retrieving the agreement controller or referencing an ambiguous subject number representation) as opposed to accounts that attribute errors to initial incorrect encoding of the subject number.

One scenario that could lead to pre-verbal slowdowns in correct utterances is that these slowdowns reflect time-consuming internal revision in which participants stop themselves before producing an error. Indeed, we observed fairly high rates of overt revisions of verb agreement errors in both experiments (approximately 43–45% of verb number errors were revised), suggesting that participants in our experiments were capable of detecting agreement errors mid-articulation. In this scenario, the revision process is engaged when an error is initially planned and then detected prior to articulation (e.g. by a post-encoding editor; Baars et al., 1975; Butterworth, 1981; Kempen & Huijbers, 1983; Levelt, 1989). This scenario predicts a greater likelihood of slowdowns in mismatch trials, where errors are more common, and that these slowdowns occur non-uniformly, since the conditions necessary to prompt revision are unlikely to be met on all mismatch trials. This scenario additionally predicts the observation from our exploratory post-hoc analysis that agreement errors tended to be produced more quickly than correct verbs (pre-verbal pauses were less likely), as errors reflect cases when incorrect number agreement is not detected by the error-monitoring process and thus revision is not engaged.

In order to instantiate this revision hypothesis within a model of agreement attraction, the model must be able to reach different conclusions about the verb form at different times (e.g. during initial computation and later when monitoring for errors). These different conclusions could arise if the representation used to check the agreement during monitoring differs from that used during the initial computation (e.g. if verb agreement is computed using grammatical number through agreement control but is checked against notional number via agreement concord). Differences may also arise if the agreement computation is performed again during a reanalysis check. This hypothesis requires i) that the process leading to attraction errors be repeated during reanalysis and ii) that there exists variability in the pressures leading to attraction such that different conclusions can be reached when computing the agreement a second time (see **Table 8** for ways to incorporate this within representational and retrieval frameworks). These requirements arise naturally if attraction errors result from the agreement process itself (as the generator references the subject head). If attraction effects arise from errors that occur earlier in sentence planning, such as an error in the

initial encoding of the subject number, in order to reach a different conclusion about verb number, it must be assumed that the subject number is re-encoded during the verb agreement check.

Another possible scenario to account for the timing effect in correct utterances is that slowdowns reflect the same pressures that lead to attraction errors. In this scenario, pre-verbal slowing can occur if these pressures result from time-consuming processes engaged during sentence planning (see **Table 8** for possible time-consuming processes within representational and retrieval frameworks). The non-uniformity of the slowing arises if there is gradation in these pressures, thereby allowing for uneven effects between trials (see **Table 8** for potential ways to account for timing variability). This scenario can be borne out in any attraction model that assumes interference whenever a mismatching attractor is present (whether or not it ultimately leads to an error) and that selecting the correct representation for agreement takes time. Under this hypothesis, in order for the attraction pressure to engender a timing effect observable directly prior to verb articulation, the time-consuming process that leads to attraction (e.g. inspection/encoding of the subject number feature, retrieval of the agreement controller) must be engaged during or shortly before verb planning. This is a natural prediction of accounts that attribute attraction effects to the computation of agreement itself (e.g. retrieving the agreement controller or referencing an ambiguous number feature). Accounts that attribute errors to initial incorrect encoding of the subject number must assume either that subject number encoding is only initiated during or after articulation of the subject phrase or that subject number is encoded a second time close to or during verb planning.

In this scenario, in order to explain the mismatch effect, the pressures leading to errors must be significantly stronger or only active in the mismatch conditions. This requirement is met by a marking and morphing account, in which the subject number representation should be unambiguous in match trials. Competition-based attraction frameworks (such as number encoding, percolation, and retrieval accounts), on the other hand, must assume that the competition process is either more difficult or only active when the attractor has a different number feature from the subject head. The first assumption can be borne out if the attractor's distinctive number feature boosts its activation in working memory, leading to greater competition in mismatch conditions. The second may arise if number features of the same value do not compete with each other. This hypothesis predicts no competition during subject number feature encoding in match conditions within a representational framework. It can also be reconciled with a retrieval framework if the retrieval mechanism looks in memory for a number feature to use for agreement, rather than retrieving the agreement controller itself and then inspecting it for its number feature.[11]

[11] While this framework represents a departure from traditional retrieval accounts, it is supported by prior findings that number and gender agreement errors are independent (Antón-Méndez et al., 2002), suggesting that an attractor's number feature can be accessed separately from other features of the attractor (both number and gender should be available if the attractor itself is retrieved and its features inspected for agreement). It additionally fits with evidence from comprehension studies showing that in cases of number agreement errors, the attractor is not treated as the semantic subject later in the sentence (Lau et al., 2008; Schlueter et al., 2019).

Framework	Time-consuming agreement process	Process activated during agreement formulation/ reanalysis	Sources of variability
Marking & morphing (Representational)	Lengthy subject number determination when plurality is close to the mid-point on the continuum between singular and plural More time for decision-making may result in fewer errors; more time for evidence accumulation may result in less ambiguous number representations	(Re)referencing or (re)valuation of an ambiguous subject number representation	Probabilistic error likelihood (accounts for variability in errors only) Variability in the initial bias of the computation process or in the rate of information accumulation informing the computation (e.g. Ratcliff et al., 1999) (accounts for variable errors and timing)
Subject number encoding error (Representational)	Lengthy competition between simultaneously-activated subject head and attractor representations when encoding the subject number feature More time to resolve competition may result in fewer encoding errors	(Re)encoding of the subject number	Noise in the activation of the attractor and subject head (accounts for variability in errors and timing)
Percolation (Representational)	Lengthy competition between the number features of the subject head and the upward-percolated attractor number feature during encoding More time to resolve competition may result in fewer encoding errors	(Re)encoding of the subject number	Noise in the activation of the attractor and subject head features (accounts for variability in errors and timing)
Retrieval	Lengthy competition between the subject head and the attractor for retrieval More time to resolve competition may result in fewer retrieval errors	(Re)retrieval of the agreement controller	Noise in the likelihood of erroneous retrieval and/or activation of the attractor and subject head (accounts for variability in errors and timing)

Table 8: Possible mechanisms to produce the timing results observed in our study in representational and retrieval models of attraction.

We believe that the interpretations of the gap effect in the two scenarios we introduce (as reflecting either i) internal revision or ii) attraction pressures active shortly before/during verb planning) are explained most naturally if the attraction effect arises at the verb. Nevertheless, the present results do not exclude the possibility that attraction pressures might occur earlier in the sentence. We did not test for timing effects earlier in the sentence, as earlier differences between the match and mismatch conditions cannot be confidently interpreted as attraction effects. Speech onset slowdowns in the mismatch conditions could reflect difficulty planning the subject phrase, as the speaker must keep track of two distinct number features (Staub, 2009), and timing effects in the duration of the subject phrase preceding the verb could be confounded with differences in word length between the match and mismatch conditions (especially in the SS vs. SP comparison). Nevertheless, in Kandel and Phillips (2022), we compared the articulation time-course of the in-lab responses (Experiment 1) in the mismatch and match conditions and observed longer mismatch durations for the second DP of the subject phrase (*the + N2*) and for the agreeing verb (*is/are*). The slowdowns at the verb likely reflect the same process underlying the gap effect. If the slowdown at the end of the subject phrase reflects attraction pressures, it could be attributed to either time-consuming encoding of the subject number feature or to verb planning occurring during articulation. We consequently still cannot distinguish from the available timing data whether the locus of the attraction effect is in the representation of the subject or in the agreement computation process of referencing/matching the subject features.

5.3.2 Speed-accuracy trade-off in agreement errors

The pattern of results across our experiments suggests that whether a speaker produces an attraction error in a given utterance may be subject to a speed-accuracy trade-off. We elicited fewer agreement errors but more pre-verbal pauses in the web-based experiment, where participants had a longer response window and articulated their responses more slowly than in the in-lab experiment. The hypothesis of a speed-accuracy tradeoff for agreement errors is supported by the preliminary evidence from our post-hoc exploratory analysis agreement errors tend to be produced more quickly than correct verb forms in attraction-inducing environments.[12] The influence of time pressure on error likelihood in our study aligns with evidence that attraction error effects are modulated by inhibitory control (Nozari & Omaki, 2022), which in turn is influenced by response pressure (e.g. Endres et al., 2020). These result patterns provide evidence that the two ways attraction pressure manifested in our study (errors and timing) are related and trade off with each other.

[12] It is important to keep in mind that our exploratory post-hoc analysis contained relatively few samples in the error condition, which could lead to an underestimation of slowdowns in error responses. Furthermore, forced-aligners may not be sensitive to lag times between words below a certain duration threshold, meaning that some of the gradation in the measurement of pre-verbal slowdowns was likely lost. In fact, the results of our analysis contrast with those of Staub (2009), who found similar response times in a forced-choice task for correct responses and agreement errors in attraction-inducing environments. The timing results for agreement errors in our study should thus be treated as preliminary.

A speed-accuracy trade-off can be reconciled with the two scenarios we proposed to explain slowdowns before correct verbs in attraction-inducing environments. If slowdowns reflect internal revision, a speed-accuracy trade-off could arise if the error-monitoring process that prompts revisions needs sufficient time to function effectively. In this scenario, increased time pressure may increase the likelihood that verb agreement errors will escape revision, resulting in a speed-accuracy trade-off. Alternatively, if slowdowns reflect the pressures leading to attraction effects, reducing planning time may make interference from an attractor more likely to be above the threshold required to result in an error at the point when a speaker makes a decision about verb form (**Table 8** presents possible time-consuming processes within representational and retrieval frameworks that could lead to greater accuracy with additional time).

6 Conclusion

The present study shows the effectiveness of a naturalistic scene-description task to elicit agreement attraction effects in both in-lab and web-based experimental settings. Our paradigm replicated the basic attraction error effect, and we observed parallel slowdowns in the articulation time-course of correct verbs, providing evidence that error measures underrepresent the extent of attraction effects. We additionally found evidence that the two ways that attraction pressure manifests (errors and timing) trade off, suggesting that they are related. Our timing results are naturally captured within an approach to agreement attraction when the attraction pressure arises at the point of computing agreement, rather than reflecting earlier errors in the encoding of the subject number (though the results can be reconciled with both types of accounts). Our study elicited similar interference from both singular and plural attractors, contradicting the conventional notion of a binary markedness effect and suggesting that the asymmetry may be more graded in naturalistic production. While more graded or absent markedness effects appear elsewhere in the literature, a binary markedness effect is undeniably present in many production studies (typically those using preamble paradigms) as well as some comprehension studies (though it is much less studied in in comprehension tasks). A challenge for future research is to define and explain the variability in the manifestation of the markedness effect.

The similarities between the results elicited in our in-lab and online experiments provide another demonstration that it is possible to conduct fairly complex and open-ended production tasks in an unsupervised web-based setting (see also Ziegler & Snedeker, 2018). The recordings collected were clear enough not only to understand what participants said but also to analyze their speech further to get more in-depth data about the production planning process itself. Web-based experimentation will allow for more efficient and easier participant recruitment in a part of the field that has not traditionally used this method, adding more flexibility and greater accessibility to the global population in production research than ever before.

Data accessibility statement

Data and Supplementary Materials are available at https://osf.io/jwnsz/.

Acknowledgements

We are very grateful to Akira Omaki, Bethany Dickerson, and Shota Momma for their advice regarding experiment design and/or analyses, to Phoebe Gaston for assistance with data collection, to Patrick Mair for his statistical advice, and to the reviewers and acting editor of this article for their helpful comments. This research was supported in part by National Science Foundation grant DGE-1448915 to the Maryland Language Science Center, C. Phillips, PI.

Competing interests

The authors have no competing interests to declare.

Author contributions

MK: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Project administration, Visualization, Writing – original draft, Writing – review & editing.

CW: Conceptualization, Data curation, Investigation, Software, Writing – original draft.

CP: Conceptualization, Supervision, Writing – original draft, Writing – review & editing.

References

Acuña-Fariña, J. C., Meseguer, E., & Carreiras, M. (2014). Gender and number agreement in comprehension in Spanish. *Lingua*, *143*, 108–128. DOI: https://doi.org/10.1016/j.lingua.2014.01.013

Almeida, D., & Tucker, M. (2017). The complex structure of agreement errors: Evidence from distributional analyses of agreement attraction in Arabic. In A. Lamont & K. Tetzloff (Eds.), *NELS 47: Proceedings of the Forty-Seventh Annual Meeting of the North-East Linguistic Society* (Vol. 3, pp. 45–54). GLSA (Graduate Linguistics Student Association), Department of Linguistics, University of Massachusetts.

Antón-Méndez, I., Nicol, J. L., & Garrett, M. F. (2002). The relation between gender and number agreement processing. *Syntax*, *5*, 1–25. DOI: https://doi.org/10.1111/1467-9612.00045

Baars, B. J., Motley, M. T., & MacKay, D. G. (1975). Output editing for lexical status in artificially elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior*, *14*, 382–391. DOI: https://doi.org/10.1016/S0022-5371(75)80017-X

Badecker, W., & Kuminiak, F. (2007). Morphology, agreement, and working memory retrieval in sentence production: Evidence from gender and case in Slovak. *Journal of Memory and Language*, *56*(1), 65–85. DOI: https://doi.org/10.1016/j.jml.2006.08.004

Bańkosz, Z., Nawara, H., & Ociepa, M. (2013). Assessment of simple reaction time in badminton players. *Trends in Sports Sciences*, *1*(20), 54–61.

Barker, J., Nicol, J., & Garrett, M. (2001). Semantic factors in the production of number agreement. *Journal of Psycholinguistic Research*, *30*(1), 91–114. DOI: https://doi.org/10.1023/A:1005208308278

Bergen, L., & Gibson, E. (2012, March). *Agreement errors as rational encoding errors*. [Poster presentation]. 25th Annual CUNY Conference on Human Sentence Processing, New York, NY, USA.

Bock, K., & Cutting, J. (1992). Regulating mental energy: Performance units in language production. *Journal of Memory and Language*, *31*(1), 99–127. DOI: https://doi.org/10.1016/0749-596X(92)90007-K

Bock, K., & Eberhard, K. M. (1993). Meaning, sound, and syntax in English number agreement. *Language and Cognitive Processes*, *8*(1), 57–99. DOI: https://doi.org/10.1080/01690969308406949

Bock, K., Eberhard, K. M., & Cutting, J. C. (2004). Producing number agreement: How pronouns equal verbs. *Journal of Memory and Language*, *51*(2), 251–278. DOI: https://doi.org/10.1016/j.jml.2004.04.005

Bock, J. K., Eberhard, K. M., Cutting, J. C., Meyer, A. S., & Schriefers, H. (2001). Some attractions of verb agreement. *Cognitive Psychology*, *43*, 83–128. DOI: https://doi.org/10.1006/cogp.2001.0753

Bock, K., & Miller, C. (1991). Broken agreement. *Cognitive Psychology*, *23*(1), 45–93. DOI: https://doi.org/10.1016/0010-0285(91)90003-7

Bock, K., Nicol, J., & Cutting, J. (1999). The ties that bind: Creating number agreement in Speech. *Journal of Memory and Language*, *40*(3), 330–346. DOI: https://doi.org/10.1006/jmla.1998.2616

Brehm, L., & Bock, K. (2013). What counts in grammatical number agreement? *Cognition*, *128*(2), 149–169. DOI: https://doi.org/10.1016/j.cognition.2013.03.009

Bresnan, J. (2000). Optimal syntax. In J. Dekkers, F. van der Leeuw & J. van de Weijer (Eds.), *Optimality theory: Phonology, syntax and acquisition* (pp. 334–385). Oxford University Press.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. DOI: https://doi.org/10.3758/BRM.41.4.977

Butterworth, B. (1981). Speech errors: Old data in search of new theories. *Linguistics*, *19*, 627–662. DOI: https://doi.org/10.1515/ling.1981.19.7-8.627

Clifton, C., Frazier, L., & Deevy, P. (1999). Feature manipulation in sentence comprehension. *Rivista di Linguistica*, *11*, 11–39.

Corley, M., & Scheepers, C. (2002). Syntactic priming in English sentence production: Categorical and latency evidence from an internet-based study. *Psychonomic Bulletin & Review*, *9*(1), 126–131. DOI: https://doi.org/10.3758/BF03196267

Den Dikken, M. (2001). “Pluringulars”, pronouns, and quirky agreement. *The Linguistics Review*, *18*, 19–41. DOI: https://doi.org/10.1515/tlir.18.1.19

Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, *69*(2), 85–103. DOI: https://doi.org/10.1016/j.jml.2013.04.003

Dillon, B., Staub, A., Levy, J., & Clifton, C. (2017). Which noun phrases is the verb supposed to agree with?: Object agreement in American English. *Language*, *93*(1), 65–96. DOI: https://doi.org/10.1353/lan.2017.0003

Duffield, C. J. (2013). Beyond the Subject: The Interaction of Syntax and Semantics in the Production of English Verb Agreement. [Unpublished doctoral dissertation]. University of Colorado.

Eberhard, K. M. (1993). *The specification of grammatical number in English* [Unpublished doctoral dissertation]. Michigan State University.

Eberhard, K. M. (1997). The marked effect of number on subject–verb agreement. *Journal of Memory and Language*, *36*(2), 147–164. DOI: https://doi.org/10.1006/jmla.1996.2484

Eberhard, K. M. (1999). The accessibility of conceptual number to the processes of subject–verb agreement in English. *Journal of Memory and Language*, *41*, 560–578. DOI: https://doi.org/10.1006/jmla.1999.2662

Eberhard, K. M., Cutting, J. C., & Bock, K. (2005). Making syntax of sense: Number agreement in sentence production. *Psychological Review*, *112*(3), 531–559. DOI: https://doi.org/10.1037/0033-295X.112.3.531

Endres, D. N., Byrne, K. A., Anaraky, R. G., Adesegun, N., Six, S. G., & Tibbett, T. P. (2020). Stop the clock because I can't stop: time pressure, but not monitoring pressure, impairs response inhibition performance. *Journal of Cognitive Psychology*, *32*(7), 627–644. DOI: https://doi.org/10.1080/20445911.2020.1810692

Enochson, K., & Culbertson, J. (2015). Collecting psycholinguistic response time data using Amazon Mechanical Turk. *PLoS ONE*, *10*(3), e0116946. DOI: https://doi.org/10.1371/journal.pone.0116946

Fairs, A., & Strijkers, K. (2021, April 27). Can we use the internet to study speech production? Yes we can! Evidence contrasting online versus laboratory naming latencies and errors. DOI: https://doi.org/10.31234/osf.io/2bu4c

Francis, W. N. (1986). Proximity concord in English. *Journal of English Linguistics*, *19*(2), 309–317. DOI: https://doi.org/10.1177/007542428601900212

Franck, J., Soare, G., Frauenfelder, U. H., & Rizzi, L. (2010). Object interference in subject–verb agreement: The role of intermediate traces of movement. *Journal of Memory and Language*, *62*(2), 166–182. DOI: https://doi.org/10.1016/j.jml.2009.11.001

Franck, J., Vigliocco, G., & Nicol, J. (2002). Subject-verb agreement errors in French and English: The role of syntactic hierarchy. *Language and Cognitive Processes*, *17*(4), 371–404. DOI: https://doi.org/10.1080/01690960143000254

Gelman, A., Jakulin, A., Grazia, P., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*(4), 1360–1383. DOI: https://doi.org/10.1214/08-AOAS191

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*, 8051–8056. DOI: https://doi.org/10.1073/pnas.1216438110

Gibson, E., & Fedorenko, E. (2013). The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, *28*(1–2), 88–124. DOI: https://doi.org/10.1080/01690965.2010.515080

Gibson, E., Piantadosi, S. T., & Fedorenko, E. (2012). Quantitative methods in syntax/semantics research: A response to Sprouse and Almeida (2012). *Language and Cognitive Processes*. DOI: https://doi.org/10.1080/01690965.2012.704385

Gillespie, M., & Pearlmutter, N. J. (2011). Hierarchy and scope of planning in subject-verb agreement production. *Cognition*, *118*(3), 377–397. DOI: https://doi.org/10.1016/j.cognition.2010.10.008

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2020). rstanarm: Bayesian applied regression modeling via Stan. R package version 2.21.1. https://mc-stan.org/rstanarm

Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1411–1423. DOI: https://doi.org/10.1037/0278-7393.27.6.1411

Greenberg, J. H. (1966). *Language universals*. De Gruyter Mouton.

Hartsuiker, R. J., Schriefers, H. J., Bock, K., & Kikstra, G. M. (2003). Morphophonological influences on the construction of subject-verb agreement. *Memory & Cognition*, *31*(8), 1316–1326. DOI: https://doi.org/10.3758/BF03195814

Haskell, T. R., & MacDonald, M. C. (2003). Conflicting cues and competition in subject–verb agreement. *Journal of Memory and Language*, *48*(4), 760–778. DOI: https://doi.org/10.1016/S0749-596X(03)00010-X

Haskell, T. R., Thornton, R., & MacDonald, M. C. (2010). Experience and grammatical agreement: Statistical learning shapes number agreement production. *Cognition*, *114*, 151–164. DOI: https://doi.org/10.1016/j.cognition.2009.08.017

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. DOI: https://doi.org/10.1017/S0140525X0999152X

Jaccard, J. (2001). *Interaction Effects in Logistic Regression*. Sage. DOI: https://doi.org/10.4135/9781412984515

Kaan, E. (2002). Investigating the effects of distance and number interference in processing subject-verb dependencies: An ERP study. *Journal of Psycholinguistic Research*, *31*, 165–193. DOI: https://doi.org/10.1023/A:1014978917769

Kandel, M., & Phillips, C. (2022, April 5). *Number attraction in verb and anaphor production*. OSF. DOI: https://doi.org/10.31219/osf.io/d97gf

Kempen, G., & Huijbers, P. (1983). The lexicalization process in sentence production and naming: Indirect election of words. *Cognition*, *14*, 185–209. DOI: https://doi.org/10.1016/0010-0277(83)90029-X

Lago, S., Shalom, D., Sigman, M., Lau, E., & Phillips, C. (2015). Agreement attraction in Spanish comprehension. *Journal of Memory and Language*, *82*, 133–149. DOI: https://doi.org/10.1016/j.jml.2015.02.002

Lau, E., Wagers, M., Stroud, C., & Phillips, C. (2008). Agreement and the subject of confusion. [Spoken presentation], 21st Annual CUNY Conference on Human Sentence Processing, Chapel Hill, NC, USA.

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

Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. *Proceedings of the 13th Conference on Empirical Methods in Natural Language Processing*, 234–243. DOI: https://doi.org/10.3115/1613715.1613749

Linnman, C., Carlbring, P., Åhman, Å., Andersson, H., & Andersson, G. (2006). The stroop effect on the internet. *Computers in Human Behavior*, *22*(3), 448–455. DOI: https://doi.org/10.1016/j.chb.2004.09.010

McAuliffe, M., Socolof, M., Stengel-Eskin, E., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner [Computer program]. Version 1.0.0, retrieved 05 May 2017 from http://montrealcorpustools.github.io/Montreal-Forced-Aligner/

McCarthy, J. (2002). *A thematic guide to optimality theory*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511613333

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, *29*(2), 111–123. DOI: https://doi.org/10.1023/A:1005184709695

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, *48*(1), 67–91. DOI: https://doi.org/10.1016/S0749-596X(02)00515-6

Nicol, J., Forster, K., & Veres, C. (1997). Subject–verb agreement processes in comprehension. *Journal of Memory and Language*, *36*(4), 569–587. DOI: https://doi.org/10.1006/jmla.1996.2497

Nozari, N., & Omaki, A. (2022, January 14). An investigation of the dependency of subject-verb agreement on inhibitory control processes in sentence production. DOI: https://doi.org/10.31234/osf.io/9pcmg

Paspali, A., & Marinis, T. (2020). Gender agreement attraction in Greek comprehension. *Frontiers in Psychology*, *11*, Article 717. DOI: https://doi.org/10.3389/fpsyg.2020.00717

Pearlmutter, N., Garnsey, S., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and Language*, *41*(3), 427–456. DOI: https://doi.org/10.1006/jmla.1999.2653

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods*. DOI: https://doi.org/10.3758/s13428-018-01193-y

Pfau, R. (2009). *Grammar as Processor: A Distributed Morphology Account of Speech Errors*. John Benjamins Publishing Company. DOI: https://doi.org/10.1075/la.137

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, *106*(2), 261–300. DOI: https://doi.org/10.1037/0033-295X.106.2.261

Reips, U.-D. (2002). Standards for internet-based experimenting. *Experimental Psychology*, *49*(4), 243–256. DOI: https://doi.org/10.1026/1618-3169.49.4.243

Ryskin, R. A., Bergen, L., & Gibson, E. (2021, October 8). Agreement errors are predicted by rational inference in sentence processing. DOI: https://doi.org/10.31234/osf.io/uaxsq

Schlueter, Z., Parker, D., & Lau, E. (2019). Error-driven retrieval in agreement attraction rarely leads to misinterpretation. *Frontiers in Psychology*, *10*. DOI: https://doi.org/10.3389/fpsyg.2019.01002

Shen, E., Staub, A., & Sanders, L. (2013). Event-related brain potential evidence that local nouns affect subject-verb agreement processing. *Language and Cognitive Processes*, *28*(4), 498–524. DOI: https://doi.org/10.1080/01690965.2011.650900

Skitka, L., & Sargis, E. (2006). The internet as psychological laboratory. *Annual Review of Psychology*, *57*, 529–555. DOI: https://doi.org/10.1146/annurev.psych.57.102904.190048

Slioussar, N. (2018). Forms and features: The role of syncretism in number agreement attraction. *Journal of Memory and Language*, *101*, 51–63. DOI: https://doi.org/10.1016/j.jml.2018.03.006

Slioussar, N., & Malko, A. (2016). Gender agreement attraction in Russian: Production and comprehension evidence. *Frontiers in Psychology*, *7*, Article 1651. DOI: https://doi.org/10.3389/fpsyg.2016.01651

Smith, G., Franck, J., & Tabor, W. (2018). A self-organizing approach to subject-verb number agreement. *Cognitive Science*, *42*(S4), 1043–1074. DOI: https://doi.org/10.1111/cogs.12591

Smith, G., Franck, J., & Tabor, W. (2021). Encoding interference effects support self-organized sentence processing. *Cognitive Psychology*, *124*, 101356. DOI: https://doi.org/10.1016/j.cogpsych.2020.101356

Solomon, E. S., & Pearlmutter, N. J. (2004). Semantic integration and syntactic planning in language production. *Cognitive Psychology*, *49*(1), 1–46. DOI: https://doi.org/10.1016/j.cogpsych.2003.10.001

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*(1), 155–167. DOI: https://doi.org/10.3758/s13428-010-0039-7

Staub, A. (2009). On the interpretation of the number attraction effect: Response time evidence. *Journal of Memory and Language*, *60*(2), 308–327. DOI: https://doi.org/10.1016/j.jml.2008.11.002

Staub, A. (2010). Response time distributional evidence for distinct varieties of number attraction. *Cognition*, *114*(3), 447–454. DOI: https://doi.org/10.1016/j.cognition.2009.11.003

Stemberger, J. P. (1985). Bound morpheme loss errors in normal and agrammatic speech: One mechanism or two? *Brain and Language*, *25*(2), 246–256. DOI: https://doi.org/10.1016/0093-934X(85)90084-7

Tanner, D., Nicol, J., & Brehm, L. (2014). The time-course of feature interference in agreement comprehension: Multiple mechanisms and asymmetrical attraction. *Journal of Memory and Language*, *76*, 195–215. DOI: https://doi.org/10.1016/j.jml.2014.07.003

Thornton, R., & MacDonald, M. C. (2003). Plausibility and grammatical agreement. *Journal of Memory and Language*, *48*(4), 740–759. DOI: https://doi.org/10.1016/S0749-596X(03)00003-2

Veenstra, A., Acheson, D., Bock, K., & Meyer, A. (2014). Effects of semantic integration on subject-verb agreement: Evidence from Dutch. *Language, Cognition, and Neuroscience*, *29*(3), 355–380. DOI: https://doi.org/10.1080/01690965.2013.862284

Veenstra, A., Acheson, D., & Meyer, A. (2014). Keeping it simple: Studying grammatical encoding with lexically reduced item sets. *Frontiers in Psychology*, *18*, Article 783. DOI: https://doi.org/10.3389/fpsyg.2014.00783

Vesker, M., Bahn, D., Degé, F., Kauschke, C., & Schwarzer, G. (2019). Identification of emotional facial expressions in a lab and over the internet. *Journal of the Higher School of Economics*, *16*(3), 571–583. DOI: https://doi.org/10.17323/1813-8918-2019-3-571-583

Vigliocco, G., Butterworth, B., & Semenza, C. (1995). Constructing subject–verb agreement in speech: The role of semantic and morphological factors. *Journal of Memory and Language*, *34*(2), 186–215. DOI: https://doi.org/10.1006/jmla.1995.1009

Vigliocco, G., & Franck, J. (1999). When sex and syntax go hand in hand: Gender agreement in language production. *Journal of Memory and Language*, *40*(4), 455–478. DOI: https://doi.org/10.1006/jmla.1998.2624

Vigliocco, G., Hartsuiker, R., Jarema, G., & Kolk, H. (1996). One or more labels on the bottles? Notional concord in Dutch and French. *Language and Cognitive Processes*, *11*, 407–442. DOI: https://doi.org/10.1080/016909696387169

Vigliocco, G., & Nicol, J. (1998). Separating hierarchical relations and word order in language production: Is proximity concord syntactic or linear? *Cognition*, *68*(1), B13–B29. DOI: https://doi.org/10.1016/S0010-0277(98)00041-9

Villata, S., & Franck, J. (2020). Similarity-based interference in agreement comprehension and production: Evidence from object agreement. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(1), 170–188. DOI: https://doi.org/10.1037/xlm0000718

Villata, S., Tabor, W., & Franck, J. (2018). Encoding and retrieval interference in sentence comprehension: Evidence form agreement. *Frontiers in Psychology*, *9*. DOI: https://doi.org/10.3389/fpsyg.2018.00002

Vogt, A., Hauber, R. C., Kuhlen, A. K., & Abdel Rahman, R. (2021, February 9). Internet based language production research with overt articulation: Proof of concept, challenges, and practical advice. DOI: https://doi.org/10.31234/osf.io/cyvwf

Wagers, M. (2008). The structure of memory meets memory for structure in linguistic cognition. [Unpublished doctoral dissertation]. University of Maryland, College Park.

Wagers, M., Lau, E., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, *61*(2), 206–237. DOI: https://doi.org/10.1016/j.jml.2009.04.002

Wilkinson, R. T., & Allison, S. (1989). Age and simple reaction time: Decade differences for 5,325 subjects. *Journal of Gerontology*, *44*(2), 29–35. DOI: https://doi.org/10.1093/geronj/44.2.P29

Zehr, J., & Schwarz, F. (2018). PennController for Internet Based Experiments (IBEX). https://doi.org/10.17605/OSF.IO/MD832

Ziegler, J., & Snedeker, J. (2018). How broad are thematic roles? Evidence from structural priming. *Cognition*, *179*, 221–240. DOI: https://doi.org/10.1016/j.cognition.2018.06.019