Journal of Memory and Language 127 (2022) 104370

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Number attraction in verb and anaphor production

Margaret Kandel [a,*], Colin Phillips [b]

[a] Harvard University, Cambridge, MA 02138, USA
[b] University of Maryland, College Park, MA 20742, USA

ARTICLE INFO

Keywords:
Agreement attraction
Verb agreement
Reflexive pronouns
Anaphora
Language production

ABSTRACT

Prior production research using the preamble-completion paradigm has elicited similar number attraction effects for both verbs and anaphora. However, this paradigm relies on comprehension and memory processes in addition to language production, making it difficult to assess the extent to which the observed attraction effects are caused by factors active during more natural production. In four production experiments, we compared number attraction effects on subject–verb and reflexive–antecedent agreement using a novel scene-description task in addition to a more traditional preamble elicitation paradigm. While the results from the preamble task align with prior findings, the more naturalistic scene-description task produced a contrast between the two dependency types, with robust verb attraction but very low rates of anaphor attraction. In addition to analyzing agreement error distributions, we also analyzed the production time-course of participant responses, finding timing effects that pattern with error distributions, even when no error is present. We discuss potential sources of variable susceptibility to number attraction, suggesting that differences may arise from the time-course of information processing across tasks and linguistic dependencies.

Introduction

Sentences require the formation of relations between words and phrases, often signaled via inflected word forms that mark agreement with another part of the sentence. In a simple world, these relations would be formed in the same fashion across languages, types of linguistic relations, and tasks, such as speaking and understanding. In that case, we might expect relationship formation to display similar behavioral profiles across all these domains. Nevertheless, psycholinguistic findings suggest that the behavioral profiles of dependency formation are not quite so uniform (see e.g. Franck et al., 2002; Omaki et al., 2014; Vasishth et al., 2010 for cross-linguistic differences in dependency formation; Parker & Phillips, 2016; Xiang et al., 2009 for cross-dependency differences; Slioussar & Malko, 2016 for production–comprehension differences; inter alia). Behavioral contrasts might point to variability in the underlying processes of relationship formation, however it is also possible that the same underlying mechanisms may yield different behavioral profiles as a function of dependency type or task. In the present study, we investigate the formation of number agreement in two types of dependencies: reflexive–antecedent dependencies and subject–verb agreement. Using a mixture of evidence from error profiles and speech timing, we observe a contrast between the two dependency types in their susceptibility to attraction effects, counter to previous findings (e.g. Bock et al., 2006; Bock et al., 1999). We argue that this variability may reduce to the time-course of information processing across different tasks and linguistic dependencies.

Attraction effects occur when nearby or intervening material interferes with normal agreement processes (Jesperson, 1913/1916). This phenomenon has primarily been investigated in the processing of subject–verb agreement relations. Consider the sentence in (1), an example of number attraction in subject–verb agreement; in this sentence, the verb *are* agrees in number with the more proximal, or local, plural noun *cabinets* instead of the singular subject head noun *key*.

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

Verb agreement errors like that in (1) have been well-documented across a range of languages both in the lab as well as in written texts and natural speech (Bock & Cutting, 1992; Bock & Eberhard, 1993; Bock & Miller, 1991; Den Dikken, 2001; Francis, 1986; inter alia), and many studies have been devoted to establishing what factors influence the likelihood of producing such errors (see Eberhard et al., 2005 for review). Number attraction errors occur when the presence of other (typically nearby) nouns disrupts the processing of the agreement

* Corresponding author.
E-mail address: mkandel@g.harvard.edu (M. Kandel).

https://doi.org/10.1016/j.jml.2022.104370
Received 13 September 2021; Received in revised form 4 August 2022; Accepted 5 August 2022
Available online 13 September 2022
0749-596X/© 2022 Elsevier Inc. All rights reserved.

relation between the subject and verb. They are commonly found in environments where one or more nouns intervene between the head subject noun and the verb, allowing the interfering noun(s) to compete with the head noun when determining verb agreement (e.g. Bock & Miller, 1991; Eberhard et al., 2005; Franck et al., 2002). Attraction effects are indexed by a greater number of errors when the head and interfering nouns mismatch in number (hereafter mismatch environments) than when they have the same number (match environments). It has been noted that agreement errors are often more common in mismatch environments with a singular head noun and plural interfering noun than in those with a plural head noun and singular interfering noun; this difference is known as the markedness effect, and it has been found to be fairly reliable across languages (e.g. Bock & Miller, 1991; Eberhard, 1997; Hartsuiker et al., 2003; Vigliocco et al., 1996; inter-alia; but cf. Franck et al., 2002). Grammatical illusion or facilitation effects have been found for verb attraction errors in a variety of comprehension measures including eye-tracking, self-paced reading, and EEG (e.g. Clifton et al., 1999; Dillon et al., 2013; Kaan, 2002; Pearlmutter et al., 1999; Shen et al., 2013; Wagers et al., 2009; inter alia).

In a side-by-side comparison of pronoun and verb production, Bock et al. (1999) found that reflexive pronoun number errors, as in (2), can be elicited in the lab and occur in the same environments at roughly the same rate as verb agreement errors (this finding was replicated by Bock et al., 2006 for both British and American English speakers). The reflexive pronoun error rates observed by Bock et al. (1999) coincide with mean verb number error rates in a meta-analysis of many other studies (Eberhard et al., 2005).

(2) *The actor in the soap operas watched themselves. (Bock et al. 1999)

At first blush, it is not surprising that reflexive number agreement should behave similarly to verb number agreement, as reflexive–antecedent and subject–verb dependencies share a number of surface similarities.[1] Both verbs and reflexives agree with the person, number, and (depending on the language) gender of the subject of the same clause (English reflexives may also have non-subject antecedents, but such cases are more rare). However, it is not clear that verb and reflexive number processing always occurs in a similar fashion: empirical evidence for a comprehension analog of reflexive attraction is more mixed than verb attraction, with some studies showing a lack of grammaticality illusions for reflexive errors (e.g. Cunnings & Sturt, 2014; Dillon et al., 2013, Experiment 2; Sturt, 2003; Xiang et al., 2009), whereas others elicit consistent illusions for reflexives (e.g. Parker & Phillips, 2017; though this required increasing the level of feature mismatch) or find that there is no attraction difference between reflexives and verbs (Jäger et al., 2020). In contrast to the comprehension literature, the production generalization appears remarkably straightforward.

Furthermore, while verb attraction effects in production have been found in corpora of spontaneous speech (see discussion in Bock & Miller, 1991; Haskell et al., 2010; Pfau, 2009), replicated multiple times in in-lab experiments (Eberhard et al., 2005), and elicited with different production paradigms (see *Previous production paradigms*), to our knowledge, the presence of reflexive attraction has not been attested in speech corpora and has only been assessed experimentally by Bock et al. (1999) and Bock et al. (2006). The responses in Bock et al.'s (1999) and Bock et al.'s (2006) experiments were elicited using a preamble completion paradigm (described in *Previous production paradigms*), which differs in a number of ways from naturalistic production. Crucially, the paradigm involves memory and language comprehension in addition to production, as participants must parse and repeat a preamble fragment to be completed as a full sentence. For the reflexive responses, this preamble fragment included the entire sentence with the exception of the sentence-final reflexive (e.g. *the actor in the soap operas watched*), leaving only one word for participants to generate and plan on their own (similar to a fill-in-the-blank task). This task could lead to performance that differs from naturalistic production, in which speakers' sentences are guided by a thought or idea they would like to convey rather than prepackaged linguistic material. Indeed, the additional task demands of the preamble paradigm make it difficult to disentangle the extent to which the observed attraction effects reflect processes active during natural production as opposed to deficits in parsing or memory processes (see e.g. Ryskin et al., 2021 for arguments that attraction errors in preamble completion paradigms may stem from preamble misinterpretations). We consequently set out to assess whether reflexives still show attraction effects comparable to verbs when they are elicited in a more naturalistic fashion.

In a series of four experiments, we compared the attraction susceptibility of the two dependency types in both the traditional preamble elicitation paradigm applied by Bock et al. (1999) as well as a more naturalistic scene-description paradigm. We not only analyzed error rates, but we also applied a forced aligner (McAuliffe et al., 2017) to look for attraction effects in production time-course, even when no error was produced. In both paradigms, we found evidence of verb attraction effects, both in the error rate measure as well as in timing measures of correct productions. The reflexive attraction profile, on the other hand, resembled that of verbs only in the preamble paradigm. The scene-description paradigm did not elicit obvious reflexive attraction effects reflected in either errors or production time-course. These results reinforce the importance of considering paradigm-specific task demands when interpreting results and suggest that reflexive–antecedent and subject–verb dependencies involve distinct formation processes with different susceptibility to attraction.

Models of attraction

Contemporary models of attraction can be roughly divided into two classes: representational accounts, which attribute errors to the representation of the features of the subject noun phrase, and retrieval accounts, which attribute errors to the retrieval of the features of the subject noun phrase. While we do not aim to distinguish between representational or retrieval accounts of attraction in the present paper, we introduce them here as a framework in which to consider our findings. In the *General discussion*, we address how our results may be explained under either class of model. We focus here on number agreement, which has received the most attention, but cross-linguistic studies show that attraction effects are by no means limited to number attraction (e.g. Acuña-Fariña et al., 2014; Badecker & Kuminiak, 2007; Paspali & Marinis, 2020; Slioussar & Malko, 2016; Vigliocco & Franck, 1999).

Representational accounts of attraction suggest that the representation of subject number is either incorrect or ambiguous. There are different theories of how this incorrect or ambiguous representation arises. Percolation accounts (e.g. Bock & Eberhard, 1993; Franck et al., 2002; Nicol et al., 1997; Vigliocco & Nicol, 1998) suggest that the number feature of the local noun percolates upward in the syntactic structure so that it is retrieved in place of the subject noun number feature when the parser or generator checks or retrieves the number of the subject phrase. Other accounts rely on encoding errors to explain the misrepresentation of the subject phrase number feature (e.g. Gillespie & Pearlmutter, 2011; Solomon & Pearlmutter, 2004), suggesting that the subject number feature is encoded incorrectly due to the simultaneous activation of the features of both the local and subject noun phrases.

[1] In this paper, we use the term number agreement to refer to the matching of number features between a target (e.g. a verb or reflexive) and a controller with which it covaries (e.g. a subject or linguistic antecedent). By referring to feature matching of both verbs and pronouns as number agreement, we are not committing to the claim that this matching occurs via the same set of mechanisms in both subject–verb and antecedent–pronoun dependencies.

Continuum accounts such as the Marking and Morphing model (e.g. Eberhard et al., 2005) propose that plurality is a continuous property, with the plurality of the subject phrase depending on the head noun number, local noun number, and notional factors such as collectivity. Under this account, attraction errors are more likely when the plural value for the subject phrase is more ambiguous, such as when the head and local noun phrases mismatch in number (see Hammerly et al., 2019 for a theory linking continuous representation of number and evidence accumulation rate using drift diffusion modeling).

Retrieval accounts, on the other hand, claim that the representation of the subject is neither incorrect nor ambiguous, but rather that errors arise during agreement formation. Retrieval accounts of attraction errors (e.g. Badecker & Kuminiak, 2007; Dillon et al., 2013; Wagers et al., 2009; inter alia) are based on theories of content-addressable memory systems with cue-based memory retrieval (McElree, 2000; McElree et al., 2003). Under such models, when the parser or generator accesses the agreement controller to check or form an agreement relation, it retrieves the item in memory that best matches some set of retrieval cues. Retrieval accounts suggest that this process can sometimes go awry, causing the incorrect element to be retrieved; retrieval errors are particularly likely when there are multiple elements in memory with similar features that match (or partially match) the retrieval cues. In a retrieval framework, production errors can occur if the generator accidentally retrieves the wrong item's number feature for agreement when planning the form of the agreement target. In comprehension, grammaticality illusions occur for sentences with agreement errors when the parser encounters the agreement target and retrieves the local noun's number feature when checking the features of the controller; the local noun matches the number feature on the verb, thus satisfying the verb agreement.

An important factor that distinguishes between models of attraction is whether attraction errors arise during the number agreement computation itself or whether they reflect an earlier problem in the encoding of the subject number. For instance, in retrieval accounts and some instantiations of Marking and Morphing, errors are produced when the agreement process goes awry, either due to picking out the incorrect item from memory for agreement or choosing the incorrect number when referencing an ambiguous subject number representation. By contrast, in representational accounts such as percolation and encoding error accounts, agreement errors occur because the subject is assigned the wrong number. This incorrect number leads to an agreement error later in the sentence when it is referenced during the computation of the agreement target's number. In this latter circumstance, the process leading to an attraction error is not the number agreement computation itself; the error is a consequence of an earlier problem in sentence processing.

Speech timing data can help identify when the process leading to attraction errors arises during sentence planning. If the pressure that at times results in attraction errors is active during the number agreement computation, then we may expect to see slowdowns in the production time-course of the agreement target even when the correct form is produced. As discussed by Kandel et al. (2022), slowdowns localized directly before the production of correct verbs may reflect interference in the agreement computation (e.g. slowing due to a lengthy competition process or lengthy number determination when referencing a subject phrase with more ambiguous plurality). Alternatively, slowdowns could reflect an internal revision process as the speaker stops themselves from producing an error, which requires that the generator reach different conclusions about the correct form of the agreement target at different times (during initial computation and in a later checking of the form), a requirement easily met in frameworks in which attraction errors arise from pressures active during the number agreement computation itself (Kandel et al., 2022). If attraction errors reflect pressures active earlier in sentence planning during initial encoding of the subject number, on the other hand, this pressure is less likely to influence the timing of number computation directly and thus may be less likely to result in slowdowns before the agreement target. In the present study, we analyze not only agreement errors but also speech timing data in order to investigate whether attraction pressure is active at similar times for reflexives and verbs. We additionally use timing data to investigate the influence of elicitation paradigm on planning and whether task demands influence when attraction pressures are active.

Previous production paradigms

The majority of in-lab production studies investigating attraction effects have used some variation of the preamble elicitation paradigm (e.g. Bock & Cutting, 1992; Bock & Eberhard, 1993; Bock & Miller, 1991). In this paradigm, a participant hears a sentence fragment, or preamble, which they are instructed to repeat and complete as a full sentence; for instance, a participant may produce a sentence like (1) in response to the preamble *the key to the cabinets*. Different versions of this paradigm have been used to elicit verb agreement errors. Verb-eliciting preambles typically consist of a complex subject phrase containing two or more nouns whose number is manipulated to create the different experimental conditions. In early applications of the paradigm (e.g. Bock & Miller, 1991), participants could complete preambles using whatever sentence completions they desired. Later studies have constrained the ways in which participants can complete preambles — for instance, by telling participants to use a specific verb (e.g. *to be*) (e.g. Franck et al., 2002), by providing participants with a verb stem or infinitive and instructing that completions use these verbs in a specified tense (e.g. Hartsuiker et al., 2001; Thornton & MacDonald, 2003), or by providing an adjective or past participle to be included in the preamble completion with an inflected form of the verb *to be* (e.g. Barker et al., 2001; Brehm & Bock, 2013; Hartsuiker & Barkhuysen, 2006; Veenstra, Acheson, & Meyer, 2014; Vigliocco et al., 1996; Vigliocco & Nicol, 1998). Other variations of the paradigm include asking participants not to simply repeat and complete the preamble but to perform other tasks with the provided linguistic material, such as forming a question (e.g. Haskell & MacDonald, 2003; Vigliocco & Nicol, 1998).

Bock et al. (1999) used the preamble paradigm to investigate reflexive pronoun number agreement in addition to verb agreement (Table 1). The verb-eliciting preambles consisted of a complex subject noun phrase containing both a head and local noun, which participants were instructed to repeat and complete as a full sentence. The reflexive-eliciting preambles used the same subject phrases as the verb-eliciting preambles plus a past-tense verb; participants were instructed to repeat and complete these preambles using a reflexive pronoun. The number of the head and interfering nouns in the preambles were manipulated to create four conditions: SS, SP, PP, and PS, where the first letter stands for the number of the head noun and the second for the number of the interfering, local noun (S = singular, P = plural). Using this paradigm, Bock et al. (1999) observed similar number error rates for verbs and reflexives in the same environments; this effect was replicated in Bock et al. (2006), and these error rates align with the mean verb number error rates in Eberhard et al.'s (2005) meta-analysis of verb attraction studies (Table 2).

However, as previously mentioned, this task involves processes that

Table 1
Sample verb-eliciting and reflexive-eliciting preambles from Bock et al. (1999).

Condition	Match Condition	Verb Preamble	Reflexive Preamble
SS	match	The actor in the soap opera…	The actor in the soap opera watched…
SP	mismatch	The actor in the soap operas…	The actor in the soap operas watched…
PP	match	The actors in the soap operas…	The actors in the soap operas watched…
PS	mismatch	The actors in the soap opera…	The actors in the soap opera watched…

Table 2
Agreement error rates from Bock et al.'s (1999) simple count noun conditions (total response counts in parentheses) and the mean verb error rates for sentences with simple count nouns estimated in Eberhard et al.'s (2005) meta-analysis.

Condition	Match Condition	Verb Error Rate (Bock et al. 1999)	Reflexive Error Rate (Bock et al. 1999)	Mean Verb Error Rate (Eberhard et al., 2005)
SS	match	2 % (154)	2 % (361)	1 %
SP	mismatch	10 % (140)	17 % (246)	13 %
PP	match	1 % (152)	1 % (281)	2 %
PS	mismatch	1 % (161)	4 % (327)	3 %

differ from those engaged during natural speech production. In natural production, speakers typically start out with a thought or idea that they then convert into linguistic output, selecting the appropriate lexical items and structure to express their message and assembling them together. In the preamble paradigm, on the other hand, much of the sentence material is already provided to participants, particularly in the reflexive conditions, in which the preambles consist of all but the final word of the sentence. Participants must comprehend, remember, and repeat a preamble before completing it. The task thus involves comprehension as well as a memory task, which could result in sentence completion strategies that differ from natural production. For instance, the relative timing of constituent planning may differ if participants focus on first remembering and repeating the preamble before planning their sentence continuation, which could result in sentence elements that appear after the preamble structure being planned later than they otherwise would under more natural circumstances. Indeed, since this task requires participants to interpret a preamble structure to complete their responses, it is unclear the extent to which elicited attraction errors reflect errors in sentence planning as opposed to misinterpretation of the preamble structure (e.g. Ryskin et al., 2021).

Some recent studies of verb agreement have moved away from the preamble paradigm, applying picture description tasks to elicit verb number agreement errors. Some tasks, while not providing participants with any pre-packaged linguistic material to remember and repeat, still involve operations that do not resemble naturalistic language production. For instance, Gillespie and Pearlmutter (2011) observed attraction errors in a task that used pictures to elicit complex subject phrases that were then completed by the participant as full sentences using whatever completion they desired, similar to a preamble paradigm. Although participants were not explicitly given linguistic material, the experiment's visual cues provided a word-for-word outline of the sentences' subject phrases, and the images in the experiment did not provide a message for the participants to convey in their responses. Even the form of the subject phrases was determined based on task-specific criteria rather than a guiding message: when presented with a display of two pictures, participants were told to name the picture appearing with a colored outline as the head noun, and the color of the outline indicated the preposition (*for* or *near*) that the participant should use to link it with the other picture name.

Other studies have elicited attraction effects using more naturalistic picture description paradigms in which participants generate a message from the pictures, rather than use them to create a form of preamble. For instance, Veenstra, Acheson, and Meyer (2014, Experiment 1) instructed participants to describe pictures of colored shape configurations using the construction *the [colored shape(s)] with/next to the [gray shape(s)] is/are [color]* (e.g. "the star next to the circles is blue"). Paralleling the observations made in preamble experiments, Veenstra, Acheson, and Meyer (2014) found that participants made more errors in the mismatch conditions (SP, PS) than the match conditions (SS, PP) and saw evidence of the markedness effect (this effect was manifested as stronger attraction in the SP condition than the PS condition rather than an absence of attraction in the PS condition). Nozari and Omaki (2022) also observed classic verb attraction effects in a similar picture description paradigm. In their experiment, participants saw slides displaying multiple groups of multicolored animals; each group contained two different animal types, with one or two animals of each type. In each trial, participants were cued to describe the color of an animal or animal pair (the target) to a confederate, whose copy of the slide displayed some of the animals in grey. Since the confederate's slide contained more than one animal of each type in grey, participants referenced the other animal(s) in the target's group in order to disambiguate the target; for instance, rather than simply saying "the snake is green", the participant would need to disambiguate which snake in the scene was the target by using a sentence such as "the snake next to the purple elephants is green". This need for disambiguation provided a natural motivation for participants to produce the desired sentence structure. Nozari and Omaki (2022) observed more errors in the mismatch conditions, with evidence of a markedness effect similar to Veenstra, Acheson, and Meyer (2014).

Veenstra, Acheson, and Meyer (2014) and Nozari and Omaki (2022) thus provide evidence that verb attraction can be observed in more naturalistic production paradigms in addition to the traditional preamble paradigm. To our knowledge, no prior studies have applied such description paradigms to elicit reflexive pronoun agreement. In the present study, we introduced a novel scene-description production paradigm to elicit verb number agreement as well as reflexive number agreement, allowing for a side-by-side comparison. We show in other work that it is possible to elicit verb attraction using this scene-description paradigm in both in-lab and web-based settings (Kandel et al., 2022; Experiment 1 reported in Kandel et al., 2022 is the same as Experiment 1 in the present study).

Timing as an index of attraction

In our study, we not only measured error rates, but we also looked for attraction effects within the production time-course of sentences in the absence of overt errors. We are interested in understanding how dependency formation proceeds in general, yet number attraction studies have typically made inferences about dependency formation based solely on errors that only ever occur on a subset of trials. The production time-course of correct responses may provide clues to what is happening on the majority of trials that are discarded in typical analyses. A handful of previous studies have investigated the relationship between response time and verb number attraction errors, finding that the latency to produce or select verb forms is influenced by the same type of environments that induce agreement errors. Thus, production time-course may provide another index of attraction effects even in responses with no errors.

Haskell and MacDonald (2003) collected both error rate and response time data in their elicitation paradigm. In each trial, participants read a preamble subject phrase (e.g. *the actor in the weekend performance*) and an adjective (e.g. *famous*), which they were instructed to use to form a question asking whether the subject had the adjective property (e.g. "Was the actor in the weekend performance famous?"). It had been previously demonstrated that questions display similar verb attraction patterns to declarative sentences (Vigliocco & Nicol, 1998). In addition to measuring error rate, Haskell and MacDonald (2003) measured speech onset latency; as the questions produced in their experiment began with the inflected verb, they reasoned that this onset latency should reflect the time needed to produce the agreement. They observed longer speech initiation latencies for sentences with correct verb agreement in the same mismatch environments where attraction errors were more common. Haskell and MacDonald (2003) interpret this latency effect as representing the cost of resolving competition between partially activated singular and plural verb forms. Staub (2009) points out, however, that the speech onset latencies in Haskell and MacDonald's (2003) experiments might not solely reflect verb planning time but could also reflect planning of later sentence constituents, such as the subject phrase, which could be slower to plan in the mismatch condition as the speaker must keep track of two distinct number features.

Staub (2009, 2010) measured response time to verb form selection in a speeded forced-choice task. Participants were presented with subject phrases using rapid serial visual presentation, followed by a display with the verb forms *is* and *are* presented to either side of the screen; participants were instructed to select the verb form that would grammatically complete the sentence. Staub (2009, 2010) found longer verb selection response times and reliably more verb agreement errors (with evidence of a markedness effect) for preambles with mismatching head and interfering noun number; latencies were not significantly different for correct and incorrect responses in the mismatch conditions. Veenstra, Acheson, and Meyer, (2014, Experiment 2) and Veenstra, Acheson, Bock, and Meyer (2014, Experiment 2) applied the forced-choice task developed by Staub (2009, 2010) with new stimuli and obtained similar results (though they only analyzed response times for correct verb selections).

Brehm and Bock (2013) applied a similar approach, using a modification of the preamble paradigm in which participants silently read subject phrase preambles presented on a screen and then produced a completion aloud using one of four possible adjectives. Brehm and Bock (2013) measured the latency between the end of the preamble display and the onset of the participant responses. They found evidence of verb attraction (more errors after SP preambles than SS preambles) as well as longer response times for grammatical sentence completions following SP preambles. Veenstra, Acheson, Bock, and Meyer (2014, Experiment 1) observed comparable results to Brehm and Bock (2013) when using the same paradigm with new stimuli. Veenstra, Acheson, Bock, and Meyer (2014) used the same preambles in both their application of Brehm and Bock's (2013) modified preamble paradigm (Experiment 1) and Staub's (2009, 2010) forced-choice paradigm (Experiment 2), finding similar results between the two paradigms.

There is thus evidence that speakers are slower to produce verb agreement in the same contexts where they are more likely to produce errors, suggesting that similar agreement processes occur whether or not an error is ultimately produced. These findings suggest that the production time-course of correct sentences can provide another index of attraction effects in addition to error rates. In the present study, we applied a forced aligner to determine onset and offset times of each word in the responses and looked for utterance-medial delays directly preceding the production of the agreement target (i.e. the verb or the reflexive pronoun). This method allowed us to investigate slowdowns localized directly before the verb or reflexive within complete sentence productions, as compared to Staub (2009, 2010) or Brehm and Bock's (2013) experiments in which participants only selected or produced a sentence completion. By using declarative responses in which the agreement targets appear at the end of the sentence, we also avoided the potential confound that any observed delays reflect the planning of other, later sentence constituents (cf. Haskell & MacDonald, 2003). To our knowledge, we are the first to use production time-course as a measure of attraction in a more naturalistic description elicitation paradigm (in the present study and Kandel et al., 2022) as opposed to one that relies on comprehension of a preamble, and the present study is the first to investigate the production time-course of reflexive pronoun agreement.

Looking for utterance-medial pauses before the verb and reflexive allows us to assess whether pressures leading to attraction are active directly prior to production of the agreement target, which could indicate that the process leading to attraction is implicated in the number agreement computation itself as opposed to initial encoding of the subject number (e.g. Kandel et al., 2022). In a follow-up exploratory analysis, we also look for slowdowns across different parts of the sentence, which additionally allows us to assess when during sentence processing attraction pressures are active, whether pressures are similarly active for each dependency type, and how elicitation paradigm influences planning.

The present study

In the present study, we revisit the finding that reflexive pronouns display robust number attraction effects similar to verbs in production. The observation of verb-like reflexive attraction in production (Bock et al., 2006; Bock et al., 1999) is potentially surprising given that number is part of the pronoun's lexical representation. A pronoun's form heavily influences its interpretation (as opposed to simply marking a dependency), meaning that there is a high potential for misinterpretation by the listener in cases of form errors. The stakes for correct anaphor agreement in production are thus very high. It has been proposed that pronoun form may be informed directly from the message rather than through an inflection process (e.g. Eberhard et al., 2005), thereby reducing the potential for influence from the production mechanism and the opportunity for number agreement to go awry. Furthermore, unlike verb attraction, reflexive attraction effects have not been tested in numerous different in-lab elicitation paradigms, and to our knowledge there is not ample evidence of reflexive pronoun attraction errors occurring regularly outside of a lab setting (cf. Bock & Miller, 1991; Haskell et al., 2010; Pfau, 2009 for verbs). In addition, the preamble paradigm that has been used to elicit such errors in the lab in the past differs in a number of ways from natural production, including providing participants with all the linguistic structure of the sentence except for the final word. The observation of reflexive pronoun number attraction in production thus merits revisiting.

In this study, we apply a novel paradigm to investigate whether previously-observed verb and reflexive attraction patterns hold in a more naturalistic scene-description task (Experiments 1 & 2). We looked for evidence of attraction in both the distribution of errors and in the production time-course of error-free sentences. We also applied a preamble paradigm to elicit the same sentences as the scene-description paradigm (Experiments 3 & 4), again investigating both overt number agreement errors as well as the production time-course of correct sentences. This series of four experiments allowed us to perform side-by-side comparisons of both dependency types as well as both production paradigms.

Data availability

The data, analysis code, model objects, and Supplementary Materials for the present study are available from https://osf.io/ysn5j/.

Experiment 1

The goal of Experiment 1 was to replicate previously-observed verb attraction effects using a more naturalistic production paradigm. We investigated whether subject phrases containing nouns of different number would lead to verb agreement errors and to delays in verb articulation, even when agreement was produced correctly. Experiment 1 is also reported as Experiment 1 in Kandel et al. (2022).

Materials and methods

Participants

The participants were 45 native English speakers (M_{age} = 21.1 years, SD = 4.5 years, 34 F, 11 M) from the University of Maryland community. Participants completed the experiment in exchange for payment or course credit. An additional four omitted participants completed the task: two were omitted for being non-native speakers of English, one was omitted for not following task directions (the participant did not respond in full sentences), and one was omitted because over 1/3 of their trials were omitted (see Experiment 1 *Analysis* for trial omission criteria). For all experiments in this paper, no participant took part in more than one experiment.

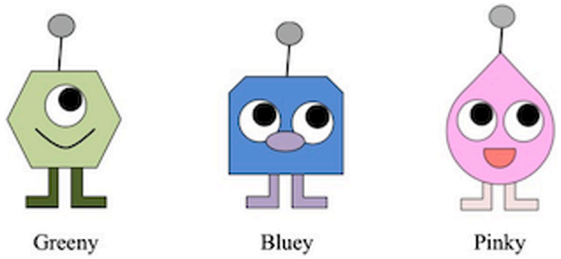

Fig. 1. *Alien types.* The three types of alien used in the scene-description paradigm.

Materials

Experimental materials consisted of short, animated scenes designed to elicit 96 different target sentences (a complete list of target sentences is available in the Supplementary Materials). The scenes depicted cartoon aliens performing a made-up action called *mimming*. The experiment contained three types of aliens (Fig. 1): *greenies* (aliens with a green body), *blueys* (aliens with a blue body), and *pinkies* (aliens with a pink body). Each alien type has an antenna protruding from the center of its head; when an alien mims, its antenna lights up. We decided to use a novel action in our study so that we could flexibly alter the argument structure of the corresponding verb in our target sentences, allowing us to elicit different target sentence frames in the different experiments while holding the verb constant.

The scenes in the experiment showed two groups of aliens separated by a centered vertical line (Fig. 2). Each group was made up of two alien types, with one or two aliens per type (2–4 aliens total). After 1 s of preview time, mimming occurred in one of the two groups on the screen; the antenna(e) of one alien type within the group lit up. The antenna(e) remained lit for 3 s, after which the scene ended. At the onset and offset of each scene, a short reference sound was played to facilitate identification of scene onset and offset times in the recorded data.

The target sentences elicited by the scenes took the form of a complex subject phrase, consisting of two noun phrases linked by a preposition, followed by the verb *mim* conjugated in the present progressive tense (*is mimming* or *are mimming*). The form of the target sentences can be described by the formula: *the* + *N1* + *preposition* + *the* + *N2* + *is/are* + *mimming*. The inclusion of two groups of aliens within each scene provided a pragmatic incentive for participants to use this sentence structure; referencing multiple aliens in their descriptions allowed participants to disambiguate exactly which alien was the subject (as opposed to simply saying, e.g., *the pinky is mimming*).

Six different nouns appeared in the target sentences: *bluey*, *blueys*, *greeny*, *greenies*, *pinky*, and *pinkies*. N1 and N2 were always different alien types. The scenes were designed to elicit four spatial prepositions: *above*, *below*, *to the left of*, and *to the right of*. The prepositions were selected as opposite pairs so that the subject and preposition of the target sentences would not be predictable from the configuration of aliens in the scenes, thereby preventing participants from being able to plan their responses in advance of the mimming action. For instance, if a group of aliens showed a bluey above a greeny, the participant would not be able to predict from this configuration which alien would be the subject and whether they would need to describe the action using *above* or *below*. The inclusion of multiple alien types in the experiment as well as the presence of two groups of aliens in each scene further decreased the predictability of the target sentence, thereby discouraging participants from planning any part of their responses before mimming had occurred.

The number of N1 and N2 was manipulated to create four conditions: SS, SP, PP, and PS, where the first letter stands for the number of N1, and the second letter refers to the number of N2 (Table 3). There were 24 target sentences in each condition. The SS and PP conditions are referred to as the match conditions (N1 and N2 have the same number), and the SP and PS conditions are referred to as the mismatch conditions (N1 and N2 have different number). Experiment 1 contained no filler trials; there was little need to conceal the primary task from participants, and fillers may have made it harder for participants to successfully produce the target structures. Furthermore, Nozari and Omaki (2022) demonstrated that a similar task without filler trials was successful in eliciting verb attraction errors.

The order of the 96 target sentences was pseudorandomized in four lists. The order in each list complied with the following constraints on trials: i) no more than 2 consecutive match or mismatch conditions, ii) no more than 2 consecutive trials of the same condition (SS, SP, PP, PS), iii) no more than 3 consecutive target sentences with the same N1 number (i.e. the same verb agreement; *is mimming* vs. *are mimming*), iv) no more than 2 consecutive sentences with the same preposition, v) no more than 3 consecutive sentences with the same alien type as N1, vi) no more than 3 consecutive sentences with the same alien type as N2, and vii) no more than 2 consecutive sentences involving the same alien pair (blueys and pinkies, greenies and blueys, pinkies and greenies).

Scenes were then created for each of the four lists. Whether the mimming action occurred in the group on the left or right side of the scene was pseudorandomized within each list such that an equal number of trials displayed mimming on the left and on the right and mimming did not occur on the same side on more than 3 consecutive trials. In each scene, the critical group (the one in which mimming occurred) was

Table 3
Experiment 1 response conditions.

Condition	Match Condition	Example Target Sentence
SS	match	the bluey above the greeny is mimming
SP	mismatch	the bluey above the greenies is mimming
PP	match	the blueys above the greenies are mimming
PS	mismatch	the blueys above the greeny are mimming

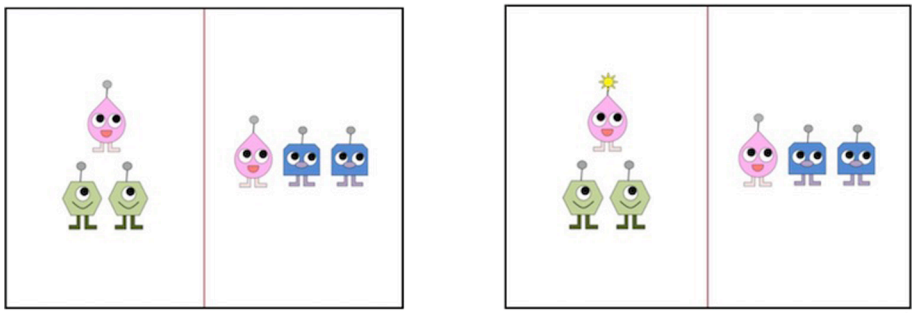

Fig. 2. *Example Experiment 1 scene.* Stills from an Experiment 1 scene before and after the onset of mimming. The target sentence elicited by this scene is *the pinky above the greenies is mimming*.

pseudorandomly paired with a group of aliens of a different configuration (there were 48 possible group configurations, with each alien type and number appearing above, below, to the left of, and to the right of each other alien type and number) such that the same scene configuration did not appear twice in a row. In a given experimental list, an equal number of each possible group configuration appeared on either side of the screen.

Each list was presented with the same 8 practice trials (always presented in the same order), divided into two practice sessions. The practice trials elicited a subset of the experiment target sentences (practice target sentences are identified in the Supplementary Materials). The practice target sentences were chosen such that each preposition type was used once per session and each alien type appeared as both N1 and N2 in each session. The side of the critical group was balanced for left and right within each practice session, and the scene configurations were created using the same criteria as the experimental trial scenes, with the added constraint that the scene configurations used in the practice trials not appear in any of the four experimental lists.

Procedure

For all experiments in this paper, stimuli were presented using PsychoPy v1.85.3 (Peirce & MacAskill, 2018) on a 2013 15″ or 13″ Retina Display MacBook Pro at a natural distance from the participant. The stimuli were presented 720px by 540px. Participant speech was recorded during the experiment in Audacity v2.2.1 at 44100 Hz using a Rode NT1 microphone with a Blue Icicle USB interface.

Participants in Experiment 1 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. At the start of the experiment, participants were introduced to the three types of aliens and to the action mimming. Participants then completed the two practice sessions (4 trials each) before starting the experimental trial block (96 trials). For each scene, participants were asked to describe who is mimming. Since two groups of aliens appeared in each scene, participants were asked to use spatial prepositions to make clear which aliens in the scenes were performing the action.

The first practice session was untimed to give participants the opportunity to familiarize themselves with the paradigm and the format of the target sentences. Participants pressed a button to initiate each trial. After 1 s of preview, the mimming action occurred. Participants gave their descriptions and, when ready, pressed a button to reveal the target sentence on screen; the target sentence appeared below the group of aliens in which the action occurred. Participants pressed again to end the trial.

The second practice session followed the same format as those in the experimental trial block (Fig. 2). Participants pressed a button to initiate each trial. Participants were told that mimming only lasts for 3 s, so they must speak quickly. The limited response time was intended to apply extra pressure to the participant and to decrease potential sentence revision time during the planning process, thereby increasing the likelihood of production errors. Participants were not shown the target sentences for the trials in the second practice session.

Participants were given verbal feedback at the end of each practice trial during the practice sessions. This feedback never referenced number agreement or whether participants produced correct or incorrect agreement in their responses. After completing the practice sessions, the participant proceeded to the experimental trial block. Participants were allowed to take breaks as necessary and were notified when they had completed half of the experimental trials. In Experiment 1 and all following experiments, the experimenter remained in the room for the duration of the experiment.

Analysis

Responses to each trial were transcribed and coded for their inclusion of a number agreement error or other type of error. Agreement errors included both unrevised errors (e.g. “the bluey above the greenies are mimming”) and revised errors (e.g. “the bluey above the greenies are is mimming”, “the bluey above the greenies are mimming is mimming”). Incomplete productions of an agreement error (e.g. “the blueys above the greeny i- are mimming”) were coded as agreement errors with revision.

Responses were omitted from the analysis if the verb form was unidentifiable, if the response did not follow the target sentence formula *the* + *N1* + *preposition* + *the* + *N2* + *is/are* + *mimming* (e.g. “the bluey is mimming above the greenies”), or if the response expressed a meaning that did not match the corresponding scene (i.e. if the participant produced incorrect number marking on one or both of the noun phrases or used the wrong alien name or preposition). Using a non-target preposition with the same meaning as the target (e.g. “under” in place of “below”) was not considered an error. If the participant corrected an error that would result in trial omission in a single revision (e.g. “the greeny- bluey above the greenies is mimming”), the response was not omitted and was instead 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”).

Responses containing no errors (agreement or disfluency) were forced-aligned to their transcriptions, obtaining onset and offset times for each word in the sentences. For all experiments in the present study, responses were forced-aligned using the Montreal Forced Aligner v1.0.0 (McAuliffe et al., 2017). Sentences with a non-zero difference between the offset of N2 and the onset of the verb *is/are* were coded as containing a gap. To assess the reliability of the forced aligner’s identification of speech gaps before the agreement target (*is/are* in Experiments 1 & 3 and *itself/themselves* in Experiments 2 & 4), we randomly selected a subset of 100 responses (25 per experiment; 50 per dependency type) that the forced aligner had identified as containing a gap and 100 responses (25 per experiment; 50 per dependency type) that the forced aligner had identified as not containing a gap. These 200 responses were hand-coded for the presence of a gap (the coder was blind to the forced aligner’s classification on a trial-by-trial basis). After initial classification, the coder inspected the forced-alignments for the responses where the classifications diverged. After reconciling differences, the forced aligner and the hand coder agreed on 91.5 % of responses with a Cohen’s κ of 0.83. Divergences in classification primarily occurred in the reflexive trials (13/17 total disagreements), where the forced aligner and the hand coder disagreed in internally-consistent fashions about verb offset identification. More details about the forced aligner reliability check are available in the Supplementary Materials.

We performed the same set of analyses for each experiment (described below). The data from the experiments were initially analyzed separately. We compare results across experiments in *Experiment comparison analyses* using post-hoc analyses combining the data from all four experiments. The results of our exploratory time-course analysis are presented for all experiments in *Exploratory production time-course analyses*. All statistical analyses in the present paper were performed in R v4.1.0 (R Core Team, 2021), and all statistical models were computed using the package {lme4} v1.1–27.1 (Bates et al., 2015). For Experiment 1 and all following experiments, the overall effects reported for categorical fixed effects variables entered into models with an interaction (e.g. match & N1 number) were derived from models with effects coding, allowing comparison to the grand mean (analogous to main effects). The reported values for the interaction effects were derived from these same models. Reported values for between-condition comparisons (SS vs. SP, PP vs. PS) were obtained from models utilizing dummy-coded categorical variables. Match effect plots for the error and gap analyses are available in the Supplementary Materials.

All analyses were also computed in a Bayesian framework using the same model structures. Descriptions of the Bayesian analyses and results are available in the Supplementary Materials. For the majority of the

analyses, the results patterns were the same for both the frequentist and Bayesian models (applying a 95 % confidence threshold). We identify in our results any instances when the two sets of results differed. In such instances, we report the posterior median of the estimated parameter coefficient values and the credible interval (CrI) for the coefficient (highest density intervals were used as CrIs). A CrI indicates the range of values in which the regression coefficient is likely to sit; for hypothesis testing, if zero is not in the 95 % CrI, we can be 95 % confident that the parameter had a non-zero effect.

Error distribution analysis. The error distribution analyses in our experiments investigated the likelihood of producing an error in the different target sentence conditions using logistic mixed effects analyses. This analysis assesses the presence of a canonical attraction error effect. We constructed generalized linear mixed models with a binomial distribution and logit link. Unless otherwise specified, for all error distribution analyses in this paper, the models contained fixed effects of match (SS/PP vs. SP/PS), N1 number (i.e. whether the target verb or reflexive was singular or plural), and their interaction, with random intercepts for target sentence and participant as well as a random slope of match by participant. For all experiments, we analyzed both all non-omitted responses as well as a restricted dataset without disfluency errors. When the pattern of findings was the same for both analyses, we report only the results from the analysis of the non-restricted dataset, which had a higher number of observations.

Gap Analyses. The gap distribution analyses in our experiments used logistic mixed effects analyses to investigate the likelihood of producing a gap before onset of the agreement target (the verb in Experiments 1 & 3 and the reflexive in Experiments 2 & 4) in the different conditions. We constructed generalized linear mixed models with a binomial distribution and logit link. For all experiments, the models in the gap distribution analyses always contained the same effects structure as their corresponding error distribution analyses. In addition to analyzing gap distribution, we also investigated the duration of the non-zero gaps in an exploratory post-hoc analysis. For all gap duration analyses in this paper, we constructed generalized linear mixed effects models with a gamma distribution and log link. The models had the same fixed effects as the error and gap distribution analyses with a random intercept for participants as well as a random slope for match by participant (adding a random intercept for the target sentence resulted in convergence errors). Analyzing the likelihood of gaps (and their durations when they occur) can be used to identify whether there are localized slowdowns in speech caused by processing agreement in the same environments where errors are more likely, even when the correct form is reached.

Results

Error distribution analysis

Out of the 4320 total responses, 304 responses were omitted from the analysis. Of the remaining 4016 responses, 249 responses contained disfluency errors. We obtained a total of 489 number agreement errors; 49 of these errors occurred in sentences that also contained a disfluency error. The distribution and percentage of agreement errors in each condition is presented in Table 4; participant error rates by condition (including responses with disfluency errors) are presented in Fig. 3.

The overall effect of match was significant, with errors more likely in the mismatch conditions ($\beta = 1.732$, SE = 0.186, $z = 9.324$, $p < 0.0001$). The estimated error probability in the match conditions was 0.7 % (SE = 0.3 %) compared to 18.5 % (SE = 2.9 %) in the mismatch conditions. There was also an overall effect of N1 number such that errors were significantly more likely in conditions with plural subjects ($\beta = 0.467$, SE = 0.141, $z = 3.306$, $p < 0.001$). The estimated error probability for the singular subject head conditions was 2.4 % (SE = 0.7 %) compared to 5.9 % (SE = 1.2 %) for the plural head conditions. Our analysis revealed a significant interaction between match and N1 number ($\beta = -0.589$, SE = 0.141, $z = -4.170$, $p < 0.0001$). The SS vs. SP ($\beta = 4.641$, SE = 0.572, $z = 8.109$, $p < 0.0001$) and PP vs. PS ($\beta = 2.287$, SE = 0.329, $z = 6.961$, $p < 0.0001$) contrasts were both significant, with errors were more likely in the mismatch conditions, though the SS vs. SP difference was greater than the PP vs. PS difference. The estimated error probabilities were 0.2 % (SE = 0.1 %) for the SS condition, 20.4 % (SE = 3.3 %) for the SP condition, 2.0 % (SE = 0.6 %) for the PP condition, and 16.7 % (SE = 2.9 %) for the PS condition.

Table 4
Experiment 1 agreement error and response distributions (values omitting responses with disfluency errors are given in parentheses).

Condition	Match Condition	Error Count	Response Count	Error Rate
SS	match	4 (4)	1046 (992)	0.4 (0.4) %
SP	mismatch	246 (224)	990 (917)	24.8 (24.4) %
PP	match	30 (25)	994 (941)	3.0 (2.7) %
PS	mismatch	209 (188)	986 (917)	21.2 (20.5) %

Gap analysis

One participant was omitted from the gap analysis because their responses could not be forced-aligned. For the 3267 remaining responses containing no errors (number agreement or disfluency), we analyzed the likelihood of pausing immediately prior to verb articulation in the different conditions. We obtained a total of 326 gaps, defined as a non-zero difference between the offset of N2 and the onset of the verb. The distribution and percentage of gaps in each condition are presented in Table 5 (a plot of participant gap rates by condition is available in the Supplementary Materials).

There was an overall effect of match such that gaps were significantly more likely in the mismatch conditions ($\beta = 0.838$, SE = 0.117, $z = 7.130$, $p < 0.0001$). The estimated gap probability in the match conditions was 2.4 % (SE = 0.7 %) compared to 11.5 % (SE = 2.2 %) in the mismatch conditions. There was also a significant overall effect of N1 number such that gaps were more likely in conditions with plural subjects ($\beta = 0.186$, SE = 0.067, $z = 2.761$, $p = 0.006$). The estimated gap probability in the singular subject head conditions was 4.0 % (SE = 0.9 %) compared to 5.7 % (SE = 1.3 %) in the plural head conditions. The interaction between match and subject NP number was not significant ($\beta = -0.020$, SE = 0.067, $z = -0.292$, $p = 0.771$).

The overall effect of match on gap duration was trending but not significant ($\beta = 0.155$, SE = 0.082, $t = 1.897$, $p = 0.058$). The estimated mismatch gap duration was 98 ms (SE = 9 ms), and the estimated match gap duration was 73 ms (SE = 13 ms). The overall effect of N1 number was significant ($\beta = 0.270$, SE = 0.056, $t = 4.846$, $p < 0.0001$), with longer gaps in sentences with plural subjects. The estimated plural subject gap duration was 110 ms (SE = 12 ms), and the estimated singular subject gap duration was 69 ms (SE = 8 ms). The interaction between match and subject NP number was not significant ($\beta = -0.095$, SE = 0.056, $t = -1.696$, $p = 0.090$).

Experiment 1 discussion

Experiment 1 replicated previously-observed verb attraction effects: we observed more number agreement errors in mismatch sentences than match sentences. While we saw evidence of a markedness effect, with more errors in the SP condition than the PS condition, our paradigm induced more PS errors than are typically found for sentences with simple count nouns (Table 2). We will address this finding further in the *General discussion* (see also Kandel et al., 2022 for discussion).

The distribution of pre-VP gaps in the sentences containing no disfluencies roughly paralleled the distribution of number errors: gaps were more common in the mismatch conditions than in the match conditions, although there was no evidence of a markedness effect. The exploratory gap duration analysis did not reveal a reliable difference in gap duration between the match and mismatch conditions. We thus see an effect of attraction-inducing environments even when participants are not

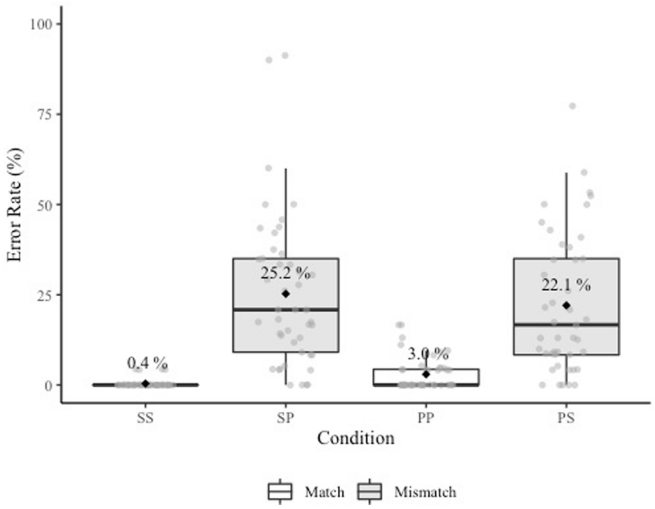

Fig. 3. ***Experiment 1 (scene-description, verbs) participant agreement error rates.*** Boxplot of participant agreement error rates by condition (including responses with disfluency errors). The round points represent participant rates. Mean error rates are labeled and represented by diamonds.

Table 5
Experiment 1 gap distributions in responses without errors or disfluencies.

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

overtly producing errors, suggesting that the same underlying processes that cause attraction are at play even when a speaker produces the correct verb form. The fact that slowdowns immediately prior to the agreement target pattern similarly to agreement errors indicates that production time-course can be used as an index of attraction effects even when no error is produced. Our results coincide with previous time-course analyses (Brehm & Bock, 2013; Staub 2009, 2010; Veenstra, Acheson, Bock, & Meyer, 2014; Veenstra, Acheson, & Meyer, 2014), though the present analysis goes beyond prior analyses by allowing for detection of utterance-medial effects in full sentences elicited by a more naturalistic paradigm. As discussed by Kandel et al. (2022), the presence of a pre-VP gap effect suggests that attraction pressures are active during the verb number agreement computation as opposed to during earlier encoding of subject number.

Experiment 2

The results from Experiment 1 show that our scene-description paradigm elicits verb attraction errors when expected and that the time-course analyses can detect an influence of attraction-inducing environments even when no error is produced. The goal of Experiment 2 was to apply this same paradigm to investigate whether the attraction effect on reflexive pronoun production observed by Bock et al. (1999) would replicate in a more naturalistic task. In addition to eliciting sentences with reflexive pronouns, we also elicited sentences containing simple object pronouns in order to investigate whether the level of attraction susceptibility observed for reflexives extends to other anaphors. As in Experiment 1, we analyzed both error likelihood and the time-course of correct productions, probing whether participants were more likely to pause before producing the anaphor in mismatch environments even when no number agreement error was produced.

Materials and methods

Participants

The participants were 43 native English speakers (M_{age} = 20.1 years, SD = 1.6 years, 29 F, 14 M) from the University of Maryland community. Participants completed the experiment in exchange for payment or course credit. An additional six omitted participants were run in the task: two were omitted for being non-native speakers of English, three were omitted because over 1/3 of their trials were omitted (see Experiment 2 *Analysis* for trial omission criteria), and one participant was omitted for not producing consistent anaphor number agreement.[2]

Materials

Experimental materials consisted of short, animated scenes designed to elicit 192 different target sentences containing reflexive and simple object pronouns (see Supplementary Materials for a complete list). As in Experiment 1, the scenes depicted greenies, blueys, and pinkies performing a made-up action called *mimming*. In Experiment 2, mimming was introduced as an action that the aliens could perform to themselves or others. If an alien mimmed itself, it pulsed and then its own antenna lit up; if an alien mimmed another alien, it pulsed and then the other alien's antenna lit up.

As in Experiment 1, each scene showed two groups of aliens separated by a centered vertical line, and the mimming action occurred in one of the two groups on screen (Fig. 4). After 1 s of preview time, the alien(s) performing the action pulsed for 1 s, after which the same alien (s)' antenna(e) lit up or another alien(s)' antenna(e) lit up. The antenna (e) remained lit for 3 s, after which the scene ended. At the onset and offset of each scene, a short reference sound was played to enable identification of trial onset and offset times in the recorded data.

The target sentences elicited by the scenes took the form of a complex subject phrase, consisting of two noun phrases linked by a preposition, followed by the verb *mimmed* plus either a reflexive pronoun (*itself, themselves*) or an object pronoun (*it, them*). The form of target sentences can be described by the formula: *the* + *N1* + *preposition* + *the* + *N2* + *mimmed* + *anaphor*. We elicited the past tense of the verb *mim* in order to parallel the preamble structure used for reflexive elicitation in Bock

[2] Out of 191 complete responses (95 of which had plural target anaphors), the participant produced only 5 plural anaphor forms. The 5 plural forms were all reflexive; the participant produced no plural object pronouns. All 5 elicited plural forms occurred in the first half of the experiment; 4 of the forms were produced within the first 20 trials (out of 192 total trials).

et al. (1999). The use of a past tense verb avoids the need to compute overt verb agreement with the subject.

The target sentences used the same set of nouns and prepositions as Experiment 1. We manipulated the number of N1 and N2 as well as whether the sentence contained a reflexive pronoun ("reflexive trials") or object pronoun ("pronoun trials") (Table 6). There were 24 target sentences in each condition. The pronoun trials provided a point of comparison to see whether the patterns observed for reflexives are specific to reflexive pronouns or whether they apply more generally to different anaphor types. The mixture of different trial types further served to focus participants' attention on describing the mimming action rather than on the forms of the anaphors in their responses, thus increasing the likelihood of form errors.

The order of the 192 target sentences was pseudorandomized in four lists. The order in each list complied with the same constraints as the pseudorandomization in Experiment 1 plus the additional constraint that no more than 3 of the same trial type (reflexive or pronoun) appear consecutively. Given the large number of trials, the lists were divided into 4 blocks of 48 trials, and participants were notified after the completion of each block. Within each block, there were the same number of reflexive and pronoun trials (24 each). Scenes were created for each of the four lists following the same method and constraints as Experiment 1.

Each list was presented with the same 8 practice trials (always presented in the same order), divided into two practice sessions. The practice trials elicited a subset of the experiment target sentences (practice target sentences are indicated in the Supplementary Materials). The practice trials were selected and created following the same constraints as Experiment 1, with the added property that each practice session contained 2 reflexive and 2 pronoun trials.

Procedure

Participants were distributed across the four presentation lists: 12 participants saw list 1, 10 participants saw list 2, 11 participants saw list 3, and 10 participants saw list 4. At the start of the experiment, participants were introduced to the three types of aliens and to the action mimming. Participants then completed two practice sessions (4 trials each) before starting the experimental trials (182 trials, divided into 4 blocks of 48 trials). For each scene, participants were asked to describe which alien mimmed whom, using spatial prepositions to make clear which aliens in the scenes were performing the action.

The practice sessions followed the same format as in Experiment 1. The first practice session was untimed, and participants saw the intended target sentences after each trial. The second practice session used the same timing as the experimental trials (Fig. 4). Participants were told that mimming only lasts for 3 s, so they must speak quickly. During both practice sessions, participants were given verbal feedback at the end of each trial. This feedback never referenced number agreement or whether participants produced correct or incorrect agreement in their responses. After completing the practice sessions, participants

Table 6
Experiment 2 response conditions.

Trial Type	Condition	Match Condition	Example Target Sentence
Reflexive	SS	match	the bluey above the greeny mimmed itself
	SP	mismatch	the bluey above the greenies mimmed itself
	PP	match	the blueys above the greenies mimmed themselves
	PS	mismatch	the blueys above the greeny mimmed themselves
Pronoun	SS	match	the bluey above the greeny mimmed it
	SP	mismatch	the bluey above the greenies mimmed them
	PP	match	the blueys above the greenies mimmed them
	PS	mismatch	the blueys above the greeny mimmed it

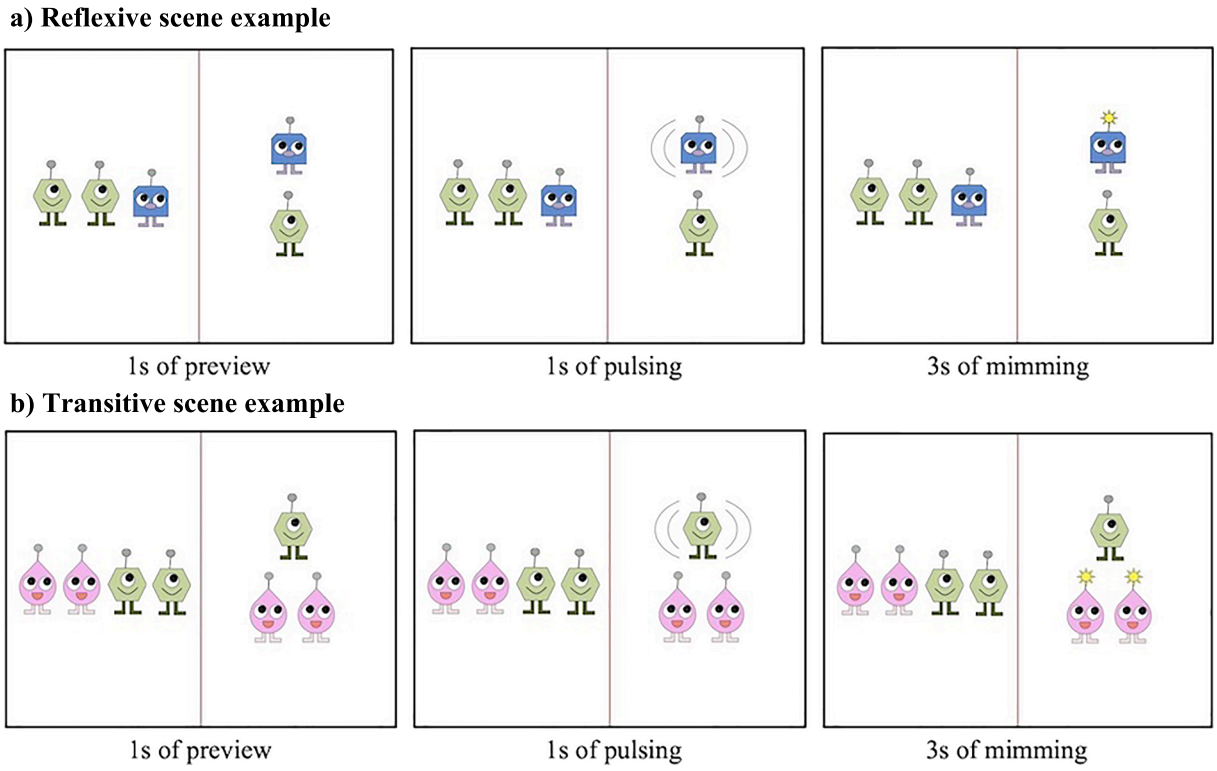

Fig. 4. ***Example Experiment 2 scenes.*** a) Scene eliciting the target sentence *the bluey above the greeny mimmed itself*. Grey lines represent the pulsing action. b) Scene eliciting the target sentence *the greeny above the pinkies mimmed them.* Grey lines represent the pulsing action.

proceeded to the experimental trials. Participants were allowed to take breaks as necessary and were notified after completing each block.

Analysis

Responses to each trial were transcribed and coded for their inclusion of a number agreement error or other type of error. Agreement errors included both unrevised errors (e.g. "the bluey above the greenies mimmed themselves") and revised errors (e.g. "the bluey above the greenies mimmed themselves itself"). Incomplete productions of an agreement error (e.g. "the bluey above the greenies mimmed themselitself") were considered agreement errors with revision. Note that incomplete productions were coded as errors even when the form of the incomplete anaphor was not clearly identifiable as the same type of anaphor (reflexive vs. object pronoun) as the target (e.g. "the bluey above the greenies mimmed the- itself"). In such cases, it is possible that the participant was revising an error of anaphor type (e.g. correcting "them" to "itself") as opposed to number (e.g. correcting "themselves" to "itself"); we decided nevertheless to include them in the analysis as number agreement errors so as not to accidentally exclude potential agreement errors. There were 16 such instances (1 reflexive trial, 15 pronoun trials) amongst the non-omitted responses in the dataset (see below for omission criteria).

There were also several responses in which an object pronoun was produced directly before a second anaphor (e.g. "the blueys above the greeny mimmed it themselves", "the blueys before the greeny mimmed them it"). It is unclear in these responses whether the first anaphor is in fact a complete production of an object pronoun or an incomplete production of a reflexive pronoun. We decided to code this type of response as containing an agreement error when the first and second anaphor had different number, although these responses may actually be instances of anaphor type revision. There were 29 such instances in the total error counts (22 reflexive trials, 7 pronoun trials).

Responses were omitted from the analysis if the anaphor form or NP number marking was unidentifiable, if the response did not follow the target sentence formula *the + N1 + preposition + the + N2 + verb + anaphor*, if the response used a non-standard anaphor type (e.g. itselves, himselves, themself, themselfs, himselfs, imselves, em) (following Bock et al., 1999), or if the response expressed a meaning that did not match the scene (i.e. if the participant produced incorrect number marking on one or both of the noun phrases, used the wrong alien name, preposition, or anaphor type, or erroneously described the action as reflexive or transitive when it was actually the other action type). Using a non-target preposition, anaphor, or determiner with the same meaning as the target (e.g. "under" in place of "below"; "him"/"himself" in place of "it"/"itself"; "a" in place of "the"), using an alternative pronunciation of *mimmed* (e.g. "mimed" or "meemed"), and using a different verb tense from the target *mimmed* (e.g. "the bluey above the greenies mims itself") were not considered errors. As in Experiment 1, if the participant corrected an error that would result in trial omission in a single revision, the response was not omitted and was instead coded as containing a disfluency error. Disfluency errors also included omitting a determiner, repeating a word or the beginning of a word, false starts to a word, revising a word (with the exclusion of anaphor revisions coded as number agreement errors), or saying the color of an alien instead of its name (e.g. "the blues").

Responses containing no errors (number agreement or disfluency) were forced-aligned to their transcriptions. Sentences with a non-zero difference between the offset of the verb and the onset of the anaphor were coded as containing a gap.

The responses to the reflexive and pronoun trials were analyzed separately.

Results: Reflexive trials

Error distribution analysis

Out of the 4128 total reflexive responses, 344 responses were omitted from the analysis. Of the remaining 3784 responses, 389 responses contained disfluency errors. We obtained a total of 47 number agreement errors (including the 23 trials in which it was ambiguous whether the error was an initial pronoun type error); 7 of these errors occurred in sentences that also contained a disfluency error. The distribution and percentage of agreement errors in each condition is presented in Table 7; participant error rates by condition (including responses with disfluency errors) are presented in Fig. 5.

The overall effect of match was not significant ($\beta = 0.437$, SE = 0.388, $z = 1.126$, $p = 0.260$). The estimated error probabilities were 0.3 % (SE = 0.2 %) in the match conditions and 0.7 % (SE = 0.3 %) in the mismatch conditions. There was a significant overall effect of N1 number such that errors were more likely in conditions with plural subjects ($\beta = 0.950$, SE = 0.236, $z = 4.024$, $p = < 0.0001$). The estimated error probability for the singular subject head conditions was 0.2 % (SE = 0.1) compared to 1.1 % (SE = 0.4 %) in the plural head conditions. The interaction between match and N1 number was not significant ($\beta = 0.121$, SE = 0.236, $z = 0.511$, $p = 0.609$).

Gap analysis

For the 3355 responses containing no errors (number agreement or disfluency), we analyzed the likelihood of pausing immediately prior to pronoun articulation. We obtained a total of 377 gaps, defined as a non-zero difference between the offset of the verb and the onset of the reflexive pronoun. The distribution and percentage of gaps in each condition are presented in Table 8 (a plot of participant gap rates by condition is available in the Supplementary Materials).

The overall effect of match was not significant ($\beta = 0.014$, SE = 0.089, $z = 0.156$, $p = 0.876$). The estimated gap probabilities were 5.3 % (SE = 1.3 %) in the match conditions and 5.5 % (SE = 1.3 %) in the mismatch conditions. There was a significant overall effect of N1 number such that gaps were more likely in conditions with plural subjects ($\beta = 0.545$, SE = 0.065, $z = 8.387$, $p < 0.0001$). The estimated gap probability for the singular subject head conditions was 3.2 % (SE = 0.8 %) compared to 9.0 % (SE = 2.0 %) in the plural head conditions. There was a significant interaction between match and N1 number ($\beta = 0.148$, SE = 0.065, $z = 2.279$, $p = 0.023$), though neither the SS vs. SP difference ($\beta = -0.268$, SE = 0.251, $z = -1.070$, $p = 0.285$) nor the PP vs. PS difference ($\beta = 0.324$, SE = 0.184, $z = 1.757$, $p = 0.079$) was significant. The estimated gap probabilities were 3.7 % (SE = 1.0 %) for the SS condition, 2.8 % (SE = 0.8 %) for the SP condition, 7.8 % (SE = 1.9 %) for the PP condition, and 10.5 % (SE = 2.4 %) for the PS condition.

Our post-hoc gap duration analysis found no significant overall effect of match ($\beta = 0.101$, SE = 0.101, $t = 0.993$, $p = 0.321$). The model estimated mismatch gap duration was 72 ms (SE = 12 ms), and the estimated match gap duration was 59 ms (SE = 8 ms). The overall effect of N1 number was significant ($\beta = -0.187$, SE = 0.049, $t = -3.803$, $p < 0.001$), with longer gaps in sentences with singular subjects. The estimated plural subject gap duration was 58 ms (SE = 7 ms), and the estimated singular subject gap duration was 84 ms (SE = 11 ms). The interaction between match and subject NP number was not significant ($\beta = -0.010$, SE = 0.049, $t = -0.210$, $p = 0.834$).

Results: Pronoun trials

Out of the 4128 total pronoun responses, 349 responses were omitted

Table 7
Experiment 2 reflexive trial agreement error and response distributions (values omitting responses with disfluency errors given in parentheses).

Condition	Match Condition	Error Count	Response Count	Error Rate
SS	match	2 (1)	973 (881)	0.2 (0.1) %
SP	mismatch	4 (3)	940 (832)	0.4 (0.4) %
PP	match	10 (9)	943 (853)	1.1 (1.1) %
PS	mismatch	31 (27)	928 (829)	3.3 (3.3) %

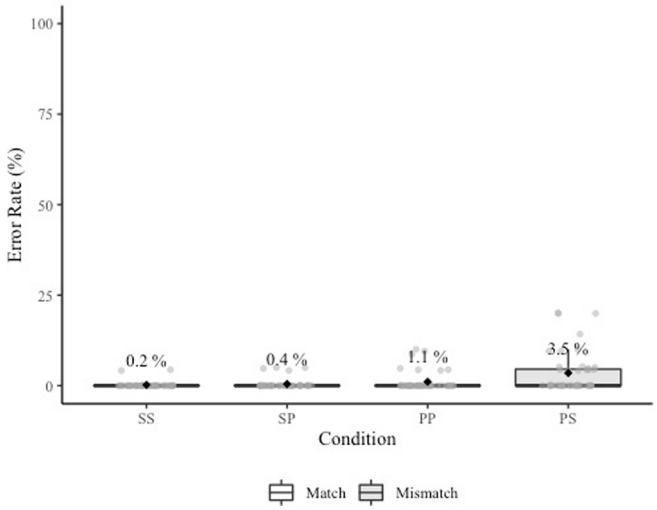

Fig. 5. ***Experiment 2 (scene-description, reflexive trials) participant agreement error rates.*** Boxplot of participant agreement error rates by condition (including responses with disfluency errors). The round points represent participant rates. Mean error rates are labeled and represented by diamonds.

Table 8
Experiment 2 reflexive trial gap distributions in responses without errors or disfluencies.

Condition	Match Condition	Gap Count	Response Count	Gap Rate
SS	match	70	880	8.0 %
SP	mismatch	53	829	6.4 %
PP	match	118	844	14.0 %
PS	mismatch	136	802	17.0 %

from the analysis. Of the remaining 3779 responses, 423 responses contained disfluency errors. We obtained a total of 110 number agreement errors (including the 22 trials in which it was ambiguous whether the error was an initial pronoun type error); 23 of these errors occurred in sentences that also contained a disfluency error. The distribution and percentage of agreement errors in each condition is presented in Table 9; participant error rates by condition (including responses with disfluency errors) are presented in Fig. 6. Note that for the pronoun trials, N2 is the antecedent of the pronoun and N1 is the potential distractor. Thus, in this section we compare conditions with the same N2 but different match (i.e. SS vs. PS and PP vs. SP).

We used a model structure containing fixed effects of match, N2 number (i.e. antecedent number), and their interaction, with random intercepts for target sentence and participant as well as a random slope of match by participant. The overall effect of match was significant such that errors were significantly more likely in the mismatch condition ($\beta = 0.890$, SE = 0.249, $z = 3.579$, $p < 0.001$). The estimated error probability for the match conditions was 0.5 % (SE = 0.2 %) compared to 2.8 % (SE = 0.6 %) in the mismatch conditions. There was also a significant overall effect of N2 number such that errors were more likely in conditions with plural antecedents ($\beta = 0.616$, SE = 0.155, $z = 3.972$, $p < 0.0001$). The estimated error probability for the conditions with singular antecedents was 0.6 % (SE = 0.2 %) compared to 2.2 % (SE = 0.6 %) for conditions with plural antecedents. The interaction between match and N2 number was not significant ($\beta = -0.280$, SE = 0.155, $z = -1.806$, $p = 0.071$).

Table 9
Experiment 2 pronoun trial agreement error and response distributions (values omitting responses with disfluency errors given in parentheses).

Condition	Match Condition	Error Count	Response Count	Error Rate
SS	match	4 (4)	962 (842)	0.4 (0.5) %
PS	mismatch	29 (24)	940 (827)	3.1 (2.9) %
PP	match	22 (17)	928 (829)	2.4 (2.1) %
SP	mismatch	55 (42)	949 (858)	5.8 (4.9) %

Due to the forced aligner's difficulty distinguishing the word "it" from the reference sound at trial offset, we are unable to report reliable timing analyses for the pronoun trials.

Experiment 2 discussion

Number errors were much less common for reflexives in Experiment 2 than they were for verbs in Experiment 1. In our analysis, we were as generous as possible when determining what constituted an agreement error (counting even ambiguous cases where the participant may have been correcting the anaphor type from an object pronoun to a reflexive pronoun rather than correcting a number agreement error), yet the data still only contained 47 reflexive number errors. By contrast, Experiment 1 elicited 488 verb number errors. The error analysis did not reflect a significant attraction effect, and the distribution of reflexive errors clearly differs from that of verb errors in Experiment 1 and of reflexive errors in Bock et al. (1999) (Table 2). We thus do not see a robust attraction effect for reflexive pronouns in the standard error measure. While it is possible that there exists a smaller reflexive attraction effect that we are underpowered to detect in the current study, it is evident that reflexives in our scene-description paradigm do not behave similarly to verbs (this conclusion is supported by the post-hoc analysis in *Experiment comparison analyses*), contra the findings of Bock et al. (1999).

As we saw in Experiment 1, latencies in sentence production can reflect the influence of attraction-inducing environments when no error is produced. Thus, even though there were fewer number agreement errors elicited in Experiment 2, we might still have expected to find evidence of attraction effects in the production time-course of the sentences produced. Nevertheless, we did not observe an effect of match environment on the likelihood or duration of pre-reflexive gaps, further reinforcing the distinction between reflexives and verbs in our paradigm. The lack of a slowdown effect directly prior to production of the reflexive may indicate that there are no attraction pressures active at that point of sentence processing.

It is important to note that the trial structures in Experiments 1 and 2 result in participants receiving the information required to plan their responses at different times in the two experiments. In Experiment 1, all of the information required to describe the scene became available at the same time: when an alien started mimming, its antenna lit up, revealing the subject of the target response at the onset of the mimming event. In Experiment 2, there was a 1 s delay between when the agent and the

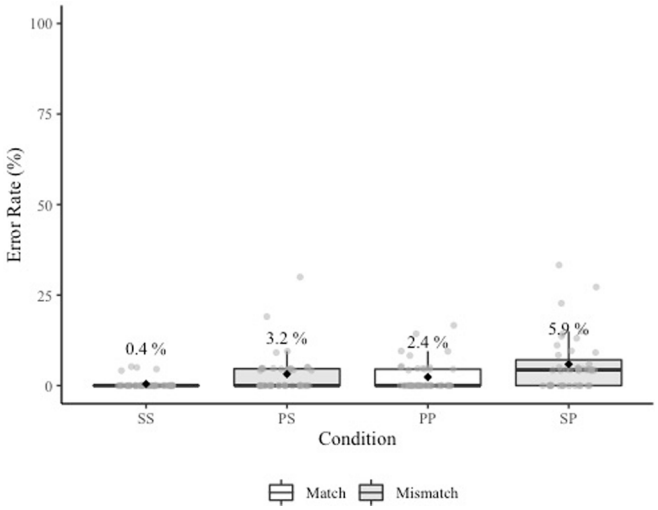

Fig. 6. ***Experiment 2 (scene-description, pronoun trials) participant agreement error rates.*** Boxplot of participant agreement error rates by condition (including responses with disfluency errors). The round points represent participant rates. Mean error rates are labeled and represented by diamonds.

patient of the mimming event were revealed: the mimming alien began pulsing at the onset of the mimming event, identifying the subject of the target response, and then an alien's antenna lit up 1 s afterwards, revealing whether the sentence requires a reflexive or object pronoun. Thus, in Experiment 2, but not Experiment 1, it was possible to begin planning a response before it was known how it should end. However, we believe it is unlikely that the contrast in attraction susceptibility between verbs and reflexive pronouns observed in our study can be attributed to the timing differences in our two experiments. In both Experiment 1 and Experiment 2, the information required to construct the subject phrase of the response was identifiable at the same time. As soon as the subject alien was revealed (either by lighting up its antenna in Experiment 1 or pulsing in Experiment 2), the prepositional phrase modifier of the response could also be identified based on the visual information in the scene. This means that the head NP of the subject phrase had a similar temporal relation to the modifier in Experiment 2 and Experiment 1. The lack of disruption from the NP in the prepositional phrase modifier in Experiment 2 thus cannot be explained by a temporal advantage for the subject head.

Furthermore, while participants may have been able to get a 1 s planning head start in Experiment 2, the maximal sentence structure that can be planned prior to knowing who is being mimmed is the subject phrase plus the past-tense verb. This is the same amount of structure provided to participants in Bock et al.'s (1999) preamble paradigm. Given that participants in Bock et al.'s (1999) study still produced reflexive pronoun errors when given this structure in advance, it is unlikely that having already planned this structure before planning the anaphor in our experiment should have limited the number of reflexive pronoun errors. In fact, having already planned the subject phrase prior to seeing the disambiguating antenna action may even increase the likelihood of errors, as it guarantees that N2 is in a position to cause interference at the point when participants plan the reflexive pronoun. If participants plan the anaphor before the subject phrase structure is fully formed, on the other hand, interference from N2 may be minimized because it has not yet had a chance to influence the representation of subject number (under a representational account of attraction) or because it is not yet available to compete with N1 during cue-based retrieval (under a retrieval account). In addition, we find evidence of attraction for the object pronoun trials in Experiment 2, showing that it is still possible to get attraction effects with the experiment's trial timing. Thus, the trial timing of Experiment 2 is unlikely to have reduced the number of reflexive pronoun attraction errors and could possibly even have inflated the number.

While we did not find a strong attraction effect for reflexive pronouns, we did observe number error distributions in the pronoun trials in line with attraction effects: errors were significantly more likely in the mismatch conditions than the match conditions. This effect was small, particularly when compared to verb attraction effects, though we have since confirmed in follow-up experiments that the effect is reliable (Wyatt et al., 2021). Number attraction errors have previously been observed for unbound tag pronouns, whose features can be influenced by an interfering noun intervening between the tag pronoun and its antecedent, the subject of the preceding clause (e.g. Bock et al., 2006; Bock et al., 2004; Bock et al., 1999). To our knowledge, however, such attraction effects have not been previously demonstrated for simple object pronouns. The number of the object pronouns in our sentences is influenced by the number feature of the head subject noun phrase (N1). Note that non-intervening attractors have previously been found to influence verb number agreement (e.g. Bock & Miller, 1991; Franck et al., 2006; Staub, 2009, 2010 for production; e.g. Wagers et al., 2009 for comprehension). Interestingly, errors were more common in the SP condition than the PS condition — the opposite pattern of what the markedness effect would predict. The markedness effect predicts that plural distractors (in this case, plural N1s) should be more disruptive than singular distractors (singular N1s), yet we see the greatest number of object pronoun number errors when the non-intervening distractor is singular (the SP condition). This apparent reversal may be due to frequency differences between *it* and *them* resulting in a bias towards producing *it* (e.g. there are 18896.31 occurrences per million for *it* vs. 1778.82 per million for *them* in the SUBTLEX-US corpus; Brysbaert & New, 2009). Indeed, we observed an effect of N2 number consistent with participants producing *it* erroneously more frequently than producing *them* inappropriately. Nevertheless, the attraction pattern is visible both in the conditions where the target pronoun is *it* (SS vs. PS) and in the conditions where the target pronoun is *them* (PP vs. SP), suggesting that the attraction effect is not reducible to an *it* bias.

The observed object pronoun attraction effect is striking in that it leads participants to produce sentences that resemble violations of Binding Principle B (Chomsky, 1981), as the pronoun agrees with the features of the local subject. We have found through follow-up investigations that these errors do not arise as a result of requiring participants to produce the object pronouns within the same experiment in which they produce reflexive pronouns, due to the generous method of coding agreement errors used here, or because participants produce the same sentence frame containing a pronoun in every trial of the experiment (Wyatt et al., 2021). The fact that we observe an influence of mismatch environments on number error likelihood for object pronouns shows that lack of a reflexive attraction effect that reaches the

significance threshold is unlikely to reflect a limitation of our elicitation paradigm and could suggest that reflexive and object pronoun dependencies are computed differently. In a post-hoc analysis of the error data (see Supplementary Materials), we tested for an interaction between match condition and trial type (reflexive or pronoun), but it was not reliable (though there was a significant overall effect of trial type such that errors were more likely in pronoun trials than reflexive trials). The present experiment was not designed to detect a match effect difference between the two trial types, however, and given the small sizes of the reflexive and pronoun match effects, we would likely need more power to reliably test for such an interaction. In addition, the reflexive and pronoun trials had different structures (the antecedents and distractors were in different syntactic positions), making it difficult to compare reflexive and pronoun attraction directly (indeed, there is evidence that attraction from intervening and non-intervening distractors may arise via different mechanisms; Staub, 2010). Future confirmatory research is thus necessary to assess the extent to which attraction susceptibility differs between the two anaphor types. Nevertheless, even if reflexives show attraction effects similar to object pronouns, this attraction effect is clearly on a different scale from the robust verb attraction elicited in Experiment 1.

In summary, our scene-description paradigm produces a contrast between verbs and reflexive pronouns. In Experiment 1, we saw robust verb number agreement error rates, with errors significantly more likely in mismatch environments, whereas in Experiment 2 we observed very few reflexive number agreement errors, with little difference between the match and mismatch conditions. The time-course analyses parallel our behavioral findings: in Experiment 1, we found that participants were more likely to pause before producing the verb in the same environments where they are more likely to make agreement errors, but participants in Experiment 2 were not any slower to produce reflexive pronouns in the mismatch environments. The contrast we observe between verbs and reflexive pronouns in our novel paradigm reflects a departure from previous results (e.g. Bock et al., 2006; Bock et al., 1999). To confirm that this departure was due to paradigm differences rather than differences related to our stimuli (while the task was more natural, the target sentences and characters were less so, and the responses were rather repetitive), we used the stimuli from Experiments 1 and 2 to create preambles for preamble elicitation experiments.

Experiment 3

The goal of Experiment 3 was to determine whether we obtain similar results to Experiment 1 when our target sentences are elicited in a preamble paradigm. Again, we investigated both error likelihood and production time-course.

Materials and methods

Participants

The participants were 25 native English speakers (M_{age} = 20.8 years, SD = 1.4 years, 18 F, 7 M) from the University of Maryland community. Participants completed the experiment in exchange for payment or course credit. An additional three omitted participants were run in the task: two participants were omitted for having over 1/3 of their trials omitted (see Experiment 3 *Analysis* for omission criteria), and one participant was omitted for providing the same sentence completion ("are mimming") in every trial.

Materials

The stimuli consisted of 96 auditory preambles that were designed to elicit the 96 target sentences from Experiment 1. Each preamble consisted of the complex subject noun phrase from an Experiment 1 target sentence. The preamble structure can be described by the formula *the* + *N1* + *preposition* + *the* + *N2*.

The number of N1 and N2 were manipulated to create the same four conditions from Experiment 1: SS, SP, PP, and PS (Table 10). There were 24 preambles in each condition (see Supplementary Materials). The order of the 96 preambles was arranged in three lists, following the same order as their corresponding target sentences in lists 1–3 in Experiment 1. Each list was presented with the same 4 practice trials presented in the same order. The practice trials used four of the preambles later heard in the experiment (practice preambles are indicated in the Supplementary Materials).

Procedure

Participants were distributed across the presentation lists: 11 participants saw list 1, 10 participants saw list 2, and 4 participants saw list 3.

Preambles were presented auditorily. After hearing a preamble, participants repeated it and completed the preamble as a full sentence. The instructions given to participants were adapted from the instructions used in Bock et al. (1999). At the start of the experiment, participants were introduced to the three alien types and to the action mimming. Participants were told that they would hear phrases involving these aliens and that their job was to use each phrase as the beginning of a sentence and to complete it as a full sentence. The instructions illustrated the desired completion type using the verb *mim* through an example: "If you hear the phrase, '*The blueys to the right of the greeny*', you should complete it, '*The blueys to the right of the greeny are mimming*'". The instructions did not mention number agreement or provide examples of number agreement errors.

Participants pressed a button to initiate each trial. During each trial, the screen displayed a white X on a black background. The preamble recording started playing 100 ms after trial onset. After hearing the preamble, participants gave their response and ended the trial with another button press.

Participants completed one practice session (4 trials) prior to the experimental trial block (96 trials). The practice session followed the same format as the experimental trials. Participants were given verbal feedback at the end of each practice trial. This feedback never referenced number agreement or whether participants produced correct or incorrect agreement in their responses. If participants provided a sentence completion using a verb other than *mim*, the experimenter advised them to use the verb *mim* in their sentences.

After completing the practice session, participants proceeded to the experimental trial block. Participants were allowed to take breaks as necessary and were notified when they had completed half of the experimental trials.

Analysis

Responses to each trial were transcribed and coded for their inclusion of a number agreement error or other type of error. As in Experiment 1, agreement errors included both unrevised errors and revised errors, and incomplete productions of an agreement error were considered agreement errors with revision.

Responses were omitted from the analysis if the participant did not correctly reproduce the preamble, if verb form was unidentifiable, or if

Table 10
Experiment 3 response conditions and example preambles.

Condition	Match Condition	Example Preamble	Target Response
SS	match	the bluey above the greeny	"the bluey above the greeny is mimming"
SP	mismatch	the bluey above the greenies	"the bluey above the greenies is mimming"
PP	match	the blueys above the greenies	"the blueys above the greenies are mimming"
PS	mismatch	the blueys above the greeny	"the blueys above the greeny are mimming"

the response did not follow the target sentence formula *the + N1 + preposition + the + N2 + is/are + mimming*. Responses were coded as containing a disfluency error if the participant revised an error that would result in omission in a single revision, repeated a word or the beginning of a word, said a false start to a word, or revised a word (with the exclusion of verb revisions coded instead as agreement errors).

Responses containing no errors (number agreement or disfluency) were forced-aligned to their transcriptions. As in Experiment 1, sentences with a non-zero difference between the offset of N2 and the onset of the verb *is/are* were coded as containing a gap.

Results

Error distribution analysis

Out of the 2400 total responses, 181 responses were omitted from the analysis. Of the remaining 2219 responses, 98 responses contained disfluency errors. We obtained a total of 173 number agreement errors; 17 of these errors occurred in sentences that also contained a disfluency error. The distribution and percentage of agreement errors in each condition are presented in Table 11; participant error rates by condition (including responses with disfluency errors) are presented in Fig. 7.

The overall effect of match was significant, with errors more likely in the mismatch conditions ($\beta = 0.622$, SE = 0.154, $z = 4.038$, $p < 0.0001$). The estimated error probability for the match conditions was 2.2 % (SE = 0.7 %) compared to 7.4 % (SE = 2.1 %) in the mismatch conditions. The overall effect of N1 number was not significant ($\beta = -0.127$, SE = 0.109, $z = -1.165$, $p = 0.244$). The estimated error probabilities were 4.6 % (SE = 1.3 %) for the singular subject head conditions and 3.6 % (SE = 1.0 %) for the plural subject head conditions. Our analysis revealed a significant interaction between match and N1 number ($\beta = -0.250$, SE = 0.109, $z = -2.307$, $p = 0.021$). The SS vs. SP ($\beta = 1.745$, SE = 0.376, $z = 4.637$, $p < 0.0001$) and PP vs. PS ($\beta = 0.743$, SE = 0.377, $z = 1.970$, $p = 0.049$) contrasts were both significant such that errors were more likely in the mismatch conditions, though the SS vs. SP difference was greater than the PP vs. PS difference. The estimated error probabilities were 2.0 % (SE = 0.7 %) in the SS condition, 10.4 % (SE = 3.0 %) in the SP condition, 2.5 % (SE = 0.9 %) in the PP condition, and 5.2 % (SE = 1.7 %) in the PS condition.

Gap analysis

For the 1965 responses containing no errors (number agreement or disfluency), we analyzed the likelihood of pausing immediately prior to verb articulation. We obtained a total of 385 gaps, defined as a non-zero difference between the offset of N2 and the onset of the verb. The distribution and percentage of gaps in each condition are presented in Table 12 (a plot of participant gap rates by condition is available in the Supplementary Materials).

Our analysis revealed an overall effect of match such that gaps are significantly more likely in the mismatch conditions ($\beta = 0.514$, SE = 0.088, $z = 5.812$, $p < 0.0001$). The estimated gap probability for the mismatch conditions was 9.3 % (SE = 2.3 %) compared to 22.2 % SE = (3.4 %) in the mismatch conditions. The overall effect of N1 number was not significant ($\beta = -0.030$, SE = 0.066, $z = -0.461$, $p = 0.645$). The estimated gap probabilities were 14.6 % (SE = 2.9 %) in the singular subject head conditions and 13.9 % (SE = 2.8 %) in the plural head conditions. The interaction between match and N1 number was not significant ($\beta = -0.043$, SE = 0.066, $z = -0.654$, $p = 0.513$).

Our post-hoc gap duration analysis revealed a significant overall effect of match ($\beta = 0.161$, SE = 0.069, $t = 2.330$, $p = 0.020$). The model estimated mismatch gap duration was 146 ms (SE = 17 ms), and the estimated match gap duration was 105 ms (SE = 18 ms). The overall effect of N1 number was not significant ($\beta = 0.064$, SE = 0.049, $t = 1.310$, $p = 0.190$). The estimated plural subject gap duration was 136 ms (SE = 18 ms), and the estimated singular subject gap duration was 123 ms (SE = 16 ms). The interaction between match and N1 number was not significant ($\beta = -0.048$, SE = 0.049, $t = -0.989$, $p = 0.323$).

Table 11
Experiment 3 agreement error and response distributions (values omitting responses with disfluency errors given in parentheses).

Condition	Match Condition	Error Count	Response Count	Error Rate
SS	match	19 (18)	576 (548)	3.3 (3.3) %
SP	mismatch	86 (78)	552 (532)	15.6 (14.7) %
PP	match	22 (19)	552 (525)	4.0 (3.6) %
PS	mismatch	46 (41)	539 (516)	8.5 (7.9) %

Experiment 3 discussion

The results of Experiment 3 parallel those of Experiment 1, showing evidence of agreement attraction both in the error and gap analyses. Number agreement errors were significantly more likely when the head and local nouns in the preamble mismatched in number, and participants were significantly more likely to pause (and to pause for longer) before producing the verb in mismatch environments in correct productions. We observed evidence of a markedness effect in error likelihood (a greater contrast between the SS vs. SP conditions and the PP vs. PS conditions), though as in Experiment 1, this contrast was smaller than has been observed in prior experimentation. We discuss the markedness effect further in the *General discussion* (see also Kandel et al., 2022 where we address reduced markedness effects in detail).

Although Experiment 3 shows the same pattern of findings as Experiment 1, there are differences between the two sets of results. Participants in Experiment 3 produced fewer attraction errors than in Experiment 1, though attraction error rates were within the range seen in prior verb attraction experiments. The reduced error rates in the SP and PS conditions compared to Experiment 1 may have resulted from the removal of time pressure; while participants in Experiment 1 only had 3 s to say each target sentence, Experiment 3 was untimed, paralleling Bock et al.'s (1999) task. Participants in Experiment 3 produced preverbal gaps at higher rates in all conditions compared to Experiment 1, and the mean duration of these non-zero gaps was longer in Experiment 3 (M = 160 ms, SD = 210 ms) than in Experiment 1 (M = 134 ms, SD = 233 ms). While more plentiful and longer gaps in the mismatch conditions may in part reflect a speed–accuracy trade-off for verb attraction (participants are more accurate but slower; see Kandel et al., 2022 for evidence of a speed–accuracy trade-off for verb attraction), gaps were more common and longer than in Experiment 1 even in the match conditions where agreement should be easier to compute (Exp 1 match $M_{gap\ duration} = 173$ ms, Exp 3 match $M_{gap\ duration} = 138$ ms), meaning that the increased average time to produce the verb in Experiment 3 is not fully attributable to additional verb processing in attraction environments. The increased time to produce the verb across conditions may instead indicate that participants are focusing on correctly repeating the preamble before they plan their verb completion as opposed to planning the full response prior to onset. We discuss how elicitation paradigm may influence production in the *General discussion*.

Experiment 4

The goal of Experiment 4 was to investigate whether we obtain similar results for reflexives to those observed in Experiment 2 when the same target sentences are elicited in a preamble paradigm. We investigated both error likelihood and production time-course.

Materials and methods

Participants

The participants were 25 native English speakers ($M_{age} = 20.3$ years, SD = 3.1 years, 14 F, 11 M; age data missing from one participant) from the University of Maryland community. Participants completed the experiment in exchange for payment or course credit. An additional three omitted participants were run in the task: two participants were

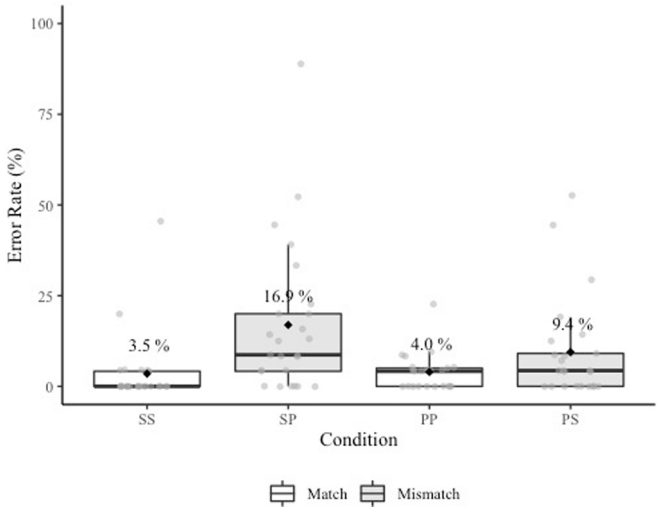

Fig. 7. ***Experiment 3 (preamble elicitation, verbs) participant agreement error rates.*** Boxplot of participant agreement error rates by condition (including responses with disfluency errors). The round points represent participant rates. Mean error rates are labeled and represented by diamonds.

omitted for failing our native speaker test, and one participant was omitted for having over 1/3 of their trials omitted (see Experiment 4 *Analysis* for trial omission criteria).

Materials

The stimuli consisted of 96 preambles designed to elicit the 96 reflexive target sentences from Experiment 2. The preambles consisted of the reflexive target sentences from Experiment 2 with the reflexive pronoun omitted; the preamble structure can thus be described by the formula *the* + *N1* + *preposition* + *the* + *N2* + *mimmed*. The form of the preambles parallels those in Bock et al. (1999) and Bock et al. (2006).

The number of N1 and N2 were manipulated to create the same four conditions used in the previous experiments (SS, SP, PP, PS) with 24 preambles in each condition (Table 13; see Supplementary Materials). The 96 preambles were arranged in three lists, following the same order as their corresponding preambles (i.e. those with the same subject phrase) in the presentation lists used in Experiment 3. Each list was presented with the same 4 practice trials presented in the same order. The practice trials used 4 of the preambles later heard in the experiment and matched the practice preambles used in Experiment 3.

Procedure

Participants were distributed across the presentation lists: 9 participants saw list 1, 11 participants saw list 2, and 5 participants saw list 3.

At the start of the experiment, participants were introduced to the three alien types and to the action mimming. Participants were told that they would hear phrases involving these aliens and that their job was to use each phrase as the beginning of a sentence and to complete it as a full sentence using a reflexive pronoun. The instructions given to participants were adapted from those used in Bock et al. (1999). These instructions did not mention number agreement or provide examples of number agreement errors.

Participants completed one practice session (4 trials) prior to the experimental block (96 trials). The practice and experiment trials in Experiment 4 followed the same procedure as Experiment 3. Participants were allowed to take breaks as necessary and were notified when they had completed half of the experimental trials.

Table 12
Experiment 3 gap distributions in responses without errors or disfluencies.

Condition	Match Condition	Gap Count	Response Count	Gap Rate
SS	match	75	530	14.2 %
SP	mismatch	123	454	27.1 %
PP	match	70	506	13.8 %
PS	mismatch	117	475	24.6 %

Analysis

Responses to each trial were transcribed and coded for their inclusion of a number agreement error or other type of error. Agreement errors included both unrevised errors and revised errors; incomplete productions of an agreement error were considered errors with revision.

Responses were omitted from the analysis if the participant did not correctly reproduce the preamble, if the pronoun was unidentifiable, if the response used a non-standard pronoun type (following Bock et al., 1999), or if the response did not follow the target sentence formula *the* + *N1* + *preposition* + *the* + *N2* + *mimmed* + *reflexive pronoun*. Responses were coded as containing a disfluency error if the participant revised an error that would in trial omission in a single revision, repeated a word or the beginning of a word, said a false start to a word, or revised a word (with the exception of pronoun revisions coded instead as number agreement errors). Using a reflexive pronoun other than *itself* or *themselves* (i.e. *himself*, *herself*) was not considered an error.

Responses containing no errors (number agreement or disfluency) were forced-aligned to their transcriptions. As in Experiment 2, sentences with a non-zero difference between the offset of N2 and the onset of the reflexive pronoun were coded as containing a gap.

Results

Error distribution analysis

Out of the 2400 total responses, 253 responses were omitted from the analysis. Of the remaining 2147 responses, 78 responses contained disfluency errors. We obtained a total of 69 number agreement errors; 5 of these errors occurred in sentences that also contained a disfluency error. The distribution and percentage of agreement errors in each condition is presented in Table 14; participant error rates by condition (including responses with disfluency errors) are presented in Fig. 8.

The overall effect of match was significant, with errors more likely in the mismatch conditions ($\beta = 1.373$, SE = 0.425, $z = 3.230$, $p = 0.001$). The estimated error probability for the match conditions was 0.3 % (SE = 0.3 %) in the match conditions compared to 5.1 % (SE = 0.9 %) in the mismatch conditions. The overall effect of N1 number was significant in the non-restricted dataset such that errors were more likely in sentences with plural subjects ($\beta = 0.455$, SE = 0.213, $z = 2.133$, $p = 0.033$). The estimated error probability for the singular subject head conditions was

Table 13
Experiment 4 response conditions and example preambles.

Condition	Match Condition	Example Preamble	Target Response
SS	match	the bluey above the greeny mimmed	"the bluey above the greeny mimmed itself"
SP	mismatch	the bluey above the greenies mimmed	"the bluey above the greenies mimmed itself"
PP	match	the blueys above the greenies mimmed	"the blueys above the greenies mimmed themselves"
PS	mismatch	the blueys above the greeny mimmed	"the blueys above the greeny mimmed themselves"

0.8 % (SE = 0.5 %) compared to 2.1 % (SE = 0.9 %) in the plural head conditions. The overall effect of N1 number was not significant in the restricted dataset ($\beta = 0.380$, SE = 0.222, $z = 1.716$, $p = 0.086$); in the restricted dataset, the estimated error probabilities were 1.0 % (SE = 0.4 %) for the singular subject conditions and 2.2 % (SE = 0.7 %) for the plural subject conditions. The interaction between match and N1 number was not significant ($\beta = -0.248$, SE = 0.213, $z = -1.164$, $p = 0.244$).

Gap analysis

For the 2005 responses containing no errors (number agreement or disfluency), we analyzed the likelihood of pausing immediately prior to reflexive pronoun articulation. We obtained a total of 461 gaps, defined as a non-zero difference between the offset of the verb and the onset of the reflexive pronoun. The distribution and percentage of gaps in each condition are presented in Table 15 (a plot of participant gap rates by condition is available in the Supplementary Materials).

The overall effect of match was significant such that gaps were more likely in the mismatch conditions ($\beta = 0.196$, SE = 0.071, $z = 2.775$, $p = 0.006$). The gap probability in the match conditions was 14.2 % (SE = 3.8 %) compared to 19.6 % (SE = 4.7 %) in the mismatch conditions. There was also a significant overall effect of N1 number such that gaps were more likely in conditions with plural subjects ($\beta = 0.511$, SE = 0.065, $z = 7.891$, $p < 0.0001$). The gap probability for the singular subject head conditions was 10.8 % (SE = 3.0 %) compared to 25.2 % (SE = 5.6 %) for plural subject conditions. The interaction between match and N1 number was not significant ($\beta = 0.075$, SE = 0.065, $z = 1.162$, $p = 0.245$).

Our post-hoc exploratory gap duration analysis revealed a significant overall effect of match ($\beta = 0.154$, SE = 0.066, $t = 2.330$, $p = 0.020$). The model estimated mismatch gap duration was 81 ms (SE = 11 ms), and the estimated match gap duration was 58 ms (SE = 9 ms). The overall effect of N1 number was not significant ($\beta = 0.062$, SE = 0.037, $t = 1.660$, $p = 0.097$). The estimated plural subject gap duration was 74 ms (SE = 10 ms), and the estimated singular subject gap duration was 64 ms (SE = 9 ms). The interaction between match and N1 number was not significant ($\beta = 0.052$, SE = 0.037, $t = 1.399$, $p = 0.162$).

Experiment 4 discussion

When we elicited the reflexive target sentences from Experiment 2 using a preamble paradigm, we saw evidence of attraction in both the error rate and production time-course measures that displayed verb attraction effects in Experiments 1 and 3 but showed no reliable attraction effects in the reflexive trials of our scene-description paradigm in Experiment 2. Participants were significantly more likely to produce number agreement errors in the mismatch conditions than the match conditions, and they were also significantly more likely to pause (and to pause for longer) before the reflexive in the same environments, even when participants produced no number agreement error. We can consequently conclude that the lack of observed reliable reflexive number attraction in Experiment 2 is not solely the result of the elicited target sentences themselves. When the same sentences are elicited with a preamble paradigm, our agreement error measures more closely resemble those of verbs, and we see reflexive number error rates more similar to those previously reported by Bock et al. (1999) (Table 16).

Table 14
Experiment 4 agreement error and response distributions (values omitting responses with disfluency errors given in parentheses).

Condition	Match Condition	Error Count	Response Count	Error Rate
SS	match	2 (2)	567 (547)	0.4 (0.4) %
SP	mismatch	24 (22)	528 (509)	4.5 (4.3) %
PP	match	7 (5)	517 (502)	1.4 (1.0) %
PS	mismatch	36 (35)	535 (511)	6.7 (6.8) %

Nevertheless, although we see attraction effects in error rates for reflexives in Experiment 4, the observed error profile does not perfectly resemble that elicited by Bock et al. (1999). We did not see evidence of a markedness effect for reflexives in Experiment 4 (there was no significant interaction between match and N1 number), and we elicited a lower percentage of errors in the SP condition (the condition that canonically has the most verb agreement errors).

The lower error rate in Experiment 4 compared to Bock et al.'s (1999) reflexive findings may result from the limited set of lexical items used in the experiment. While verb-eliciting experiments have shown robust agreement attraction with reduced lexicons (e.g. Experiment 3; Veenstra, Acheson, & Meyer, 2014), the repetition of the verb within the preambles in Experiment 4 might make the task of repeating the preamble easier and hence might allow speakers to devote more attention to planning the reflexive. Indeed, to successfully repeat the preamble in Experiment 4, participants need only store the preamble's subject phrase and then recall that they should use the verb form *mimmed* in their responses. If participants strategically store only the subject phrase, they would have extra time to plan the form of the reflexive between recitation of the stored preamble structure and the onset of the agreement target compared to Bock et al.'s (1999) task, in which participants must also store the verb. Indeed, we discuss potential evidence for reflexive planning during verb articulation in Experiment 4 in section *Exploratory production time-course analyses*.

Despite these differences, the observed attraction profile for reflexives in Experiment 4 appears to be more similar to the verb profile in Experiment 3 than the reflexive profile in Experiment 2 was to the verb profile in Experiment 1. The results of Experiment 4 thus provide a potential bridge between our data and those of Bock et al. (1999). The results suggest that the presence or absence of a contrast between reflexives and verbs may be the result of task demands. In the following section, we report the results of post-hoc analyses investigating the influence of dependency type and task on attraction effects.

Experiment comparison analyses

In order to draw conclusions about the relative size of attraction effects for the different dependencies and tasks in our study, we computed post-hoc analyses combining the data from all four experiments for both the number agreement error likelihood analysis and the pre-agreement target gap likelihood analysis.[3] The post-hoc analyses used the same model structures as the individual experiment analyses with the addition of fixed effects of dependency type (verb or reflexive) and task (description or preamble). These fixed effects were entered in a four-way interaction with match and N1 number. For the purposes of the comparison analysis, we were interested in the presence of a three-way interaction between match, dependency type, and task, which indicates whether the match effect was influenced by dependency type and

[3] We did not compute a post-hoc comparison for the exploratory gap duration analysis (investigating the length of gaps when they occurred) due to the smaller sample size.

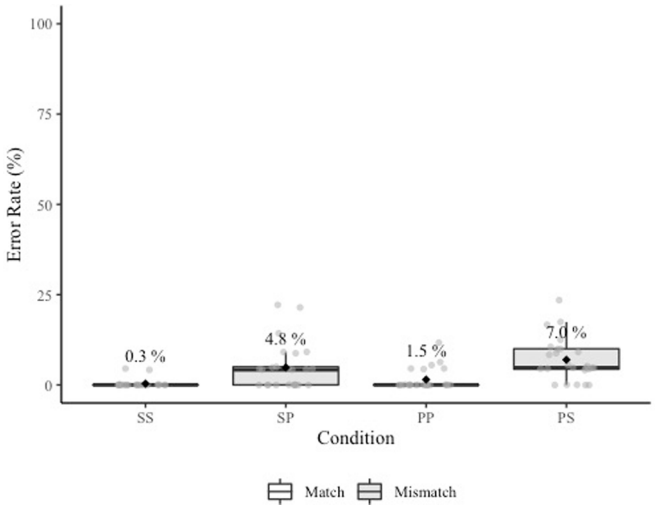

Fig. 8. ***Experiment 4 (preamble elicitation, reflexives) participant agreement error rates.*** Boxplot of participant agreement error rates by condition (including responses with disfluency errors). The round points represent participant rates. Mean error rates are labeled and represented by diamonds.

Table 15
Experiment 4 gap distributions in responses without errors or disfluencies.

Condition	Match Condition	Gap Count	Response Count	Gap Rate
SS	match	85	545	15.6 %
SP	mismatch	89	487	18.3 %
PP	match	126	497	25.4 %
PS	mismatch	161	476	33.8 %

whether that interaction was different across the two task types. We decided to retain the N1 number effect in our models due to the fact that it was a reliable predictor of our dependent variables in the individual experiment analyses for both dependency types and tasks. As in the individual experiment analyses, the fixed effects were entered into the models using effects coding. Model results are available in the Supplementary Materials. Effect plots showing the three-way interaction between match, dependency type, and task for the error and gap likelihood analyses are given in Fig. 9.

To more specifically investigate how the match effect compares in different experiments, we used the {emmeans} package v1.6.2–1 (Lenth, 2021) to compute and compare estimated marginal means (EMMs) contrasts across the different factor levels in our models. EMMs and contrasts were computed on the analysis scale. We first computed EMMs of our response variable for each level of match in each level of dependency and task (i.e. for each experiment). We contrasted these EMMs at each level of dependency type and task in a pairwise comparison to obtain the match effect for each experiment (for both the error and gap likelihood analyses, the computed match effects paralleled the results from the individual experiment analyses; see Supplementary Materials). We then applied custom effects contrast coding to contrast the computed match effects while holding either the dependency type or task constant, comparing Experiment 1 vs. Experiment 2 (verbs vs. reflexives in the scene-description task), Experiment 3 vs. Experiment 4 (verbs vs. reflexives in the preamble task), Experiment 1 vs. Experiment 3 (verbs in the scene-description task vs. the preamble task), and Experiment 2 vs. Experiment 4 (reflexives in the scene-description task vs. the preamble task).

Results

The error likelihood analysis obtained a significant three-way interaction between match, dependency type, and task ($\beta = -0.441$, SE = 0.098, $z = -4.491$, $p < 0.0001$), suggesting that the relationship between verb and reflexive attraction differed by task. The four-way interaction with N1 number was not significant ($\beta = 0.176$, SE = 0.091, $z = 1.944$, $p = 0.052$). Contrasting the match effect in each experiment, we observed a significant contrast between Experiments 1 and 2 ($\beta = 2.654$, SE = 0.585, z-ratio = 4.534, $p < 0.0001$), suggesting that verbs showed a bigger match effect in the scene-description task than reflexives. The contrast between Experiments 3 and 4 was not significant ($\beta = -0.874$, SE = 0.523, z-ratio = −1.672, $p = 0.095$). The contrast between Experiments 1 and 3 was significant ($\beta = 2.183$, SE = 0.393, z-ratio = 5.553, $p < 0.0001$), suggesting that the match effect was bigger for verbs in the scene-description task than the preamble task. The contrast between Experiments 2 and 4 was significant in the analysis with the full dataset ($\beta = -1.345$, SE = 0.679, z-ratio = −1.981, $p = 0.048$), pointing to a bigger match effect for reflexives in the preamble task than the scene-description task, however the contrast was not reliable in the analysis with no disfluency errors ($\beta = -1.251$, SE = 0.798, z-ratio = −1.568, $p = 0.117$) and was only reliable in the Bayesian analysis with 93 % confidence (median posterior estimate = −1.383, 95 % CrI [−2.82, 0.03], 93 % CrI [−2.71, −0.09]).

The gap likelihood analysis obtained a significant three-way interaction between match, dependency type, and task ($\beta = -0.116$, SE = 0.033, $z = -3.493$, $p < 0.001$), suggesting that relationship between verb and reflexive attraction differed by task. The four-way interaction with N1 number was not significant ($\beta = 0.011$, SE = 0.032, $z = 0.348$, p

Table 16
Percentage of sentences with agreement errors out of the total number of valid responses for Experiments 1–4 and Bock et al.'s (1999) simple count noun conditions. We present this measure instead of mean participant error rate to allow for a direct comparison to the reported Bock et al. (1999) data.

		Verbs			Reflexives		
Condition	Match Condition	Exp. 1	Exp. 3	Bock et al. (1999)	Exp. 2	Exp. 4	Bock et al. (1999)
SS	match	0.4 %	3.3 %	2 %	0.2 %	0.4 %	2 %
SP	mismatch	24.8 %	15.6 %	10 %	0.4 %	4.5 %	17 %
PP	match	3.0 %	4.0 %	1 %	1.1 %	1.4 %	1 %
PS	mismatch	21.2 %	8.5 %	1 %	3.3 %	6.7 %	4 %

= 0.728). There was a significant contrast in the match effect between Experiments 1 and 2 (β = 1.411, SE = 0.192, z-ratio = 7.357, p < 0.0001), suggesting that the match effect was larger for verbs than reflexives in the scene-description task. The contrast between Experiments 3 and 4 was also significant (β = 0.484, SE = 0.183, z-ratio = 2.646, p = 0.008), implying that the match effect was larger for verbs than reflexives in the preamble task. The contrast between Experiments 1 and 3 was significant (β = 0.649, SE = 0.192, z = 3.375, p < 0.001), suggesting that the match effect for verbs was larger in the scene-description task than the preamble task. The contrast between Experiments 2 and 4 was not significant (β = −0.278, SE = 0.188, z = −1.475, p = 0.140).

Experiment comparison discussion

Our post-hoc comparison analyses confirm that the attraction effects observed in our study are influenced by both dependency type and elicitation task. We observed three-way interactions between match, dependency type, and task on both the likelihood of producing a number agreement error and the likelihood of pausing directly before producing the agreement target, suggesting that the relationship between reflexive and verb attraction is influenced by the choice of elicitation paradigm.

We found evidence that in both the error and gap measures, reflexive and verb attraction effects are reliably different in the scene-description experiments. In particular, the scene-description paradigm elicited greater verb attraction effects than reflexive attraction effects, with participants more likely to produce number attraction errors and to pause before production of the agreement target in attraction-inducing environments. In the preamble experiments, on the other hand, there was no significant difference in the size of the attraction effect for reflexives and verbs in the error likelihood analysis, supporting Bock et al.'s (1999) and Bock et al.'s (2006) findings that reflexives and verbs show similar attraction error effects when elicited with a preamble paradigm. Interestingly, although there was no reliable difference in the error effect, there was a difference in the gap likelihood effect: the match effect was slightly larger for pre-VP gaps in Experiment 3 than pre-reflexive gaps in Experiment 4. Two plausible assumptions could account for this difference. Participants in the preamble paradigm may first focus on successfully recalling and repeating the stored preamble structure before planning their sentence continuations. In addition, participants in Experiment 4 might only store the subject phrase from the preamble given that the verb is the same in each trial (see *Experiment 4 discussion*). The combined effect of these factors could be that participants in Experiment 4 have additional time to plan their continuations between the offset of the stored structure and the onset of the agreement target compared to Experiment 3, where participants produced the agreement target directly after reciting the stored preamble structure (the subject phrase). Consequently, we may expect production slowdowns in Experiment 4 to be less likely to manifest directly prior to the agreement target onset, leading to fewer attraction-induced gaps before reflexives in Experiment 4 than verbs in Experiment 3 and thus a smaller match effect.

We also observed contrasts by task type in the comparison analyses. For verbs, we observed larger error and gap effects in the scene-description paradigm than the preamble paradigm. As hypothesized in the *Experiment 3 discussion*, the greater error attraction effect in Experiment 1 may be the result of the time pressure applied in the experiment. The smaller observed gap attraction effect in Experiment 3 may have been influenced by the fact that there were more gaps produced overall in the preamble task than the scene-description task (β = −0.577, SE = 0.122, z = −4.714, p < 0.0001), reducing differences between the match and mismatch conditions. We discuss potential influences of elicitation paradigm on planning in the *General discussion*.

For reflexives, there was less of a clear contrast between tasks. We observed a significant difference between Experiments 2 and 4 in the size of the error attraction effect (though note that we can only be 93 % confident of the presence of a contrast in the Bayesian analysis, and this contrast was not significant in the reduced dataset omitting sentences with disfluency errors, perhaps in part influenced by the relatively small Experiment 4 sample, N = 25). Future work should address this contrast and assess its reliability in a confirmatory study. One difference between Experiments 2 and 4 that could contribute to a potential contrast in the observed attraction error effect is that participants in Experiment 2 produced both reflexive and simple object pronouns, whereas participants in Experiment 4 only produced one type of anaphor. It is possible that eliciting two different anaphor types in Experiment 2 could have led participants to pay more close attention to anaphor form, allowing them to better avoid producing errors. However, we think that this difference is unlikely to be the primary cause of the contrast, since the object pronoun trials showed attraction effects in Experiment 2, and it is unlikely that heightened monitoring would only affect reflexive trials and not pronoun trials. In addition, the pronoun error effect was the same size when elicited in a task without reflexive trials (Wyatt et al., 2021), suggesting that the mix of trial types did not influence attraction error rates in Experiment 2. Alternatively, the different tasks may prompt different planning processes that result in differential attraction error rates. Indeed, similar to Experiment 3, we observed many more gaps after the completion of the preamble structure (i.e. between the verb and the reflexive) in Experiment 4 than in the corresponding scene-description paradigm, which could point to the fact that participants are focusing on repeating the preamble prior to planning the agreement target. Unlike Experiment 3, however, there was not much difference in average gap duration between Experiment 4 (M = 75 ms, SD = 119 ms) and the reflexive trials in Experiment 2 (M = 78 ms, SD = 153 ms). We return to the discussion of paradigm influences on planning in the *General discussion*.

We did not observe a significant difference in the gap likelihood attraction effect between the two reflexive-eliciting experiments. Looking at the gap attraction effects in Fig. 9, we see a difference in the shape of the estimated reflexive effect across the scene-description and preamble tasks. However, the standard errors for the match and mismatch gap probability estimates in Experiment 4 are very large, making it difficult to confirm the precise shape of the effect and whether it differs from the contrast of Experiment 2, which was not significant in the post-hoc model (β = 0.150, SE = 0.148, z-ratio = 1.011, p = 0.312) or in the individual experiment analysis (the post-hoc match contrasts paralleled the individual experiment analyses for all experiments). We believe that the lack of a significant difference between the two experiments is more likely to reflect uncertainty in the shape of the gap effect in Experiment 4 as opposed to reflecting that Experiment 2 showed an attraction effect similar to that in the Experiment 4 post-hoc model match contrast (β = 0.428, SE = 0.134, z-ratio = 3.188, p = 0.001) or individual experiment analysis. Future work with a larger sample of participants may be able to more precisely estimate the shape of the reflexive gap effect in the preamble paradigm. Nevertheless, while we cannot confirm from the current data that the gap attraction effect differed between Experiments 2 and 4, we can conclude from the comparison analyses that even if there is a reflexive attraction effect in Experiment 2 (in either error or gap likelihood), it is not on the same scale as the verb attraction effects in Experiment 1, as shown by the error and gap likelihood contrasts between Experiments 1 and 2. We can also conclude that the attraction contrast observed for verbs and reflexives in the scene-description experiments is different from the preamble experiments, as indicated by the observed three-way interaction between match, dependency type, and task.

Exploratory production time-course analyses

While our gap analyses probe the underlying production processes that are active directly prior to articulation of the agreement targets in our experiments, it is possible that attraction pressure on planning of the agreement target may not be constrained to this narrow time window. Therefore, to further investigate differences in production time-course

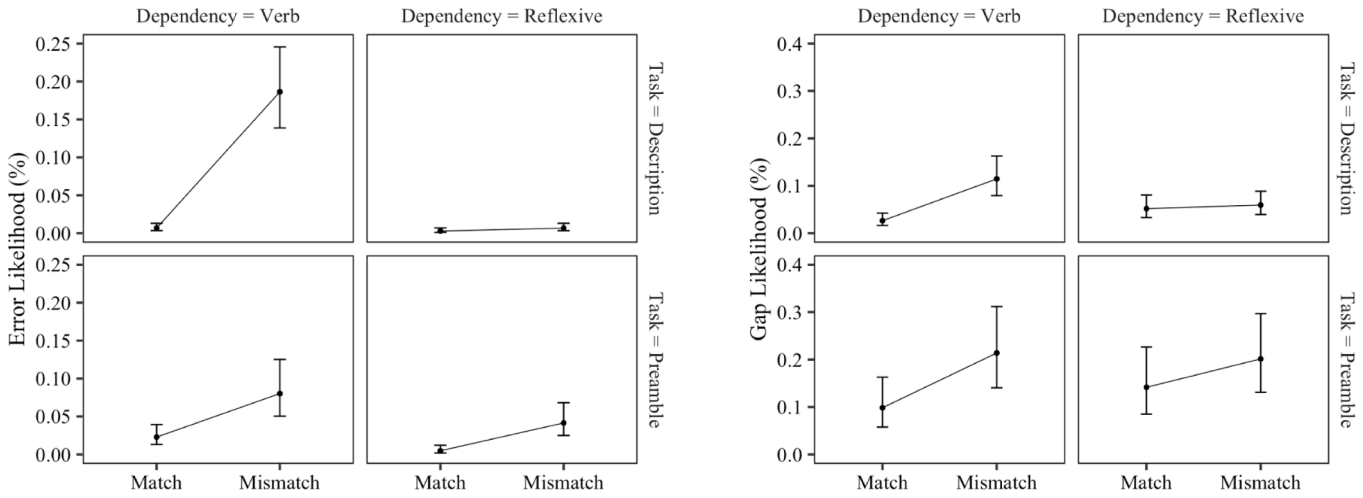

Fig. 9. ***Effect plots for the error and gap likelihood comparison analyses.*** Effect plots showing the match effect for each dependency type (reflexives, verbs) in each task (scene-description, preamble) in our study. Error bars indicate standard errors. Each panel corresponds to one experiment: Experiment 1 (Dependency = Verb, Task = Description), Experiment 2 (Dependency = Reflexive, Task = Description), Experiment 3 (Dependency = Verb, Task = Preamble), and Experiment 4 (Dependency = Reflexive, Task = Preamble).

between the different conditions in our experiments, we carried out post-hoc exploratory analyses examining the trajectory of production time-course across correct utterances. These analyses allow for a more fine-grained comparison between dependency types and elicitation paradigms. In these analyses, we used the same error-free responses from the gap analyses. We present here the production time-course analyses of the match and mismatch conditions; see the Supplementary Materials for analyses by the four sub-conditions (SS, SP, PS, PP).

For each experiment, we divided the target sentences into 5 regions (Table 17) and investigated the effect of match condition on the duration of these regions. Note that the first 3 regions are the same for the verb-eliciting experiments ("verb experiments") and the reflexive-eliciting experiments ("reflexive experiments"). We make the assumption that slower durations of a sentence region in the mismatch conditions compared to the match conditions reflect the presence of attraction pressure, providing a potential index of when the number agreement computation is active during sentence articulation.

For each utterance in our datasets, we calculated the duration of each region. Durations were measured from the offset of the previous word in the sentence to the offset of the final word in that region, with the exception of the DP1 durations, which were measured from the onset of speech to the offset of the first NP. We did not measure from the onset of the recording to the offset of DP1 due to concerns about the consistency of the recording onset times; since the experiments used continuous audio recordings that were later divided into trials rather than automatically beginning a new recording at the onset of each trial, there was the potential for unintended, inconsequential differences in trial recording onset time. Table 18 presents the grand mean sentence region durations in each condition for each experiment (calculated as means of by-participant means) as well as the grand means of the difference between the match and mismatch conditions. A plot showing the grand mean mismatch - match difference for each sentence region in each experiment is presented in Fig. 10.

We analyzed the region durations for each experiment using generalized linear mixed effects models with a gamma distribution and log link. Each experiment was analyzed separately. The models had fixed effects of region and match condition with an interaction term and random intercepts for item and participant (the inclusion of a random slope for participant by match resulted in convergence errors for some analyses, so we omitted them to allow for a standard effects structure across all experiment analyses). Fixed effects were entered into the models using effects coding. EMMs were derived from the models and compared (on the analysis scale) using the {emmeans} package v1.6.2–1 (Lenth, 2021). Table 19 presents the pairwise comparisons between the mismatch and match conditions for each region in each experiment as well as estimated Cohen's d effect sizes for each comparison (on the analysis scale). These standardized effect sizes can be used to assess differences in the size of the match effect at each region across experiments.

To more directly test for differences in the size of the match effect between experiments at regions with reliable contrasts in the individual analyses, we also computed post-hoc models pooling the data for the region of interest from the experiments we wished to compare (see Supplementary Materials for complete result summaries). As in the primary region duration analysis, the post-hoc models were generalized linear mixed effects models with a gamma distribution and log link. The models had fixed effects of match condition and experiment with an interaction and random intercepts for item and participant. Fixed effects were entered into the models using effects coding.

We first discuss the results from the verb experiments and then those from the reflexive experiments.

Results summary: Verb experiments (Experiments 1 and 3)

The production time-course analyses allow us to more precisely localize the onset of attraction pressure on verb number agreement in the sentences elicited by our verb experiments.

Aligning with our pre-VP gap analyses, we observed significant differences between the match and mismatch conditions at the *is/are* region in both Experiment 1 and Experiment 3, with longer *is/are* durations in the mismatch responses. The effect size was slightly larger for Experiment 3 (Cohen's d = 0.512) than Experiment 1 (Cohen's d = 0.476), though there was no significant interaction between match and experiment in a post-hoc test ($\beta = 0.006$, SE = 0.006, t = 0.888, p = 0.374). A match effect at the *is/are* region is consistent with our observations from the Experiment 1 and Experiment 3 gap analyses that preverbal pauses were more likely in mismatch conditions than match conditions. The effect of mismatch at the verb appears to disappear by the offset of the *is/are* region in both experiments; there were no significant differences between conditions in the duration of the verb *mimming*, suggesting that any slowdown in the VP production process resulting from attraction pressures was resolved by the time that participants finished producing the agreement target.

In Experiment 1, we additionally observed a significant *DP2* difference based on match. The *DP2* match effect in Experiment 1, although small, was approximately twice the size of that in Experiment 3 (Cohen's

Table 17
Sentence regions for the production time-course analyses.

Region	Experiments	Description	Example
DP1	1, 2, 3, 4	determiner + head noun	"the bluey"
prep	1, 2, 3, 4	prepositional head of modifier PP	"above"
DP2	1, 2, 3, 4	determiner + noun in PP modifier	"the greeny"
is/are	1, 3	auxiliary verb *is* or *are*	"is"
mimming	1, 3	the verb *mimming*	"mimming"
mimmed	2, 4	the verb *mimmed*	"mimmed"
reflexive	2, 4	reflexive pronoun	"itself"

d 0.135 vs. Cohen's d 0.067), and the interaction between match and experiment was significant in a post-hoc test ($\beta = 0.005$, SE = 0.002, t = 2.338, p = 0.019). A slowdown at *DP2* in the mismatch condition could signal extra processing difficulty planning the subject phrase, resulting from having to add a second noun phrase of differing number from the head noun (a difficulty potentially reduced in Experiment 3 by being provided the subject phrase in the preamble). Alternatively, a slowdown at *DP2* could reflect early planning of the agreement target. Looking at the production time-course of the reflexive trials in Experiment 2, whose target sentences contained the same set of subject phrases but do not require overt verb agreement marking, we do not see a significant effect of match condition at the *DP2* region. Comparing the *DP2* effect in Experiments 1 and 2, there was a significant interaction between match and experiment in a post hoc test ($\beta = 0.006$, SE = 0.003, t = 1.960, $p < 0.05$). The fact that we do not see a reliable slowdown in Experiment 2 (where the subject phrase also needs to be planned by the participant) could suggest that the mismatch slowdown in Experiment 1 was not entirely due to the construction of the subject phrase but rather was influenced by VP planning.

The fact that we do not see this same effect in Experiment 3 could indicate that the agreement target is planned later in preamble paradigm elicitations than in more naturalistic speech. If participants plan the verb only after repeating the stored preamble structure whereas verb planning starts earlier in a more naturalistic scene-description task, we would expect to see evidence of more verb planning occurring after *DP2* (the end of the preamble) in Experiment 3 than in Experiment 1. Indeed, the estimated durations of the *is/are* region were longer for Experiment 3 than Experiment 1 (the overall effect of experiment in the post-hoc test was significant; $\beta = -0.134$, SE = 0.040, t = −3.376, $p < 0.001$). The fact that region durations were not longer across the whole sentence for Experiment 3 (Table 18) suggests that the longer duration of the *is/are* region was not simply due to participants speaking more slowly in general in the preamble experiment due to lack of a time constraint. This finding is consistent with our earlier observation that pre-verbal gaps were more common and longer in Experiment 3 than Experiment 1.

The production time-course analysis thus supports the findings of our pre-VP gap analyses by serving as an additional demonstration that sentence constructions that are more likely to induce attraction errors (i.e. mismatch environments) influence the time-course of verb production, even when the correct form of the verb is ultimately produced. The match effects observed in the production time-course of the verb experiment responses suggest that verb agreement processing (and attraction pressure thereon) occurs relatively close to verb onset, potentially just in time for articulation. The analysis additionally provides evidence that verb planning may proceed differently in the preamble paradigm than it does in a more naturalistic production task.

Results summary: Reflexive experiments (Experiments 2 and 4)

As in the verb experiments, the production time-course of the reflexive experiment responses can provide evidence about when reflexive number agreement is computed and whether this timing differs across elicitation paradigms.

In neither of the reflexive analyses did we see significant timing differences in the reflexive region based on match. This result is not surprising for Experiment 2, for which we similarly observed no effect of match in our pre-anaphor gap analysis. The lack of an effect is somewhat surprising for Experiment 4, for which we observed more and longer pre-reflexive gaps in the mismatch condition than the match condition. These results could suggest that the slowdown preceding reflexive articulation observed in the Experiment 4 gap analyses did not continue through the articulation of the anaphor itself.

Although there was no mismatch difference in the reflexive region in Experiment 4, there was a trending difference in duration of the verb *mimmed* between the mismatch and match conditions (p = 0.079). This difference was reliable in the Bayesian analysis (median posterior estimate = 0.039, 95 % CrI [0.005, 0.071]). The effect size at the *mimmed* was larger in Experiment 4 (Cohen's d = 0.142) than Experiment 2 (Cohen's d = 0.035), which did not display a similar slowdown at the verb region, despite participants producing the same verb *mimmed* as in Experiment 4. Given that there is no slowdown in Experiment 2, a mismatch slowdown in Experiment 4 would be unlikely to reflect planning the invariable verb form *mimmed* (with no overt agreement) and may instead reflect attraction pressure during planning of the reflexive. An attraction effect at the verb would lend support to the hypothesis introduced in the *Experiment 4 discussion* that participants in our preamble paradigm may start computing the reflexive pronoun form before or during articulation of the verb, as a result of focusing on recalling and repeating the subject phrase of the preambles. Nevertheless, despite this trend in the data, the interaction between match and experiment was not significant in a post-hoc test ($\beta = -0.002$, SE = 0.004, t = −0.422, p = 0.673), meaning that we cannot conclusively say from the present analysis that there was a reliable difference in reflexive

Table 18
Grand mean durations and mismatch - match differences of the five sentence regions in the verb and reflexive experiments.

Exp.	Condition	Grand Mean Duration (ms)				
Verb Experiments		*DP1*	*prep*	*DP2*	*is/are*	*mimming*
1	mismatch	553 (SD = 88)	476 (SD = 79)	561 (SD = 86)	153 (SD = 49)	409 (SD = 61)
	match	540 (SD = 98)	465 (SD = 76)	532 (SD = 81)	128 (SD = 54)	421 (SD = 81)
	difference	13 (SD = 46)	10 (SD = 36)	30 (SD = 40)	25 (SD = 41)	−12 (SD = 29)
3	mismatch	514 (SD = 50)	453 (SD = 50)	587 (SD = 63)	203 (SD = 82)	454 (SD = 90)
	match	499 (SD = 57)	448 (SD = 50)	567 (SD = 60)	164 (SD = 47)	448 (SD = 88)
	difference	15 (SD = 20)	5 (SD = 24)	20 (SD = 30)	39 (SD = 47)	6 (SD = 18)
Reflexive Experiments		*DP1*	*prep*	*DP2*	*mimmed*	*reflexive*
2	mismatch	575 (SD = 99)	485 (SD = 78)	556 (SD = 69)	279 (SD = 34)	687 (SD = 68)
	match	567 (SD = 99)	477 (SD = 78)	539 (SD = 65)	274 (SD = 33)	676 (SD = 71)
	difference	8 (SD = 46)	8 (SD = 30)	16 (SD = 33)	5 (SD = 17)	11 (SD = 34)
4	mismatch	610 (SD = 61)	552 (SD = 61)	623 (SD = 60)	380 (SD = 65)	768 (SD = 92)
	match	603 (SD = 66)	541 (SD = 67)	610 (SD = 61)	361 (SD = 50)	756 (SD = 91)
	difference	7 (SD = 19)	11 (SD = 32)	13 (SD = 33)	19 (SD = 34)	12 (SD = 29)

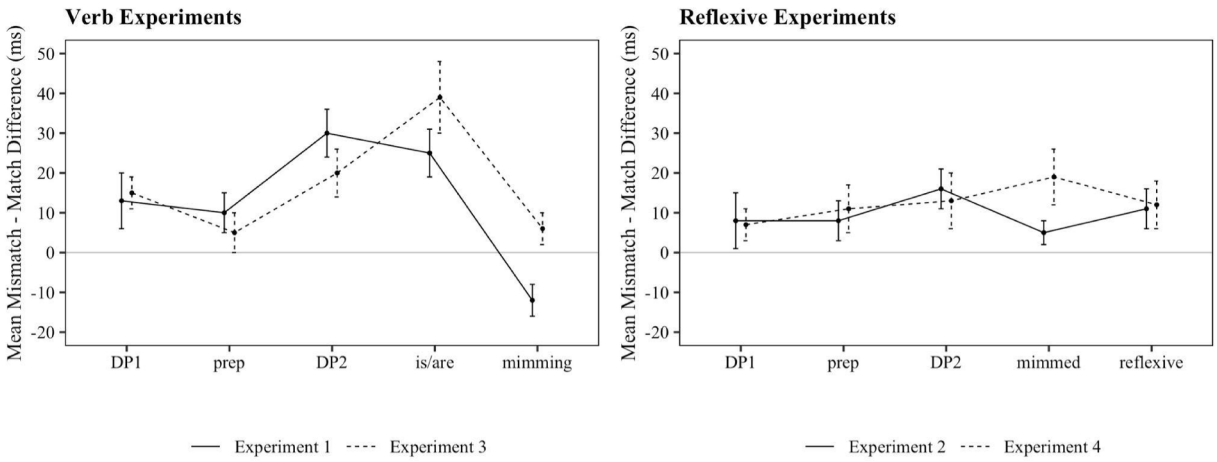

Fig. 10. Grand mean mismatch - match differences in region duration. Error bars indicate standard error of the grand mean.

planning timing between the preamble and scene-description tasks.

Comparing the raw region durations in the two experiments (Table 18), we see that all region durations tended to be longer in Experiment 4, even though the sentences produced were exactly the same. While participants in Experiment 4 may have spoken more slowly due to the fact that they did not have a time limit for articulation, this difference may also suggest that reflexive sentence planning was more difficult in the preamble task than in the scene-description paradigm. This increased difficulty may result from the need to remember, comprehend, and repeat the preamble in addition to planning reflexive form, leading production processing to proceed more slowly than it otherwise would under the average demands of natural speech.

Production time-course analyses discussion

The production time-course analyses provide a more fine-grained comparison between the elicitation paradigms and dependency types in our study. As in the gap analyses, we observed timing differences between the match and mismatch conditions in correct sentences that parallel the distributions of agreement errors, finding slowdowns during the articulation of the mismatch sentences in the same experiments that had match effects on error likelihood. The lack of match effects in the time-course of the reflexive sentences in Experiment 2 reinforces the finding that reflexive pronouns do not appear to be susceptible to agreement attraction in our scene-description paradigm.

Comparing the verb and reflexive sentences elicited in the scene-description task, we see evidence that the two dependency types are computed differently. The match effects at the *DP2* and *is/are* regions in Experiment 1 suggest that verb number planning occurs very close to verb articulation and that this planning process is susceptible to attraction pressure. By contrast, we observe no reliable effect of match condition in any sentence region in Experiment 2. This difference aligns with the contrasts observed in our error and gap distribution analyses for verbs and reflexives in the scene-description task. The fact that we see no evidence of match effects in the reflexive trial time-course but clear match effects in the verb trial time-course could arise from qualitative differences between reflexive–antecedent processing and subject–verb agreement that result in little to no interference from the local noun phrase in anaphor planning. A lack of match effects in the reflexive trial time-course analysis could also arise if reflexive planning occurs early on in utterance planning instead of proximally to target articulation. We discuss potential explanations of differential attraction susceptibility for

Table 19
Pairwise mismatch - match comparisons for the five sentence regions in the verb and reflexive experiments.

Exp.	Region	Estimate	SE	z-ratio	p-value	Cohen's d (with 95 % Confidence Interval)
Verb Experiments						
1	*DP1*	0.020	0.019	1.013	0.311	0.052 [−0.049, 0.153]
	prep	0.025	0.019	1.295	0.195	0.067 [−0.034, 0.167]
	DP2	0.051	0.019	2.631	0.009	0.135 [0.034, 0.236]
	is/are	0.179	0.019	9.257	< 0.0001	0.476 [0.375, 0.576]
	mimming	−0.025	0.019	−1.304	0.192	−0.067 [−0.168, 0.034]
3	*DP1*	0.028	0.024	1.192	0.233	0.074 [−0.048, 0.195]
	prep	0.009	0.024	0.369	0.712	0.023 [−0.099, 0.145]
	DP2	0.026	0.024	1.077	0.281	0.067 [−0.055, 0.188]
	is/are	0.197	0.024	8.244	< 0.0001	0.512 [0.390, 0.633]
	mimming	0.005	0.024	0.200	0.842	0.012 [−0.109, 0.134]
Reflexive Experiments						
2	*DP1*	0.010	0.021	0.496	0.620	0.033 [−0.097, 0.163]
	prep	0.019	0.021	0.904	0.366	0.060 [−0.070, 0.191]
	DP2	0.031	0.021	1.508	0.132	0.100 [−0.030, 0.231]
	mimmed	0.011	0.021	0.527	0.598	0.035 [−0.095, 0.165]
	reflexive	0.015	0.021	0.728	0.466	0.049 [−0.082, 0.179]
4	*DP1*	0.007	0.021	0.332	0.740	0.027 [−0.132, 0.186]
	prep	0.014	0.021	0.665	0.506	0.054 [−0.105, 0.213]
	DP2	0.020	0.021	0.950	0.342	0.077 [−0.082, 0.236]
	mimmed	0.036	0.021	1.757	0.079	0.142 [−0.017, 0.301]
	reflexive	0.011	0.021	0.533	0.594	0.043 [−0.116, 0.202]

subject–verb and reflexive–antecedent agreement in the *General discussion*.

We also observed evidence that the choice of elicitation paradigm influences number agreement planning, resulting in different time-course effects in sentences elicited in preamble and scene-description tasks. While the verb sentences from the scene-description paradigm (Experiment 1) showed a match effect at both *DP2* and the *is/are* region, suggesting that verb agreement planning begins towards the end of the complex subject phrase, the preamble-elicited verb sentences (Experiment 3) showed no evidence of a match effect until the *is/are* region. This pattern of results supports the hypothesis that participants in preamble experiments focus on reciting linguistic structure from the preamble prior to planning the completion of the sentence. The preamble-elicited reflexive sentences (Experiment 4) showed a greater trend towards a mismatch slowdown at the verb than scene-description sentences (Experiment 2) (as measured by standardized effect size). Although we cannot confirm that there was a difference in the match effect between the two experiments (the interaction in the post-hoc test was not significant), the trend in the data could suggest that reflexive planning in Experiment 4 is similarly influenced by the need to recall and repeat a preamble, leading the reflexive to be planned towards the offset of the stored preamble structure.

General discussion

In the present study, we set out to reassess whether reflexive pronouns display attraction susceptibility similar to verbs in a naturalistic scene-description paradigm. We examined number attraction effects for subject–verb and reflexive–antecedent dependencies in both our scene-description paradigm (Experiments 1 & 2) as well as a traditional preamble paradigm (Experiments 3 & 4), using the standard measure of error rate as well as the production time-course of sentences without errors. Our results can be summarized in three key observations.

First, we show that by applying a novel elicitation paradigm involving scene descriptions rather than preamble repetition and completion, we elicit a contrast between reflexives and verbs' susceptibility to attraction effects. In Experiment 1, we observed a strong agreement attraction effect for verbs, with number errors significantly more likely when the head subject noun phrase and the interfering noun phrase mismatched in number. We also saw evidence of a number markedness effect, with higher rates of attraction from plural attractors than singular attractors (though we observed reliable effects of both attractor types; see *The markedness effect and influence of singular attractors* for discussion). In Experiment 2, on the other hand, we elicited very few reflexive number errors, tallying only 47 errors total out of 3784 complete responses, even when being as generous as possible in our error coding. In contrast, Experiment 1 elicited 489 verb number errors out of 4016 complete responses. The reflexive errors were not significantly more likely in the mismatch conditions than the match conditions, showing no clear influence of the interfering attractor. Our post-hoc analysis revealed that the attraction effect was reliably different for the reflexives in Experiment 2 and the verbs in Experiment 1. Interestingly, we did observe a small but significant attraction effect on pronoun number in sentences with simple object pronouns, suggesting that the lack of a reliable attraction effect for the reflexive sentences was not simply due to an inability of our paradigm to elicit anaphor number errors. While it is possible that there exists a small reflexive attraction effect that we were underpowered to detect, it is clear that the scene-description paradigm does not elicit reflexive attraction errors on the same scale as verb attraction errors.

Our second key observation is that the production time-course of sentences with correct agreement shows the influence of attraction-inducing environments. In all experiments that found an attraction effect on error likelihood (Experiments 1, 3, and 4), we also observed slowdowns prior to production of the agreement target in the mismatch conditions even when no error was produced. In the experiment where there was no influence of the local noun on the production of the agreement target (Experiment 2, reflexive trials), there was also no influence on the timing of target articulation (though the contrast between the gap effect in Experiments 2 and 4 was not reliable in the post-hoc experiment comparison analysis, possibly due to uncertainty in the shape of the effect in Experiment 4). Our timing analyses thus appear to reflect the same attraction susceptibility observed in the error analyses. These results reinforce the finding of Staub (2009, 2010), Brehm and Bock (2013), Veenstra, Acheson, and Meyer, (2014), and Veenstra, Acheson, Bock, and Meyer (2014) that timing information can be used as a measure of agreement attraction for verbs, and they show that such effects are detectable utterance-medially in sentences elicited by a naturalistic scene-description task. Moreover, the results demonstrate that timing effects parallel behavioral results for number agreement. Such localized slowdowns prior to production of the agreement target could suggest that the elicited attraction error effects are caused by pressures active during the number agreement computation itself as opposed to reflecting earlier errors in subject number encoding (see *Interpreting the mismatch slowdowns in correct utterances*; Kandel et al., 2022 for additional discussion).

Finally, we see that the choice of elicitation paradigm affects how responses are planned and that paradigms with different task demands can lead to performance differences. Our description paradigm required participants to watch a scene, identify an action, and build a linguistic structure that conveys a message describing this action. In the preamble paradigm, participants must listen to, comprehend, remember, and repeat a preamble and complete it with a verb or reflexive pronoun. By eliciting the same target sentences in both paradigms, we were able to directly compare the scene-description paradigm with a more traditional preamble paradigm. We observed a significant three-way interaction between the match effect, dependency type, and task on the likelihood of producing a number agreement error, finding a significant contrast between reflexive and verb attraction in the scene-description experiments but not the preamble experiments. While we saw attraction effects for verbs in both paradigms, the choice of paradigm appeared to have an influence on reflexive error rates: there was little influence of attraction in Experiment 2 using the description paradigm but a significant attraction effect in Experiment 4 using a preamble paradigm. Our post-hoc comparison analysis provides preliminary evidence that this contrast is reliable, though it should be interpreted with caution. In both the reflexive and verb experiments, we observed differences in production time-course between paradigms, even though the target sentences were exactly the same. We observed more pauses before the articulation of the agreement target (i.e. after the offset of the preamble structure) in the preamble task compared to the scene-description task. The fact that participants delay more after repeating the preamble than they do in the equivalent sentence position in the more naturalistic paradigm suggests that participants may focus on repeating the preamble before planning the rest of the sentence. We also found effects of attraction on the duration of different sentence regions in the two paradigms, suggesting that the agreement target may be planned later in the preamble task. The contrasts in the error and timing analyses between paradigms suggest different sentence planning strategies prompted by the paradigms' different task demands.

In sum, our investigation reveals contrasts between dependency types as well as elicitation paradigms on number attraction susceptibility. In the following sections, we discuss our findings in more depth and explore how such contrasts may arise.

Accounting for contrasts across dependency types

Arguably the most striking finding from our study is the stark contrast between verb and reflexive attraction profiles in our scene-description paradigm: we observed strong agreement attraction for verbs but not for reflexives. While we acknowledge that there are limitations to the naturalness of the task (such as the repetitive structure of

the responses, the lexically reduced item set, the application of a time limit, and the novel characters and actions being described), we believe that the process engaged in the task of mapping from a speaker-generated message to a sentence formulation overlaps with key components of natural language production. We propose three possible accounts to explain the differential attraction susceptibility of the two dependency types. For each account, we explain how it could be reconciled with both retrieval and representational models of attraction.

One potential account for the difference between subject–verb and reflexive–antecedent dependencies is that they prioritize different information when establishing the number agreement relation. We refer to this explanation as the syntactic prioritization account. This account relies on the same basic principle applied in retrieval frameworks to explain contrasts between verbs and reflexives observed in some comprehension studies — that is, reflexive dependency resolution prioritizes structural information, whereas verb dependencies do not (e.g. Cunnings & Sturt, 2014; Dillon et al., 2013; Omaki et al., 2019). Within a retrieval model of attraction, this syntactic prioritization could manifest in different cue prioritization for reflexive pronouns compared to verbs, with reflexives prioritizing syntactic cues and verbs prioritizing morphological cues (e.g. Dillon et al., 2013). Relying on syntactic cues during retrieval may reduce the susceptibility of reflexives to attraction from non-structurally-licensed NPs by allowing the generator to more accurately retrieve the licensed antecedent. A representational model of attraction cannot capture the verb–reflexive contrast if the two dependency types target the same representation of the agreement controller. Structural prioritization for reflexives within a representational model of attraction may thus manifest as targeting only the head of the subject phrase during dependency formation rather than the representation of the subject phrase as a whole (used by verbs), thereby reducing the likelihood of agreement errors caused by the subject phrase's ambiguous or incorrect number representation.

The number source account suggests that reflexives and verbs differ in attraction susceptibility because they get their number from different sources: reflexive form is driven by semantics, whereas verb form is computed through an agreement process with the subject. This account is similar to that proposed by Eberhard et al. (2005), who suggest that reflexive number is derived from the notional number of their referents through agreement concord, whereas verbs inherit the grammatical number of the subject though agreement control. If the processes involved in determining reflexive and verb number are different and reference different representations, this could lead to differences in their attraction susceptibility. For instance, in a representational model of attraction, reflexives would not reference the same incorrect or ambiguous number feature as verbs to determine their number, thereby reducing the likelihood of error. This type of explanation can also be applied in a retrieval context by assuming that the agreement retrieval process (which is susceptible to interference) only occurs for verbs, while reflexive pronouns get their number feature from the semantics of the message. It is important to note, however, that we found that simple object pronouns do show number attraction effects; if there proves to be a reliable contrast between object pronouns and reflexives, the number source account would need to posit that reflexive anaphors and pronouns reference different sources to determine form.

Finally, the planning order account proposes that reflexive and verb forms are differentially susceptible to attraction in production because they are formulated at different times in the planning process. Under this account, reflexives are at times planned before the constituent structure has been assembled that makes the distractor available as an attractor, whereas verbs are planned after the complete subject phrase constituent has been assembled. Such a situation could arise if reflexive forms are often planned with close temporal proximity to the subject head whereas verb agreement occurs after the complete subject phrase has been planned. In the context of our experiment, this could mean that the reflexive form *itself* is planned shortly after or concurrently with the subject *the greeny* but before the prepositional phrase *above the blueys* in the sentence *the greeny above the blueys mimmed itself*. The verb, on the other hand, is planned after the constituent *the greeny above the blueys* in the sentence *the greeny above the blueys is mimming*. Earlier planning of reflexives compared to verbs may allow them to avoid attraction from distractors in a way that verbs cannot. Within a retrieval model of number agreement formation, the distractor *the blueys* is not yet present to compete in the antecedent retrieval process and lead to interference errors. Within a representational model, upwards percolation of the distractor's number feature or ambiguous encoding of the subject phrase's number has not yet had the opportunity to occur at the point when the reflexive is being planned, meaning that the representation accessed during reflexive planning is different from the incorrect/ambiguous representation accessed during verb planning.

Although some models of language production assume a direct relationship between surface word order and planning (e.g. Kempen & Hoenkamp, 1987; Levelt, 1989; Dell et al., 2008), there is evidence that sentence planning may in some cases prioritize linguistic dependencies over linear word order (e.g. Momma & Ferreira, 2019; Momma et al., 2016, 2018). It is possible that the dependency between the reflexive anaphor and the subject antecedent may be thus prioritized, leading to early planning of the anaphor before the distractor that intervenes in the surface structure. Anaphors in local subject-oriented reflexivity (LSOR; Ahn, 2015) have a close relationship to their antecedents. In prototypical reflexive constructions, LSOR anaphors and their antecedents are co-arguments of the same verb that refer to the same entity (van Hoek, 1997). Unlike non-reflexive pronouns, which are constrained by anti-locality conditions (Binding Principle B; Chomsky, 1981) and can refer to discourse antecedents outside of the sentence in which they occur, LSOR anaphors must be sufficiently proximal to their antecedents (e.g. Binding Principle A; Chomsky, 1981). LSOR anaphors and their antecedents have been proposed to be tightly semantically coupled — for example, via a processual linking relation (van Hoek, 1997). This tight syntactic and semantic relationship may lead the reflexive in our sentences to be planned with or shortly after the subject. By contrast, we do not see a reason why the verb in a sentence like *the greeny above the blueys is mimming* would be consistently planned early. In fact, Momma and Ferreira (2019) found evidence that the timing of unergative verb retrieval (e.g. in the sentence *the octopus below the lemon is swimming*) is variable (though note that timing of verb retrieval may differ from the timing of verb inflection, which may occur more regularly on a just-in-time basis). Furthermore, LSOR anaphors may be planned earlier than verbs if planning order is sensitive to hierarchical structure; it has been proposed that in LSOR constructions, the anaphor (or a copy) assumes a hierarchically-superior position close to the subject and above the verb (e.g. in a reflexive VoiceP above the vP node; Ahn, 2015).

Differences in the timing of reflexive and verb number formulation can also arise if number agreement for verbs and reflexives is instantiated at different stages of planning. In influential classic multi-stage models of language production (e.g. Bock & Levelt, 1994; Garrett, 1976, 1980), lexical selection is proposed to occur separate from (and potentially prior to) positional processes such as constituent assembly and inflection. If number is part of a reflexive anaphor's lexical representation, it is possible that selection of *itself* or *themselves* may occur during the lexical selection process. If lexical selection of the reflexive occurs relatively early in sentence processing (for any of the reasons mentioned above), it may occur at a point prior to complete constituent assembly of the subject phrase. Lexical selection of the verb, on the other hand, may only pick out the lemma for the verb *mim*, with the planning of the specific form *is mimming* or *are mimming* occurring later in sentence planning as an inflectional process, which may occur after the constituent structure of the entire verb phrase has been assembled.

The production time-course of the verb and reflexive sentences elicited in our scene-description paradigm may support the planning order hypothesis. Verb agreement sentences exhibited slowdowns in the mismatch conditions (where we expect attraction effects) shortly before articulation of the agreement target, starting at the end of the

articulation of the prepositional phrase modifier (in the *DP2* region). A comparable slowdown was not observed in this position for the reflexive sentences, suggesting that the slowdown observed in the verb sentences was related to verb planning as opposed to construction of the subject phrase. The timing data thus provides evidence that the verb number agreement computation starts after the complete subject phrase constituent has been planned and while it is being articulated. The reflexive sentences, on the other hand, did not display any major slowdowns in the mismatch conditions, leaving open the possibility that the reflexive form may be planned close to sentence onset before the distractor is available.

The presence of reliable attraction effects for object pronouns may support the hypothesis that reflexives are planned early, before the attractor is available for interference. Unlike in the reflexive trials where the anaphor captures the nature of the action (*the greeny mimmed itself* more precisely describes the event than *the greeny mimmed the greeny*), in the pronoun trials, the event could easily be described without an anaphor if the object of the action were not mentioned previously in the sentence (e.g. *the greeny mimmed the blueys*). In the pronoun trials, the choice to use an anaphor is driven by the desire to avoid repetition of a NP within close proximity to its first mention (i.e. to avoid a production such as *the greeny above the blueys mimmed the blueys*). The speaker thus likely makes the decision to pronominalize only after planning the subject phrase of the sentence containing the pronoun's antecedent, meaning that pronoun form is determined after the speaker has representations of both NPs in the sentence in memory. The fact that we observe attraction on anaphor number in a case when anaphor form is determined after the attractor is already planned supports the hypothesis that reflexives avoid attraction by being planned prior to the attractor (though note that we were unable to detect a reliable difference between object pronouns and reflexives in the present study). Indeed, when the same reflexive sentences were elicited by a preamble paradigm that encourages the reflexive to be planned after the participant already has a complete representation of subject phrase (provided in the preamble), more similar to the planning of object pronouns, we then observed attraction effects for reflexive pronouns as well (see *Influence of elicitation paradigm* for further explanation of the way that preamble paradigms may influence the timing of planning).

We thus see that the contrast in attraction susceptibility for reflexives and verbs observed in our naturalistic description paradigm can be explained by hypothesizing planning procedures that make the attractor less disruptive in reflexive sentences than verb sentences. These planning procedures may invoke distinct mechanisms of agreement formation for the two dependency types, causing reflexives to access a different representation of the subject than verbs (as in the number source account). Alternatively, the contrast between dependencies may result from the same underlying mechanism being applied to different dependency types, leading the features of the attractor to play less of a role in reflexive agreement planning (as in the syntactic prioritization account), or being applied at different times, resulting in the attractor not yet being present in the utterance representation when the reflexive is planned (as in the planning order account). A remaining question is why this same dependency type contrast is not observed in sentences elicited by a preamble paradigm.

Influence of elicitation paradigm

The contrast we observed between reflexive and verb attraction in the scene-description paradigm is a stark departure from previous preamble findings (e.g. Bock et al., 1999; Bock et al., 2006) as well as from the results of our preamble experiments, which elicited comparable attraction error rates for both dependency types. There consequently appears to be an influence of elicitation paradigm on number attraction profiles.

Many of the properties of the scene-description task that differed from prior investigations of reflexive attraction (e.g. the repetitive response structure, lexically reduced item set, and novel characters and actions) were shared by our preamble experiments. The elicited contrast in the scene-description task thus is unlikely to be due to those factors. Two factors that differed between our preamble and scene-description tasks were the application of time pressure and the inclusion of a second trial type within the experiment that elicited reflexive sentences. The time pressure in the scene-description paradigm may amplify attraction effects by increasing error likelihood (see Kandel et al., 2022 for evidence of a speed–accuracy trade-off for verb attraction). If reflexive and verb number agreement are formulated in the same fashion, we should expect to see increased error rates for both dependency types, meaning the application of time pressure should not create a contrast between reflexives and verbs. Rather, time pressure could help reveal an underlying contrast if the dependencies are formulated through processes with differential sensitivity to time pressure. The inclusion of a second trial type eliciting object pronouns in the reflexive scene-description experiment may have led to additional monitoring of anaphor form, which could reduce the size of the reflexive error attraction effect relative to verbs. However, such monitoring may have been expected to be reflected in the timing measures, and the attraction effect for the object pronoun trials does not increase when elicited without reflexive trials (Wyatt et al., 2021), suggesting that the monitoring did not reduce anaphor number attraction errors in Experiment 2. We propose that the observed difference between our scene-description and preamble experiments primarily results from the distinct task demands of the two paradigms.

The scene-description paradigm requires that participants watch a scene, identify the action, and build a linguistic structure that conveys the proper message. In the preamble paradigm, participants must listen to, comprehend, remember, and repeat a preamble and complete it with a reflexive pronoun or verb. These different task demands could result in different sentence planning strategies. Indeed, speakers in our experiments were more likely to delay after repeating the preamble than in the equivalent sentence position in the scene-description task, suggesting that participants in the preamble paradigm may first focus primarily on remembering and repeating the preamble before planning their sentence continuation. This hypothesis is consistent with the results of the production time-course analyses. For both preamble experiments, we observed evidence of mismatch slowdowns at or towards the end of the provided preamble structure. While the mismatch slowdown in the verb scene-description experiment started during articulation of the end of the subject phrase, in the verb preamble experiment, there was a match effect only after the offset of the subject phrase (i.e. the offset of the preamble). We observed a greater slowdown in this region than in the equivalent sentence position in scene-description experiment, suggesting that verb planning may have been pushed after the articulation of the subject phrase in the preamble paradigm. We additionally observed potential evidence that there may be a greater mismatch slowdown during verb articulation in the reflexive preamble experiment than in the scene-description experiment (reflected by a difference in standardized effect size). As proposed in the *Experiment 4 discussion*, since the verb is the same for every preamble, the most onerous part of correctly repeating the preamble is recalling and articulating the subject phrase. Thus, in the preamble paradigm, reflexive planning may start at the verb after reciting the demanding part of the preamble phrase (i.e. the subject phrase). While the observed trend towards a mismatch slowdown at the verb should be treated with caution (the between-experiment interaction was not significant), it would be consistent with such a planning strategy.

We offer the potentially counterintuitive suggestion that the preamble paradigm is more demanding than the scene-description paradigm. Although linguistic material is provided to the speaker, the task of accurately storing and recalling the preamble before switching attention to the completion of the sentence may be more demanding than generating the entire sentence from a scene, leading speakers not to plan later sentence material until they have successfully articulated the

preamble structure stored in their short-term memory. We propose that this strategy of deferring planning until after the recitation of the preamble (or at least the preamble structure that varies between trials) causes reflexive planning to become more susceptible to attraction. This strategy may influence reflexive planning under any of the three proposed accounts of how reflexives typically avoid attraction susceptibility. Delaying reflexive planning until after the production of the full subject phrase constituent (i.e. the preamble structure) ensures that the distractor is available for interference during reflexive number planning. Under a retrieval model of agreement, the distractor's number feature becomes available to the retrieval process, and within a representational model of attraction, the distractor number feature has had time to be incorporated in the encoding of the subject phrase number. If reflexive planning typically avoids number attraction by occurring prior to the planning of this constituent structure (as proposed by the planning order account), a delayed planning strategy would enable agreement interference that is not available to reflexives under natural conditions. Alternatively, focusing on reciting the preamble structure may lead participants not to build as rich of a structural representation of the preamble, making syntactic information less helpful when determining the number of the reflexive pronoun (reducing potential syntactic prioritization for reflexives). Within a retrieval model of attraction, an impoverished structural representation of the preamble could lead retrieval to rely more on morphological cues when resolving the reflexive dependency, resulting in attraction susceptibility similar to verbs. Within a representational model, the generator may not be able to target the head of the subject phrase as effectively, resulting in use of the same ambiguous or incorrect number representation of the subject used in verb agreement. A focus on reciting a preamble could also result in a weaker encoding of the preamble's meaning than building the same linguistic structure de novo from a message (indeed, in preamble experiments, participants typically do not see the events that they describe). A weaker semantic representation of the event may require reflexive agreement to rely on the number feature of the subject, similar to verb agreement, as opposed to deriving its number from a semantic source (thereby eliminating the potential for number source differences between reflexives and verbs).

More similar verb and reflexive attraction may also arise in preamble paradigms given that the task involves comprehension processes. Successful completion of the preamble task requires participants to parse the subject phrase provided in the preamble structure and to use this parse to guide the form of the elicited verb or reflexive. It is possible that some of the same processes that lead to attraction effects in comprehension could be at play in the interpretations of the subject phrases in the preambles. In fact, there is evidence that comparable attraction illusions can be elicited for verbs and reflexives in comprehension (Jäger et al., 2020), so it is perhaps unsurprising that a task that also involves comprehension should also show similar attraction effects for both dependencies.

We consequently attribute the presence of comparable reflexive and verb attraction effects in our preamble experiment and those of Bock et al. (1999) and Bock et al. (2006) to production strategies specific to the preamble elicitation paradigm and propose that reflexive number attraction is not a common phenomenon under more natural circumstances.

The markedness effect and influence of singular attractors

One standout finding from our verb attraction experiments was the presence of attraction from singular distractors. This finding runs counter to some previous demonstrations of the markedness effect (the finding that plural attractors are more disruptive than singular attractors), which elicited virtually no errors in sentences with singular attractors (the PS condition) (e.g. Bock et al., 2006; Bock et al., 1999; Eberhard, 1993, 1997). While we did find evidence of a markedness effect on error likelihood in our verb agreement experiments (Experiments 1 & 3), manifesting as a larger difference between the SS and SP conditions than the PP and PS conditions, our experiments demonstrated clear attraction from both plural as well as singular attractors, eliciting large numbers of errors in both the SP and PS conditions, particularly in the scene-description paradigm. We have replicated the presence of increased singular attraction using materials from Experiment 1 in a web-based experiment (Kandel et al., 2022), where we again observed a high number of PS errors and an even more reduced markedness effect, with little difference between singular and plural attraction. The presence of singular attraction in the present study thus appears to be reliable.

Although the large number of PS errors in our verb experiments may seem like an outlier, our study is not the first to demonstrate reliable attraction from singular attractors. Paradigms that share properties with our scene-description task (e.g. prompting a message through a visual display, repetitive trial structures, lexically reduced item sets, etc.) have similarly elicited reliable attraction from singular interfering nouns (e.g. Veenstra, Acheson, & Meyer, 2014; Nozari & Omaki, 2022). In a meta-analysis of attraction studies eliciting the verb *to be* that tested both singular and plural head conditions, Kandel et al. (2022) found that evidence of singular attraction is fairly common in experiments using the same agreeing verb as the present study. The investigation included eleven experiments (one of which is divided into two conditions, treated hereafter as two different experiments) from eight studies involving three different elicitation paradigms (description, forced-choice, and preamble completion) and three languages (English, Dutch, and French). All but three of the experiments showed higher error rates in the plural head mismatch conditions than the match conditions. The mean participant PS error rates in our verb experiments (22 % in Experiment 1, 9 % in Experiment 3) are within the range of PS error rates elicited in the other experiments (excluding the two experiments reported in Kandel et al., 2022 using our scene-description paradigm) (range = 3–22 %, M = 9 %, SD = 7 %), though our error rates are on the high end of the scale (particularly in Experiment 1). The difference between the mean participant PS and PP error rates in Experiment 1 (19 % points) is high compared to the other experiments surveyed by Kandel et al. (2022) (range = −1–13 % points, M = 4 % points, SD = 4 % points), though the PS-PP difference in Experiment 3 is similar to the other experiments (5 % points). Crucially, the degree of difference between plural and singular attraction (i.e. the size of the markedness effect) observed in our verb experiments falls within previously-observed ranges: the other experiments surveyed by Kandel et al. (2022) elicited a 1–20 % point difference (M = 9 % points, SD = 7 % points) between the PS-PP and SP-SS error rate percentage point differences, compared to a 6 % point difference in Experiment 1 and a 8 % point difference in Experiment 3. The reduced markedness effect we observe in our experiments thus does not appear to be an extreme outlier, and the error rates in our verb experiments are consistent with the general trend observed in the attraction literature that accuracy across conditions follows the order SS > PP > PS > SP (Staub, 2009).

Nevertheless, we recognize that there may remain potential concerns about the elevated rate of PS errors in our experiments (particularly Experiment 1). For instance, there is evidence from corpora of spontaneous speech errors that plural attraction errors are more common than singular ones (see corpora discussed in Bock & Miller, 1991; Haskell et al., 2010; Pfau, 2009). This data is consistent with a sharp markedness asymmetry. However, we believe this corpus data should be interpreted with caution, as the relative rates of plural and singular attraction errors in corpora could be influenced by the types of things individuals talk about rather than differences in agreement processes, especially when corpus measures compare what percentage of elicited errors are singular or plural (e.g. corpora discussed in Bock & Miller, 1991; Pfau, 2009) rather than how often singular and plural errors occur relative to the frequency of potential attraction-inducing environments. Haskell et al. (2010) investigate the frequency of attraction errors relative to the frequency of SS, SP, PP, and PS collective subject phrases, finding

agreement error rates of 1 % of tokens in the SS condition, 21 % of tokens in the SP condition, 0 % of tokens in the PP condition, and 3 % of tokens in the PS condition. While this distribution is consistent with a binary markedness effect, Haskell et al. (2010) report that the SP error tokens occurred almost exclusively with collective subject phrases (e.g. *a number of considerations*, *a series of consequences*), so it is unclear the extent to which the observed asymmetry is the result of markedness or collectivity, meaning that the markedness effect may be smaller in the corpus than it appears from the SP error rate. Furthermore, corpora do contain examples of spontaneous singular agreement errors (Bock & Miller, 1991; Haskell et al., 2010; Pfau, 2009), showing that singular attraction can arise in natural production.

Another potential concern arises when considering the comprehension literature on markedness effects. The markedness asymmetry is less well documented in comprehension than in production, but 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. Wagers et al., 2009; Almeida & Tucker, 2017; Dillon et al., 2017). Nevertheless, we do not believe that production attraction profiles must necessarily parallel those of comprehension given that different processes may be at play in agreement formation in comprehension and production (see Kandel et al., 2022 for discussion of how different markedness asymmetries could arise in comprehension and production). In addition, the strength of the markedness effect in comprehension has not been explored in depth, as the stark markedness effect is taken to be so reliable that comprehension studies often only include conditions with singular heads (e.g. Franck et al., 2010; Lago et al., 2015; Schlueter et al., 2019; Tanner et al., 2014; inter alia). Although similar assumptions have been made about the presence of a stark markedness asymmetry in production (many production studies also leave out the plural head conditions; e.g. Eberhard, 1999; Bock et al., 2001; Brehm & Bock, 2013; Haskell & MacDonald, 2003; Gillespie & Pearlmutter, 2011; Solomon & Pearlmutter, 2004; Veenstra, Acheson, Bock, & Meyer, 2014; inter alia), it appears that the markedness effect can have substantial variation in strength (Kandel et al., 2022). Consequently, it is possible that similar variation may exist in the strength of the markedness effect in comprehension.

There are also potential objections that the inflated rate of PS errors in Experiment 1 could be due to properties of the scene-description task. For instance, the task involves novel concepts and words, a highly reduced lexical item set, and elicits highly repetitive sentences with only one of two possible VPs (*is mimming* or *are mimming*). However, these properties are also shared by the preamble version of the task in Experiment 3, which elicited fewer PS errors, so the elevated errors in Experiment 1 cannot solely be attributable to these properties. A property of Experiment 1 not shared by Experiment 3 that may have influenced error likelihood is the time pressure applied to participants. The increased time pressure may have led to more errors in attraction-inducing environments in Experiment 1 compared to untimed preamble tasks, as speakers had less time to plan and produce their sentences. Indeed, Experiment 1 elicited higher numbers of SP errors as well as PS errors than Experiment 3 and many other preamble tasks (see Table 2 for error rates from Eberhard et al.'s, 2005 meta-analysis of preamble tasks), suggesting that avoiding errors was difficult in both mismatch conditions. Kandel et al. (2022) find evidence for a speed–accuracy trade-off for verb attraction effects in production, which supports the hypothesis that increased time pressure may influence attraction susceptibility. Even though time pressure of the sort applied in our scene-description task is less common outside of the lab, we nevertheless believe that our task still taps into natural production processes and elicits errors that reflect factors at play in natural speech contexts.

The presence of singular attraction in our verb experiments challenges the traditional interpretation of the markedness effect. The markedness effect is often explained by appealing to the notion that singular nouns are unmarked for number whereas plural nouns have a marked plural feature, which allows them to interfere with the agreement process (e.g. Bock & Eberhard, 1993; Eberhard, 1997). Our verb agreement results confute the assumption that singular nouns cannot exert any influence on number agreement due to their lack of marking. Our findings support Veenstra, Acheson, and Meyer's (2014) proposal that the effect of markedness on agreement is not binary but rather graded such that plural interfering nouns exert a stronger attraction effect than singular interfering nouns. Such a graded markedness effect accounts for the PS > SP accuracy order observed in our data and in prior literature (see Kandel et al., 2022 for discussion of how to account for a graded markedness effect in production within representational and retrieval frameworks).

The effect of head noun number

In addition to the attraction effects observed in our study, we also found significant effects of head noun number for both verbs and reflexives such that errors were more likely in conditions with plural head nouns (PP, PS) than those with singular head nouns (SS, SP). We observed this effect in all experiments except Experiment 3 (the verb preamble experiment). In the experiments with error effects, we also observed analogous effects in articulation slowdowns: pauses before the agreement target in correct utterances were more likely and (for Experiment 1) longer when the agreement controller was plural. Prior studies of verb agreement have similarly reported significant effects of head noun number on verb errors and/or an asymmetry between the match conditions, with more errors when the subject head was plural (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). In the past, this effect has been explained by an appeal to increased processing complexity for plural nouns (see Eberhard, 1997; Franck et al., 2002 for a review of evidence that plural forms are more difficult to process in language production, comprehension, and acquisition). This complexity may result from increased complexity of the plural nouns themselves (see Franck et al., 2002, who argue that plural nouns are more complex to due to having both morphological marking and semantic marking; but cf. Sauerland et al., 2008 for arguments that plurals are semantically unmarked) or because plural verb and anaphor forms may be less frequent than their singular counterparts (plural nouns are less frequent than singular nouns; Greenberg, 1966). The fact that participants in our study were more likely to pause before producing plural agreement targets supports the hypothesis that producing plural agreement forms involves an additional processing cost.

Interpreting the mismatch slowdowns in correct utterances

One key finding from the present study is that slowdowns before correct agreement targets occur in the same environments where errors are more likely. In all three experiments that elicited a reliable attraction error effect (Experiments 1, 2, and 4), pauses (i.e. gaps) before correct forms were more likely in the mismatch conditions. This finding suggests that the pressures that lead to agreement errors are active on more than the subset of trials when an error is articulated. As discussed in detail in Kandel et al. (2022), pauses directly before the agreement target may reflect either i) a time-consuming internal revision process as the speaker stops themselves from producing an error or ii) the direct influence of the same pressures that in some trials result in agreement errors. Either possibility can be easily realized within accounts of agreement attraction where errors reflect pressures that are active at the point of computing target number agreement (these contrast with accounts in which errors reflect problems that arose during initial encoding of the subject number).

For instance, in order to prompt internal revision, the generator must reach different conclusions about the target number during initial computation and in a subsequent check. This scenario arises naturally when the number agreement computation can at times result in

incorrect agreement but does so non-uniformly — for example, in a Marking and Morphing account in which errors occur probabilistically when referencing an ambiguous subject number representation or in a retrieval account in which the noise in the likelihood of erroneous retrieval results in different outcomes. If, on the other hand, agreement errors result from errors in the encoding of the subject number (e.g. in an encoding error or percolation account), it is unclear how a pre-articulation check of the agreement target would lead to a conflict that might cause an articulation delay. In order to produce a gap before the agreement target, the subject number must be re-encoded when checking the target's number; otherwise, referencing the same incorrect subject number representation during the check should result in the same incorrect conclusion. While re-encoding of the subject number is possible, we believe that parsimony favors an explanation in which errors arise from pressures active during the agreement computation itself.

If pauses prior to agreement target production do not reflect revision processes but instead directly reflect the same processes that result in errors, then similar considerations apply. The slowdowns are more straightforwardly explained as the result of pressures that are active when determining the target number, as opposed to pressures active during encoding of the agreement controller earlier in sentence planning. Under this explanation, number agreement must involve a time-consuming process that is slowed in the mismatch conditions. This requirement is easily met by a Marking and Morphing account, in which determination of the subject number is more difficult and thus slowed when the plurality of the subject phrase is close to the mid-point on the continuum between singular and plural (see Staub, 2009). In a retrieval account, slowing can result from lengthy competition between the subject head and attractor during retrieval of the agreement controller. For slowdowns to occur only in the mismatch conditions, it must be the case that this competition process is more difficult or only active when the attractor and subject head have different number features. We discuss possible ways that this could arise in Kandel et al. (2022). Timing predictions of retrieval-based accounts of agreement have been explored in more detail in comprehension studies (e.g. Engelmann et al., 2019; Jäger et al., 2020), however these predictions may not map directly onto production, as only in comprehension are agreement features on the incoming verb available as retrieval cues (Kandel et al., 2022; Slioussar & Malko, 2016).

Interestingly, in the experiments with reliable gap effects, pauses were not detected uniformly in all mismatch trials. While it is possible that there exists a smaller, more uniform slowdown effect that is not detected by the gap analysis (in our experience, the Montreal Forced Aligner v1.0.0 detects mid-word pauses of 30 ms or greater), at a minimum we can conclude that the slowdown effect is stronger for a subset of mismatch trials. This pattern suggests that the process leading to the slowdown effect may not be uniformly engaged across all mismatch trials. This variability may arise due to noise in the activation of the attractor and subject head or the initial biases of computation processes leading to differences in attraction pressure across trials (see Kandel et al., 2022 for discussion).

Conclusion

The present study compares the influence of number attraction on subject–verb and reflexive–antecedent dependencies. We compared number attraction profiles of verbs and reflexive pronouns using both a novel scene-description task and a more traditional preamble paradigm. The preamble paradigm produced number attraction effects for both verbs and reflexives, consistent with the results of previous preamble experiments (e.g. Bock et al., 1999). The scene-description task, on the other hand, displayed an asymmetry between reflexives and verbs: while verbs showed a robust agreement attraction effect in both the standard error rate measure as well as in the production time-course of sentences containing no errors, reflexives showed no evidence of attraction in either measure.

These results suggest that subject–verb and reflexive–antecedent dependences involve distinct planning procedures with different susceptibility to agreement attraction. Differential attraction susceptibility could arise if verb and reflexive number agreement invoke distinct mechanisms, derive number by referencing different sources, differentially weigh syntactic information in the dependency formation, or apply the same agreement mechanism at different times in the sentence planning process. By combining error analyses with analyses of the production time-course of correct responses, we can get a clearer overall picture of how these dependency formation processes proceed in general, rather than drawing inferences only from the subset of trials when they go awry. We suggest that the attraction asymmetry between verbs and reflexives results from different time-courses of information processing across the two dependency types and that the variability in the reflexive–verb contrast observed across paradigms reflects the demands and timing of information flow in the two tasks.

CRediT authorship contribution statement

Margaret Kandel: Conceptualization, Methodology, Investigation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Colin Phillips:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We are grateful to Cassidy Wyatt, Lalitha Balachandran, Hanna Muller, and Phoebe Gaston for their assistance with data transcription and/or collection and to Akira Omaki, Bethany Dickerson, Shota Momma, Steven Worthington, Evan DeFilippis, and Patrick Mair for advice about experiment design and/or analyses. We are also thankful for the reviewers of this paper for their helpful comments. This research was supported in part by a National Science Foundation grant [grant number DGE−1448915] to the Maryland Language Science Center, C. Phillips, PI.

References

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

Ahn, B. T. (2015). *Giving reflexivity a voice: Twin reflexives in English*. Los Angeles: University of California [Unpublished doctoral dissertation].

Almeida, D. & Tucker, M. (2017). The complex structure of agreement errors: Evidence from distributional analyses of agreement attraction in Arabic. In A. Lamont & Tetzloff, K. (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.

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. https://doi.org/10.1016/j.jml.2006.08.004

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

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

Bock, K., Butterfield, S., Cutler, A., Cutting, J. C., Eberhard, K. M., & Humphreys, K. R. (2006). Number agreement in british and american english: Disagreeing to agree collectively. *Language, 82*(1), 64–113. https://doi.org/10.1353/lan.2006.0011

Bock, K., & Cutting, J. (1992). Regulating mental energy: Performance units in language production. *Journal of Memory and Language, 31*(1), 99–127. 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. 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. 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. https://doi.org/10.1006/cogp.2001.0753

Bock, K., & Levelt, W. J. (1994). *Language production: Grammatical encoding. Handbook of psycholinguistics* (pp. 945–984). Academic Press.

Bock, K., & Miller, C. (1991). Broken agreement. *Cognitive Psychology, 23*(1), 45–93. 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. https://doi.org/10.1006/jmla.1998.2616

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

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. https://doi.org/10.3758/BRM.41.4.977

Chomsky, N. (1981). *Lectures on Government and Binding*, Foris.

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

Cunnings, I., & Sturt, P. (2014). Coargumenthood and the processing of reflexives. *Journal of Memory and Language, 75*, 117–139. https://doi.org/10.1016/j.jml.2014.05.006

Dell, G. S., Oppenheim, G. M., & Kittredge, A. K. (2008). Saying the right word at the right time: Syntagmatic and paradigmatic interference in sentence production. *Language and Cognitive Processes, 23*(4), 583–608. https://doi.org/10.1080/01690960801920735

Den Dikken, M. (2001). "Pluringulars", pronouns, and quirky agreement. *The Linguistics Review, 18*, 19–41. 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. 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. https://doi.org/10.1353/lan.2017.0003

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

Eberhard, K. M. (1997). The marked effect of number on subject–verb agreement. *Journal of Memory and Language, 36*(2), 147–164. 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. 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. https://doi.org/10.1037/0033−295X.112.3.531

Engelmann, F., Jäger, L. A., & Vasishth, S. (2019). The effect of prominence and cue association on retrieval processes: A computational account. *Cognitive Science, 43*(12), Article e12800. https://doi.org/10.1111/cogs.12800

Francis, W. N. (1986). Proximity concord in English. *Journal of English Linguistics, 19*(2), 309–317. 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. https://doi.org/10.1016/j.jml.2009.11.001

Franck, J., Lassi, G., Frauenfelder, U. H., & Rizzi, L. (2006). Agreement and movement: A syntactic analysis of attraction. *Cognition, 101*(1), 173–216. https://doi.org/10.1016/j.cognition.2005.10.003

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. https://doi.org/10.1080/01690960143000254

Garrett, M. F. (1976). Syntactic processes in sentence production. In R. J. Wales, & E. Walker (Eds.), *New approaches to language mechanisms* (pp. 231–256). North-Holland.

Garrett, M.F. (1980). Levels of processing in sentence production. In B. Butterworth (Ed.), *Language production. Vol.1: Speech and talk* (pp. 177–220). Academic Press.

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

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

Hammerly, C., Staub, A., & Dillon, B. (2019). The grammaticality asymmetry in agreement attraction reflects response bias: Experimental and modeling evidence. *Cognitive Psychology, 110*, 70–104. https://doi.org/10.1016/j.cogpsych.2019.01.001

Hartsuiker, R. J., Antón-Méndez, I., & van Zee, M. (2001). Object attraction in subject-verb agreement construction. *Journal of Memory and Language, 45*(4), 546–572. https://doi.org/10.1006/jmla.2000.2787

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

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. 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. 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. https://doi.org/10.1016/j.cognition.2009.08.017

Jäger, L. A., Mertzen, D., Van Dyke, J. A., & Vasishth, S. (2020). Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study. *Journal of Memory and Language, 111*, Article 104063. https://doi.org/10.1016/j.jml.2019.104063

Jesperson, O. (1913/1961). *A Modern English Grammar on Historical Principles,* Blackwell Munksgaard & George Allen & Unwin.

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. https://doi.org/10.1023/a:1014978917769

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

Kempen, G., & Hoenkamp, E. (1987). An incremental procedural grammar for sentence formulation. *Cognitive Science, 11*(2), 201–258. https://doi.org/10.1016/S0364−0213(87)80006-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. https://doi.org/10.1016/j.jml.2015.02.002

Lenth, R. (2021). emmeans: Estimated Marginal Means, aka Least-Squares Means. *R package version*, 1.6.2−1. https://CRAN.R-project.org/package=emmeans.

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

McAuliffe, M., Socolof, M., Mihuc, A., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *Proceedings of the Conference of the International Speech Communication Association, 18*, 498–502. https://doi.org/10.21437/Interspeech.2017−1386

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research, 29*(2), 111–123. 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. https://doi.org/10.1016/S0749−596X(02)00515−6

Momma, S., & Ferreira, V. S. (2019). Beyond linear order: The role of argument structure in speaking. *Cognitive Psychology, 114*, Article 101228. https://doi.org/10.1016/j.cogpsych.2019.101228

Momma, S., Slevc, L. R., & Phillips, C. (2016). The timing of verb selection in Japanese sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(5), 813–824. https://doi.org/10.1037/xlm0000195

Momma, S., Slevc, L. R., & Phillips, C. (2018). Unaccusativity in sentence production. *Linguistic Inquiry, 49*(1), 181–194. https://doi.org/10.1162/LING_a_00271

Nicol, J., Forster, K., & Veres, C. (1997). Subject–verb agreement processes in comprehension. *Journal of Memory and Language, 36*(4), 569–587. 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. https://doi.org/10.31234/osf.io/9pcmg.

Omaki, A., Davidson White, I., Goro, T., Lidz, J., & Phillips, C. (2014). No fear of commitment: Children's incremental interpretation in English and Japanese Wh-Questions. *Language Learning and Development, 10*(3), 206–233. https://doi.org/10.1080/15475441.2013.844048

Omaki, A., Ovans, Z., Yacovone, A., & Dillon, B. (2019). Rebels without a clause: Processing reflexives in fronted wh-predicates. *Journal of Memory and Language, 107*, 80–94. https://doi.org/10.1016/j.jml.2019.04.003

Parker, D., & Phillips, C. (2016). Negative polarity illusions and the format of hierarchical encodings in memory. *Cognition, 157*, 321–339. https://doi.org/10.1016/j.cognition.2016.08.016

Parker, D., & Phillips, C. (2017). Reflexive attraction in comprehension is selective. *Journal of Memory and Language, 94*, 272–290. https://doi.org/10.1016/j.jml.2017.01.002

Paspali, A., & Marinis, T. (2020). Gender agreement attraction in Greek comprehension. *Frontiers in Psychology, 11*, 717. 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. https://doi.org/10.1006/jmla.1999.2653

Peirce, J. W., & MacAskill, M. R. (2018). *Building Experiments in PsychoPy*. Sage.

Pfau, R. (2009). *Grammar as Processor: A Distributed Morphology Account of Speech Errors*. John Benjamins Publishing Company.

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

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

Sauerland, U., Anderssen, J. & Yatsushiro, K. (2008). The Plural is Semantically Unmarked. In S. Kepser & M. Reis (Eds.), *Linguistic Evidence* (pp. 413–434). De Gruyter Mouton. https://doi.org/10.1515/9783110197549.413.

Schlueter, Z., Parker, D., & Lau, E. (2019). Error-driven retrieval in agreement attraction rarely leads to misinterpretation. *Frontiers in Psychology, 10*, 1002. 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. https://doi.org/10.1080/01690965.2011.650900

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

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

Staub, A. (2009). On the interpretation of the number attraction effect: Response time evidence. *Journal of Memory and Language, 60*(2), 308–327. 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. https://doi.org/10.1016/j.cognition.2009.11.003

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language, 48*(3), 542–562. https://doi.org/10.1016/S0749−596X(02)00536−3

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. https://doi.org/10.1016/j.jml.2014.07.003

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

van Hoek, K. (1997). *Anaphora and Conceptual Structure*. University of Chicago Press.

Vasishth, S., Suckow, K., Lewis, R. L., & Kern, S. (2010). Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. *Language and Cognitive Processes, 25*(4), 533–567. https://doi.org/10.1080/01690960903310587

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. 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*, 783. https://doi.org/10.3389/fpsyg.2014.00783

Vigliocco, G., Butterworth, B., & Garrett, M. F. (1996). Subject-verb agreement in Spanish and English: Differences in the role of conceptual constraints. *Cognition, 61*(3), 261–298. https://doi.org/10.1016/S0010−0277(96)00713−5

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. 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. https://doi.org/10.1006/jmla.1998.2624

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. https://doi.org/10.1016/S0010−0277(98)00041−9

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

Wyatt, C., Kandel, M., & Phillips, C. (2021, March 4−6). *Number attraction in pronoun production: Evidence for antecedent feature retrieval* [Conference presentation]. 34th Annual CUNY Conference on Human Sentence Processing, University of Pennsylvania, Philadelphia, PA, United States.

Xiang, M., Dillon, B., & Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain and Language, 108*(1), 40–55. https://doi.org/10.1016/j.bandl.2008.10.002