Cognitive Psychology 110 (2019) 70–104

Contents lists available at ScienceDirect

Cognitive Psychology

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

The grammaticality asymmetry in agreement attraction reflects response bias: Experimental and modeling evidence☆

Christopher Hammerly[a,*], Adrian Staub[b], Brian Dillon[a]

[a] Department of Linguistics, University of Massachusetts Amherst, United States
[b] Department of Psychological and Brain Sciences, University of Massachusetts Amherst, United States

ARTICLE INFO

Keywords:
Agreement attraction
Grammaticality illusions
Response bias
Cue-based retrieval
Drift diffusion

ABSTRACT

Memory access mechanisms such as cue-based retrieval have come to dominate theories of the processing of linguistic dependencies such as subject-verb agreement. One phenomenon that has been regarded as demonstrating the role of such mechanisms is the *grammaticality asymmetry* in *agreement attraction*, which is the observation that nouns other than the grammatical controller of agreement can influence the computation of subject-verb agreement in ungrammatical, but not grammatical, sentences. This asymmetry is most often accounted for via the dynamics of retrieval interference. We challenge this interpretation, arguing that the asymmetry largely reflects response bias. Three forced-choice judgment experiments show that neutralizing response bias results in a decrease in the size of the grammaticality asymmetry, or its elimination altogether. Together with the response time patterns in these experiments, this result favors an account that attributes attraction effects to a continuous and equivocal representation of number, rather than to the dynamics of retrieval interference. We implement a model of grammaticality judgments that links a continuous representation of number to the rate of evidence accumulation in a diffusion process. This model accounts for the presence or absence of the grammaticality asymmetry through shifts in the decisional starting point (i.e. response bias), and highlights the importance of monitoring for response bias effects in judgment tasks.

1. Introduction

Language users draw on their grammatical knowledge during the course of routine language comprehension and production. However, the degree to which language processing is guided by a speaker's tacit grammatical knowledge remains an open question in psycholinguistics. It is clear that speakers and listeners do not always behave in a rigidly grammatical fashion: the cognitive systems that underlie language processing can fail to accurately parse linguistic input or generate well-formed linguistic output in systematic ways. Misalignments between the requirements of a speaker's grammar and what a speaker does in real time language processing have come to be called *grammatical illusions* (Lewis & Phillips, 2015; Phillips, Wagers, & Lau, 2011). Just as the vast literature on

☆ We would like to thank our colleagues at the Cognitive Science Brown Bag and the Joint Labs Meeting at UMass for feedback and discussion. Thanks also to the audience of the 31st CUNY Sentence Processing Conference at UC Davis and three anonymous reviewers for helpful input. Special thanks to Andrew Cohen for providing detailed comments on a draft of this manuscript, Jeff Starns, Ellen Lau, Heidi Lorimor, and Omer Preminger for discussion, and Jacob Prescott for assistance in running participants. Christopher Hammerly is supported by NSF GRFP (DGE-1451512). All opinions and errors are those of the authors.

* Corresponding author at: Department of Linguistics, N408 Integrative Learning Center, University of Massachusetts, 650 North Pleasant Street, Amherst, MA 01003, United States.
E-mail address: chammerly@umass.edu (C. Hammerly).

https://doi.org/10.1016/j.cogpsych.2019.01.001
Accepted 21 January 2019
Available online 22 February 2019
0010-0285/

visual illusions has played a fundamental role in the study of visual information processing, the study of these grammatical illusions has played a central role in shaping our understanding of linguistic processing. They can be studied and explored to reveal the cognitive mechanisms that underlie how our linguistic knowledge is deployed in real time (Lewis & Phillips, 2015; Phillips et al., 2011).

An important question concerns what aspects of the language processing system create these misalignments. One type of explanation appeals to errors that arise during the process of accessing linguistic representations in memory during language processing (see Lewis & Phillips, 2015 for an elucidation of this view). However, another tradition in language processing attributes these misalignments to errors or noise in the encoding of the linguistic representations used in language processing (Eberhard, Cutting, & Bock, 2005).

In this paper, we explore this contrast by investigating one of the most well studied illusions: *agreement attraction*. In Standard American English, the subject and the verb in a sentence must agree in *morphosyntactic number*, i.e., either singular or plural. For example, a sentence like *the cat is licking the cheese* is grammatical, as the singular subject (i.e. *cat*) matches in number with the auxiliary verb (i.e. *is*). Replacing this singular verb with a plural verb (*are*) results in ungrammaticality: **the cat are licking the cheese* (where "*" conventionally indicates a sentence is ungrammatical).

Despite the apparent simplicity of this constraint, speakers and listeners do not always seem to faithfully implement it in language processing. Agreement attraction occurs when the computation of subject-verb agreement is disrupted due to the presence of a distractor noun, known as the *attractor*. For example, in the classic example *the key to the cabinets is rusty* (Bock & Miller, 1991), the subject is a complex phrase consisting of a singular *head noun* (*key*) and a plural attractor noun (*cabinets*) embedded inside of a prepositional phrase modifier (introduced by the preposition *to*). In these cases, the number mismatch between the head noun and the attractor interferes with the computation of subject-verb agreement. In language production, if given this complex subject as a preamble to be completed with a verb, speakers produce a significant number of plural verbs (i.e. **the key to the cabinets* ***are*** *rusty*; Bock & Miller, 1991). In comprehension, these configurations result in the disruption of normal reading patterns (Pearlmutter, Garnsey, & Bock, 1999), and lead to errors in acceptability judgments (Clifton, Frazier, & Deevy, 1999). The phenomenon is known to be robust across a number of languages, including Spanish, French, Italian, German, Dutch, Slovak, and Russian (Badecker & Kuminiak, 2007; Hartsuiker, Schriefers, Bock, & Kikstra, 2003; Lago, Shalom, Sigman, Lau, & Phillips, 2015; Slioussar, 2018; Slioussar & Malko, 2016; Vigliocco, Hartsuiker, Jarema, & Kolk, 1996; to name a few), occurs with both grammatical gender and number features (e.g. Badecker & Kuminiak, 2007; Slioussar & Malko, 2016; Vigliocco & Franck, 1999), and has been examined using a wide variety of methodologies, among them self-paced reading, eye tracking, ERP, maze tasks, binary and scaled judgments, speed-accuracy tradeoff, free-choice production, and forced-choice production (e.g. Bock & Miller, 1991; Franck & Wagers, 2015; Nicol, Forster, & Veres, 1997; Pearlmutter et al., 1999; Staub, 2009; Tanner, Nicol, & Brehm, 2014; Wagers, Lau, & Phillips, 2009).

Despite this broad profile, it has been argued that agreement attraction illusions are constrained in a number of key ways. In comprehension, one particular constraint has been highly consequential for theories of how dependency processing proceeds: the *grammaticality asymmetry* (Wagers et al., 2009). Consider the classic examples of agreement attraction in (1).

(1)
a. **GRAMMATICAL:** *The key to the cabinets is rusty.*
b. **UNGRAMMATICAL:** **The key to the cabinets are rusty.*

In many studies, comprehenders behave in a way that suggests they are treating ungrammatical sentences (1b) as grammatical or acceptable (i.e. there is evidence of an *illusion of grammaticality*), but they do not generally behave as if grammatical sentences with a number-mismatching attractor (1a) are ungrammatical or unacceptable (i.e. there is limited evidence of an *illusion of ungrammaticality*). As we detail in our review below, this asymmetry receives a natural explanation in terms of parsing models that rely on cue-based direct-access memory architectures, and as a result, these models have come to dominate theories of agreement attraction. This type of parsing model captures the grammaticality asymmetry in terms of how memory retrieval processes necessary for comprehension interact with a *categorical* representation of number. On these models, the *plural* attractor noun in grammatical sentences like (1a) fails to match any of the retrieval cues of the *singular* verb. As a result, attractors in these sentences do not compete for retrieval, and do not create interference effects. On the other hand, the plural attractor noun in ungrammatical sentences like (1b) where the verb is also plural provides a partial match for the retrieval cues of the verb, leading to erroneous retrievals on a subset of trials, resulting in facilitatory interference. The success of these models in capturing this asymmetrical pattern in agreement attraction has contributed to a much broader deployment of memory access errors as a means of accounting for a wide variety of grammaticality illusions.

The present paper questions this interpretation of the grammaticality asymmetry, arguing instead that it is largely an artifact of response bias. This calls into question current assumptions regarding the role of memory access and cue-based retrieval in creating the agreement attraction 'illusion.' Instead, our results support an account where agreement attraction arises due to a noisy encoding of the number of the subject itself, offering a different perspective on why agreement attraction occurs. In three preregistered binary judgment experiments, we replicate the classic grammaticality asymmetry, but show that neutralizing response bias leads to the emergence of both illusions of grammaticality *and* ungrammaticality, and the elimination of the critical interaction effect indicative of asymmetrical agreement attraction.

This finding supports the claim that agreement attraction arises from a *continuous* representation of the subject's number value, as proposed within the Marking and Morphing framework (Bock, Eberhard, Cutting, Meyer, & Schriefers, 2001; Eberhard et al., 2005). Previously, such models have been thought to be inconsistent with the grammaticality asymmetry in comprehension (e.g. Wagers et al., 2009). We show that this is not necessarily the case. We fit the Ratcliff drift diffusion model (Ratcliff, 1978) to the data from the three experiments, finding support for the view that the feature mismatch between the subject head noun and attractor decreases the

rate of evidence accumulation within the decision process, regardless of whether the sentence is grammatical or ungrammatical. Crucially, we demonstrate that asymmetrical illusions emerge when the decision process is biased towards either grammatical or ungrammatical responses. Equally important, response time (RT) results from the three experiments confirm the predictions of the evidence accumulation framework, as RTs are always slower in the presence of a mismatching attractor. To begin, we present the Marking and Morphing model of agreement attraction and review the evidence that supports it.

2. A representational model of agreement attraction

The Marking and Morphing model can be said to be a representational model of agreement attraction because it locates the source of the effect in the encoding of the number of the subject. Below, we contrast representational models with alternatives that locate the source of the effect in the operations of a memory-retrieval mechanism, rather than in the representation of the subject's number itself. The family of representational models can be split into two types. First, *percolation models* (Bock & Eberhard, 1993; Eberhard, 1997; Franck, Vigliocco, & Nicol, 2002; Vigliocco, Butterworth, & Semenza, 1995) propose that the number feature of the plural embedded attractor noun phrase "percolates" up to the subject noun phrase, with a certain probability. In the cases where percolation occurs, the computation of the subject-verb agreement relation is disrupted, as the percolated plural feature is visible to the mechanism that assigns or evaluates the number of the verb. The second type, *continuous valuation models* (Eberhard et al., 2005), which are the focus of the present paper, claim that the representation of number of a subject noun phrase is continuous. In these models, the computation of the number representation of the subject is disrupted via spreading activation of plural number marking over the syntactic representation. This spreading can occur in any direction in the hierarchical syntactic representation of a sentence. On this model, agreement attraction occurs due to the equivocality of number marking on the subject, which disrupts the computation of number agreement on the verb. The only full implementation of a continuous valuation model is Marking and Morphing (Eberhard et al., 2005), which was originally proposed to account for agreement attraction effects in production. In the next section, we consider this model in more detail.

2.1. Marking and morphing

In the Marking and Morphing model, the number values of individual lexical items and pieces of morphology can range from 1 (unambiguously plural) to −1 (unambiguously singular)[1]. Consider again the classic prepositional phrase (i.e. PP) modifier *attractor mismatch* configuration, shown in (2). In the Marking and Morphing model, the number of the entire subject phrase is calculated by combining the notional number of the subject noun phrase with the spreading activation associated with the number morphology contained within the sentence. In (2), there is a single plural morpheme *-s*, located on the attractor noun *cabinets*.

(2) The key to the cabinets …

The final number value associated with the subject root node is derived by Eq. (1), adapted from the formula for spreading activation in Dell (1986). Here, the notional number of the subject is represented by $S(n)$, sources of number information by $S(m)_j$, and the weighted connection of these number sources to the root node by w_j, where the weight is a function of syntactic proximity of this number information to the root node. The final number valuation of the subject, $S(r)$, is thus the sum of the message-level valuation of the subject and the weighted sum of morpheme values.

$$S(r) = S(n) + \sum_j (w_j \times S(m)_j) \tag{1}$$

A core assumption is that $S(r)$ is the only information accessible to the mechanism that computes subject-verb agreement. If $S(r)$ either has a large value (i.e., very plural) or is very close to zero (i.e., very singular), then we may say that the number marking is relatively unambiguous. Under these conditions, subject-verb agreement can be easily computed and there will be little variation in the number marking produced. However, if $S(r)$ takes an intermediate value – that is, the number marking of the subject falls somewhere along the singular-plural continuum – then we may say that the final number is *equivocal*. This ultimately results in agreement attraction effects.

To see this, note that $S(r)$ can be mapped to probability of a plural verb by using the logistic transformation in Eq. (2), where b is a bias term set at −3.42 in order to predispose the model towards a singular default form in the absence of evidence for plurality. $S(r)$ on its own is difficult to interpret, but the logistic transformation is useful in considering how different values of $S(r)$ predict differences in the equivocality of the subject number.

$$1/\{1 + \exp - [S(r) + b]\} \tag{2}$$

[1] Note that the Marking and Morphing model as presented here applies most straightforwardly to languages like English that have only a two-way distinction in grammatical number: singular versus plural. For ease of exposition (and to establish continuity with Eberhard et al.'s model), we will simply use 'number marking' to refer to $S(r)$ on the subject phrase, as formalized for English. However, we do not wish to imply that $S(r)$ exhausts the possible values of grammatical number that might be represented cross-linguistically, especially in languages with more complex number marking systems. At present it remains an open question how to best generalize Eberhard et al.'s model to languages that have more complex grammatical number systems, and this remains an important direction for future research (though see Harrison, 2009 and Ristic, Molinaro, & Mancini, 2016 for an examination of attraction with dual and paucal attractors in Slovene and Serbian, and Badecker & Kuminiak, 2007, who discuss this issue in the context of a three-way distinction in grammatical gender in Slovak).

Returning to the example in (2), we can consider the effect of a plural morpheme embedded within a PP when the subject is a singular count noun. In the model, singular count nouns take a value of 0, plural morphemes have a value of 1.15, the weight associated with marking on the subject noun is 18.31, and the weight associated with morphology embedded in a PP is 1.39. Given that the contribution of the subject noun is zero, the value of *S(r)* for the subject phrase in (2) will simply be the product of the weight associated with PP embedding and the value of the plural morpheme: that is, 1.60. This can be contrasted with the *attractor match* case where the attractor noun is unmarked for number (i.e. *the key to the cabinet)*, where *S(r)* is 0. Applying the logistic transformation to each produces a probability of producing a plural verb in the attractor mismatch case of 0.13, and in the attractor match case of 0.01. This increased probability in the mismatch case is responsible for agreement attraction effects.

2.2. Empirical evidence for marking and morphing

Besides the standard agreement attraction effect with a PP modifier, the Marking and Morphing model predicts other effects that emerge in studies of agreement production: (i) the mismatch asymmetry; (ii) syntactic depth effects; and (iii) effects of notional plurality. The standard agreement attraction effect, as well as each of these predictions, is considered in more detail below.

Early studies on agreement attraction by Bock and colleagues (Bock & Cutting, 1992; Bock & Eberhard, 1993; Bock & Miller, 1991) revealed several conditions under which agreement production errors are likely to occur. These studies used a production elicitation technique where participants were given a complex subject preamble (e.g. *the key to the cabinets*), and were asked to repeat this preamble and provide a continuation. Agreement attraction errors appeared in the production of incorrect verb forms, and were significantly more likely when the attractor mismatched the subject head noun in number than when it matched. One of the most striking effects, known as the *mismatch asymmetry*, is that number mismatch effects are stronger when the head noun is singular (e.g. *the key to the cabinets*) than when it is plural (e.g. *the keys to the cabinet*). This effect has been quite robust in production (e.g., Eberhard, 1997; Staub, 2009), and has also been shown in comprehension measures including reading time (Patson & Husband, 2016; Wagers et al., 2009) and acceptability judgments (Häussler, 2009) in a variety of different languages.

In the Marking and Morphing model, the mismatch asymmetry is captured by the fact that embedded singular nouns do not contribute to the *S(m)* term, as they do not have a morphological realization with activation to spread. Therefore, the plural marking on the head noun itself dominates, and largely determines the *S(r)* value. The result is that a plural verb has a 0.97 probability in the plural head noun mismatch condition, and a 0.98 probability in the match condition (see Table 5 of Eberhard et al., 2005 for calculations of these probabilities).

The second critical effect predicted by Marking and Morphing is the effect of syntactic depth, first shown in production by Bock and Cutting (1992), and since replicated in a variety of production tasks and languages (Franck et al., 2002; Nicol, 1995; Solomon & Pearlmutter, 2004; and Staub, 2009; but see Gillespie & Pearlmutter, 2011; Solomon & Pearlmutter, 2004 for an alternative view based in semantic integration). In Bock and Cutting's original study, the rate of agreement errors was compared when the attractor was embedded in a PP modifier of the type previously described, and when it was in a subject relative clause modifier (e.g. *the key that opened the cabinets*). The finding is that producers make more agreement errors with mismatching attractors embedded in PP modifiers compared to relative clause modifiers. In later work, Franck et al. (2002) found more attraction from a noun in a more shallowly embedded PP (e.g. *the statue in the gardens by the mansion*) compared to a more deeply embedded one (e.g. *the statue in the garden by the mansions*). In recent work, the decrease in attraction effects with relative clause compared to PP modifiers has been shown in both scaled judgments and eye tracking measures (Hammerly & Dillon, 2017). Such effects are predicted by Marking and Morphing based on decreased weight given to morphology that is more deeply embedded in the syntactic structure.

Yet another prediction of the Marking and Morphing model is that notionally plural nouns such as collectives (e.g. *gang, team*) in the head noun position should increase the probability of a plural verb due to an increase in the value of the *S(n)* term, which represents the message-level number marking on a given head noun. Haskell and MacDonald (2003), and later Staub (2009) in a conceptual replication, showed that collective head nouns (e.g. *the gang with the dangerous rivals*) do lead to more plural verb continuations than their individual counterparts (e.g. *the leader with the dangerous rivals*) in the critical mismatch agreement attraction configuration. To our knowledge, these effects have not been tested in comprehension. Importantly, the notional number of the embedded attractor noun (i.e. *rivals* in the examples above) is not predicted to affect agreement, as it is not part of the calculation of *S(r)*—a prediction that has been validated in a number of production studies (Bock & Eberhard, 1993; Bock et al., 2001).

2.3. A challenge for representational models: the grammaticality asymmetry

As discussed in the previous section, the Marking and Morphing model captures a number of empirically attested modulations of agreement attraction effects found in both production and comprehension. However, this model of agreement attraction receives a serious challenge when we consider findings from agreement comprehension more generally. Critically, continuous valuation models predict that agreement attraction effects in comprehension should occur in both ungrammatical sentences (*the key to the cabinets are rusty*) and grammatical sentences (*the key to the cabinets is rusty*). That is, this model predicts *symmetrical* effects of mismatch in these sentences (Tanner et al., 2014; Wagers et al., 2009). This is because the number value of *S(r)* on the subject is fixed independently of the grammaticality of the sentences, which is ultimately determined by whether the number of the verb matches or mismatches the number of the subject. In other words, the number marking of the complex subject on this account is *not* determined by whether the verb is ultimately singular or plural. Therefore, equivocal number marking with attractor mismatches should occur in both ungrammatical and grammatical sentences, and should have effects on comprehension when the grammaticality of agreement is evaluated at the verb in both cases.

The literature on agreement attraction in comprehension, which we review in full detail below, largely fails to support the prediction of a *symmetrical* effect. The general consensus – despite the fact that our review reveals the effect to be surprisingly inconsistent – is that in comprehension there is a *grammaticality asymmetry* in agreement attraction effects (first observed in Wagers et al., 2009). As outlined above, the grammaticality asymmetry is the observation that in judgments (forced-choice and scaled), reading time (eye tracking and self-paced reading), and ERPs, number mismatching attractors attenuate the disruption associated with the processing of an ungrammatical verb. However, they do not disrupt the processing of a grammatical verb. It is difficult to overstate the impact that this general finding has had on the interpretation of agreement attraction effects in comprehension: The grammaticality asymmetry has been thought to uniquely favor cue-based retrieval models of agreement attraction, and has been taken as firm evidence against continuous valuation models such as Marking and Morphing (e.g. Dillon, Mishler, Sloggett, & Phillips, 2013; Schlueter, Williams, & Lau, 2018; Wagers et al., 2009). In the next section, we present cue-based retrieval models with a focus on the prediction that agreement attraction should show the grammaticality asymmetry, and then turn to a review of the literature supporting the existence of the grammaticality asymmetry.

3. Cue-based retrieval models

In the past two decades, cue-based memory models have been widely influential for theories of sentence comprehension, as they provide a general mechanism to account for the processing of a variety of long-distance dependencies beyond subject-verb agreement, including reflexive pronouns (e.g. *himself, herself, themselves*; see Dillon et al., 2013) and negative polarity items (e.g. *ever, any*; see Vasishth, Brüssow, Lewis, & Drenhaus, 2008). From linguistic theory, it is known that each of these dependencies is subject to distinct grammatical constraints. Cue-based retrieval provides a pathway for unifying the processing of these grammatically distinct dependencies (see Parker, Shvartsman, & Van Dyke, 2017 for a recent review).

The core claim is that resolving long-distance dependencies requires accessing memory. Focusing on the case of English subject-verb agreement, when the critical verb is reached, the comprehender must check to see if the form of the verb matches the features of the noun that controls the agreement relation—in this case the subject—but referred to more generally as the *agreement controller*. The hypothesis is that this checking process involves retrieving the agreement controller from memory. In a cue-based model, this retrieval process relies on a direct-access memory architecture. This architecture uses cues provided by the verb to identify the agreement controller without the need to search through the antecedents that contain non-matching cues.

Cue-based models differ in whether they instantiate memory access as a race process, where fluctuations in the activation levels of items in memory determine what is retrieved and the ultimate retrieval latency (e.g. as in the ACT-R model of Lewis & Vasishth, 2005; see also Engelmann, Jäger, & Vasishth, 2018, and Jäger, Engelmann, & Vasishth, 2017, for more recent instantiations), or by a direct-access memory retrieval process where any item that make the retrieval cues can be accessed with the same speed (e.g. Foraker & McElree, 2011; Van Dyke & McElree, 2011). However, these models are united by their prediction that interference effects arise when multiple encodings in memory match the retrieval cues. In such a model, agreement attraction occurs due to retrieval interference. Perhaps the most common view, following Wagers et al. (2009), posits that alternations in two features are responsible for agreement attraction effects: a number feature, and a feature that picks out the unique structural position of the subject noun—usually the structural case feature [+NOM] (for *nominative* case). In this system, all features are *privative* and *categorical*. For example, number is privative in the sense that plural is represented by the presence of a [+PL] feature, while singular is represented through the *absence* of a number feature, as opposed to the presence of a [+SG] feature (though see Jäger et al., 2017, and Nicenboim, Vasishth, Engelmann, & Suckow, 2018, further discussed below, for a non-privative view of number encoding using a [+SG] feature). They are categorical in the sense that a given element is either singular or plural—number cannot take an intermediate value. In this way, this explanation retains the traditional view that number is represented as categorical in both grammar and processing. Error-prone processing is attributed to errors in how these categorical grammatical representations interface with noisy memory access mechanisms (Phillips et al., 2011; Wagers et al., 2009).

Turning to the PP modifier attraction configurations introduced in (2), the assumption is that the search for an antecedent in the match sentences like **the key to the cabinet are rusty* is triggered by the main verb, *are*, which is associated with [+NOM] and [+PL] retrieval cues. The only antecedent noun that matches any of these features is *key*, which is specified for [+NOM] as the subject. However, as this noun mismatches the verb in number, the retrieval process is less likely to succeed, or may take more time, resulting in processing disruption compared to the grammatical match baseline. In the mismatch conditions, **the key to the cabinets are rusty*, each of the antecedents matches one feature of the verb: The subject *key* matches with [+NOM], and the attractor *cabinets* with [+PL]. In this situation, the distractor partially matches the retrieval cues associated with the verb, and so it may be erroneously retrieved on a subset of trials. Therefore, the presence of a distractor noun provides another encoding in memory that partially matches the retrieval cues of the verb; this additional matched encoding will on average reduce the amount of disruption induced by the ungrammatical verb *are*.

To understand how the model predicts the grammaticality asymmetry when the subject head noun is singular, consider the features of the grammatical main verb *is*. In either the match condition (i.e. *the key to the cabinet is rusty*) or the mismatch condition (i.e. *the key to the cabinets is rusty*), the search features for retrieval will only be the [+NOM] feature. As the verb is unspecified for number, changing the number features of the attractor will not allow interference effects to arise, and attraction is therefore not expected to occur in grammatical sentences. Importantly, the pattern of the asymmetry with singular subject head nouns should appear as *no* effect in grammatical sentences, and a reliable effect in ungrammatical sentences. It is not sufficient for the asymmetry to emerge only as increased attraction in the ungrammatical compared to grammatical sentences.

In a similar vein, cue-based retrieval can capture the mismatch asymmetry—the finding that mismatching attractors cause

attraction with singular, but not plural, head nouns (see Wagers et al., 2009 for evidence in comprehension). A plural head noun, being specified for both [+NOM] and [+PL], will always match more search features than the attractor, so the attractor will not cause significant interference effects. In the grammatical conditions, *the keys to the cabinet(s) are rusty*, it will match both in number and case, where the attractor can only ever match in number. In the ungrammatical sentences, *the keys to the cabinet(s) is rusty*, the search will occur with respect to [+NOM]. Since only the subject matches that feature, it will always be retrieved. The fact that it mismatches in number will lead the dependency to be evaluated as ungrammatical.

As noted above, Jäger et al. (2017) account for retrieval interference effects utilizing a non-privative system of number encoding, which encodes singular as [+SG] and plural as [−SG]. Importantly, adopting this view does not affect the prediction of grammatically asymmetrical agreement attraction: In the grammatical mismatch conditions with singular agreement controllers, the plural attractor, which is encoded with [−SG], will still fail to match the [+SG] feature of the verb. In contrast, a plural attractor will provide a match for the [−SG] verb in the ungrammatical conditions. However, a potential challenge for a non-privative account is capturing the mismatch asymmetry. If singular is specified with a [+SG] feature, then a singular attractor should be expected to generate analogous similarity-based interference effects to its plural counterpart. Due to this empirical challenge, we retain the original privative feature structure proposed by Wagers et al. (2009).

4. The grammaticality asymmetry: a review

As the preceding section makes clear, the presence or absence of the grammaticality asymmetry emerges as a critical data point that distinguishes representational approaches to agreement attraction from memory access-oriented approaches. The appropriate explanation for agreement attraction is an important issue to resolve, as the distinction between representational and memory access-oriented approaches carries broad implications for how number is represented (categorical versus continuous), and how core grammatical processes interact with memory during language comprehension.

The perception in the literature on agreement comprehension is that asymmetrical effects of grammaticality in agreement attraction are stable and widely observed (see, e.g. Schlueter et al., 2018 for a recent and clear elucidation of this view). In order to evaluate this claim, we undertook a comprehensive and systematic review of the comprehension literature, summarized in Table 1. Note that this is not a formal meta-analysis, but an informal summary of known findings in comprehension. To create the table, we split each study and experiment by dependent measure. Some experiments, for example Experiment 1 in Tanner et al. (2014), are thus split across the table: One row for accuracy in the end-of-sentence judgments, and one for the ERP measures. We then report whether the study shows a significant mismatch effect in the grammatical and ungrammatical sentences, and whether it shows the critical interaction—larger mismatch effect in ungrammatical sentences—that defines the grammaticality asymmetry. We categorize each study as either showing a significant effect, failing to show a significant effect, not reporting a result that could in principle be tested, or as not having an experimental design such that the test can be conducted (e.g. the interaction cannot be tested if the study only contains grammatical sentences).

To summarize the overall results, we found that only 27 of the 45 studies that directly tested the grammaticality asymmetry found a significant interaction between grammaticality and attractor number. Moreover, 15 of the 40 studies that reported effects of attractor number separately for grammatical and ungrammatical sentences found significant effects of attractor number in grammatical sentences (compare to 32 of 35 that found a significant effect of mismatch with ungrammatical sentences). We regard these findings as surprising given the current consensus regarding the reliability of the grammaticality asymmetry and the lack of agreement attraction in grammatical sentences. These facts set the stage for the present study.

At this juncture, we should note that our informal summary of the literature has clear limitations: We are relying on a simple "significant or not significant" criterion for many studies that may be underpowered to detect the critical interaction (see Jäger et al., 2017, who offer simulations in support of this; see Vasishth, Mertzen, Jäger, & Gelman, 2018, on the perils of using statistical significance as a categorical filter). This places limits on how strongly this should be interpreted. Even so, we believe this pattern of results is surprising given the widespread perception about the reliability of grammaticality asymmetry.

4.1. The grammaticality asymmetry in reading

In both eye-tracking and self-paced reading, ungrammatical verbs are associated with longer reading times (and more regressive saccades) at the critical verb or the word following the verb (Clifton et al., 1999; Deevy, 1999; Pearlmutter et al., 1999). Agreement attraction effects appear as either faster reading times at ungrammatical verbs in the attractor mismatch condition compared to the match condition, or less commonly (due to the purported grammaticality asymmetry) as slower reading times at grammatical verbs in the mismatch condition compared to the match (Dillon et al., 2013; Enochson & Culbertson, 2015; Franck, Colonna, & Rizzi, 2015; Lago et al., 2015; Parker & Phillips, 2017; Patson & Husband, 2016; Pearlmutter et al., 1999; Schlueter et al., 2018; Tucker, Idrissi, & Almeida, 2015; Villata, Tabor, & Franck, 2018; Wagers et al., 2009). These effects have been shown not only with PP and relative clause modifiers, where the attractor noun linearly intervenes between the subject and the critical verb, but also at the embedded verb of an object relative clause, where the attractor noun does *not* linearly intervene (e.g. *the runners the driver kindly **wave** to*; see Wagers et al., 2009).

In our review, only 11 of the 22 studies that tested for the interaction indicative of the grammaticality asymmetry found a significant effect. Of the studies that ran contrasts to test for an effect of attractor number in grammatical sentences, 7 of the 20 studies found a significant effect, while all of the 16 studies that tested for effects of attractor number in ungrammatical sentences found a significant effect. We reserve full discussion of the wider implications of this finding for the general discussion section, noting

Table 1
Summary of comprehension literature on agreement attraction. "Yes" reflects a significant effect of $p < 0.05$ or $t > 2$, "No" reflects a non-significant effect, "N/R" means the test was not run or reported, and "N/A" that the test does not apply given the design of the experiment. When relevant, studies are split by measure and construction-type, therefore the same experiment may appear on multiple lines. We used the following abbreviations: PP = prepositional phrase, RC = relative clause, SRC = subject relative clause, ORC = object relative clause.

Citation	Construction	Language	N	Grammatical mismatch effect?	Ungrammatical mismatch effect?	Grammaticality asymmetry?
Self-Paced Reading						
Wagers et al. (2009), E2	ORC	English	30	No	Yes	Yes
Wagers et al. (2009), E3	ORC	English	60	No	Yes	Yes
Wagers et al. (2009), E4	PP Modifier	English	46	Yes	Yes	Yes
Wagers et al. (2009), E5	PP Modifier	English	60	No	Yes	Yes
Wagers et al. (2009), E6	PP Modifier	English	30	No	N/A	N/A
Pearlmutter et al. (1999), E1	PP Modifier	English	80	Yes	Yes	No
Pearlmutter et al. (1999), E3	PP Modifier	English	50	Yes	N/A	N/A
Lago et al. (2015), E1	ORC	Spanish	32	No	Yes	Yes
Lago et al. (2015), E2	ORC	English	32	No	Yes	Yes
Lago et al. (2015), E3A	ORC	Spanish	32	Yes	Yes	No
Lago et al. (2015), E3B	ORC	Spanish	32	No	Yes	No
Tucker et al. (2015), E1	SRC modifier	Arabic	114	No	Yes	Yes
Schlueter et al. (2018), E2	PP Modifier	English	42	No	Yes	Yes
Schlueter et al. (2018), E5	PP Modifier	English	41	N/R	N/R	No
Franck et al. (2015), E1	ORC	French	72	Yes	N/A	N/A
Patson and Husband (2016), E1	PP Modifier	English	72	Yes	Yes	No
Enochson and Culbertson (2015), E2	PP Modifier	English	82	N/R	N/R	No
Enochson and Culbertson (2015), E3	ORC	English	60	N/R	N/R	No
Villata et al. (2018), E2	ORC	English	130	No	N/A	N/A
Tanner (2011), E3b	PP Modifier	English	80	N/R	N/R	No
Tanner (2011), E3b	RC Modifier	English	80	N/R	N/R	No
Parker, Lago, and Phillips (2015), E2	SRC Modifier	English	32	No	Yes	Yes
Parker et al. (2015), E3	SRC Modifier	English	32	No	Yes	Yes
Eye Tracking						
Pearlmutter et al. (1999), E2	PP Modifier	English	64	Yes	Yes	No
Dillon et al. (2013), E1	SRC modifier	English	40	No	Yes	Yes
Parker and Phillips (2017), E2	ORC modifier	English	30	N/R	N/R	No
Binary Acceptability Judgments						
Wagers et al. (2009), E7	PP Modifier	English	16	No	Yes	Yes
Tanner et al. (2014), E1	PP Modifier	English	24	No	Yes	Yes
Tanner et al. (2014), E2	PP Modifier	English	22	Yes	Yes	No
Tanner et al. (2014), E3	PP Modifier	English	36	No	N/A	N/A
Schlueter et al. (2018), E1	PP Modifier	English	30	N/R	N/R	Yes
Schlueter et al. (2018), E3	PP Modifier	English	30	N/R	N/R	Yes
Schlueter et al. (2018), E4	PP Modifier	English	30	N/R	N/R	Yes
Franck et al. (2015), E2	ORC	French	30	No	N/A	N/A
Franck et al. (2015), E3	ORC	French	26	No	Yes	Yes
Häussler (2009), E1	PP Modifier	German	48	Yes	No	No
Häussler (2009), E2	Possessive RC	German	32	Yes	Yes	No
Häussler (2009), E3	Possessive RC	German	64	Yes	Yes	No
Häussler (2009), E5	Possessive RC	German	40	Yes	Yes	No
Häussler (2009), E6	Possessive RC	German	40	Yes	Yes	No
Wagers (2008), E3	ORC	English	16	N/R	N/R	Yes
Wagers (2008), E5	ORC	English	24	N/R	N/R	Yes
Wagers (2008), E5	ORC	English	24	N/R	N/R	Yes
Lago et al. (2018), E1	Possessive	Turkish	44	No	Yes	Yes
Tanner (2011), E1	PP Modifier	English	17	No	Yes	Yes
Scaled Acceptability Judgments						
Hammerly and Dillon (2017), E1	PP Modifier	English	64	Yes	Yes	Yes
Hammerly and Dillon (2017), E1	RC Modifier	English	64	Yes	Yes	Yes
Dillon et al. (2013), E1	SRC modifier	English	12	No	No	N/A
EEG						
Shen et al. (2013), E2	PP Modifier	English	24	No	Yes	Yes
Tanner et al. (2014), E1	PP Modifier	English	24	No	Yes	Yes
Tanner et al. (2014), E2	PP Modifier	English	22	No	Yes	Yes
Tanner (2011), E1	PP Modifier	English	17	No	Yes	No
Tanner (2011), E2	PP Modifier	English	21	No	No	N/A

only that this finding shows the grammaticality asymmetry may also be unreliable in more implicit measures such as reading for comprehension. Other implicit measures such as ERP have shown near uniform support for the grammaticality asymmetry (Shen, Staub, & Sanders, 2013; Tanner et al., 2014), but we will argue that these findings, too, should be considered in light of a wider constellation of findings, including those we present in the current article.

4.2. The grammaticality asymmetry in judgments

Binary yes/no judgment tasks have been the most widely used methodology for gathering judgment data about agreement comprehension, with agreement attraction effects shown to be present across more or less the same range of languages and constructions as observed in reading (Dillon et al., 2013; Franck et al., 2015; Häussler, 2009; Tanner et al., 2014; Wagers, 2008; Wagers et al., 2009). In general, sentences in which the verb shows incorrect agreement with the subject are correctly judged to be ungrammatical, and sentences in which the verb shows correct agreement are correctly judged to be grammatical. Agreement attraction effects therefore appear as a decrease in proportion correct in sentences with a mismatching attractor compared to a matching attractor. If grammatically asymmetrical agreement attraction holds, then this mismatch effect of accuracy appears as an interaction such that there is no difference between match and mismatch conditions in grammatical sentences, but accuracy is lower with the ungrammatical mismatch conditions.

As in the studies of agreement comprehension in reading, judgment studies show a mixed record with respect to the grammaticality asymmetry. Of the 19 studies that directly tested for the presence of the asymmetry, 13 found a significant interaction. Of the studies that tested contrasts, exactly half (8 of 16) found a significant effect of attractor number in grammatical sentences, with 12 of 14 finding an effect of attractor number in ungrammatical sentences. Again, this suggests that agreement attraction may not be as strongly asymmetrical with respect to grammaticality as previously thought. The present study pinpoints a source of this variation: differences in response bias across experiments. To understand how response bias may matter for the grammaticality asymmetry, we describe a model of the judgment process that underlies the grammaticality decisions in the experiments reported above. It is to this that we now turn.

5. Modeling grammaticality judgments: continuous number and drift diffusion

Staub (2008, 2009; see also Brehm, 2014) proposed a linking hypothesis between the continuous number value *S(r)* from Marking and Morphing and the process of evidence accumulation that culminates in the production of a verb form. He proposed to model this evidence accumulation process with the Drift Diffusion Model (Ratcliff, 1978)—a model designed to recover the cognitive variables underlying binary decision tasks. Here, we extend this idea to grammaticality judgments. We ask: Can a diffusion model that links *S(r)* and rate of evidence accumulation account for the observed pattern of grammaticality judgments in the agreement studies? In short, we find that it can. This is surprising, and it motivates the current study: While it has been thought that continuous valuation models such as Marking and Morphing cannot account for asymmetrical agreement, we find that the grammaticality asymmetry can emerge in judgment measures when bias is present in the response patterns. Only with unbiased responders is the underlyingly symmetrical profile of agreement attraction predicted to appear in the judgment results.

Diffusion models allow for the recovery of cognitively meaningful parameters thought to underlie two-choice decision tasks, including variables related and unrelated to the decision process itself. These parameters, their relation to the decision task, and their potential relation to agreement processing, are detailed below. In contrast to cue-based retrieval, the diffusion model is not advanced as a model of the specific cognitive processes that underlie differential categorization of stimuli, but rather as a tool that allows researchers to identify the cognitive sources of response accuracy and latency distributions in decision tasks commonly used in experimental studies (see Ratcliff & McKoon, 2008, and Voss, Nagler, & Lerche, 2013, for reviews).

Fig. 1 schematizes the diffusion decision model. The top panel shows how the decision process is situated between an encoding stage, which includes processes such as visual encoding and word recognition, and a response output stage, which encompasses the execution of a physical button press. For the purposes of fitting the model to data, the two stages on either side of the decision process itself are combined in a single parameter, T_0, which represents the mean duration of all non-decisional processes.

The bottom panel of Fig. 1 schematizes the core decision process. The diffusion model assumes that decisions are made by a noisy information accumulation process. The process begins at a starting point between the two possible decision boundaries (in our example, responding either grammatical or ungrammatical), and completes when the accumulation process reaches one boundary or another. The stochastic nature of the accumulation process leads to the possibility of reaching the "wrong" boundary, as well as the characteristic right-tailed distribution of response RT, shown over each boundary in the figure. Besides non-decisional time, there are three parameters that most commonly account for variability in decision tasks. These parameters define the separation between the decision boundaries (threshold separation; a), the starting point of the diffusion process (z), and the average slope of the diffusion process (drift rate; v). Each of these parameters is known to be sensitive to distinct underlying variables.

Threshold separation (a) represents the amount of information that is needed to make a decision. The higher a is, the more information that is needed. This parameter is frequently interpreted as response caution, as fitting the model to more conservative responders recovers a higher value of a. The parameter is also known to be sensitive to speed and accuracy instructions, with greater focus on accuracy leading to higher values of a, and greater focus on speed resulting in lower values (Voss, Rothermund, & Voss, 2004).

The starting point parameter (z) represents initial bias towards one response or another. As z is only interpretable in relation to a, it is frequently presented in a relativized form, $z_r = z/a$, which can range from 0 (starting at the lower boundary) to 1 (starting at the

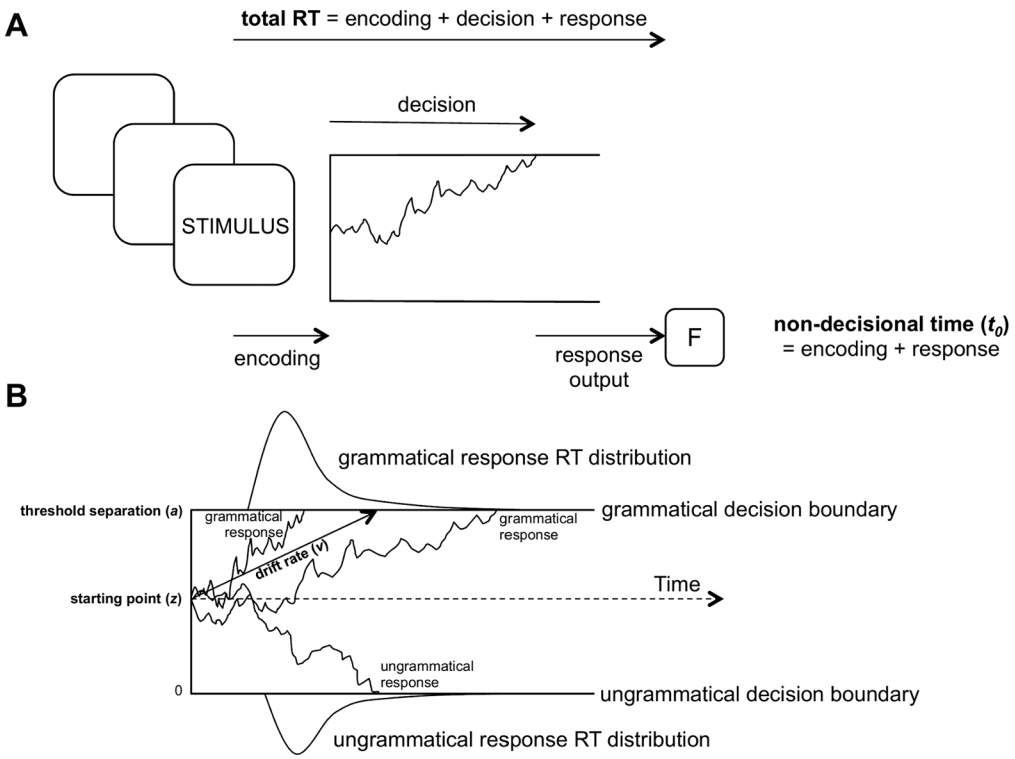

Fig. 1. Schematization of the Ratcliff drift diffusion decision model. The top panel (A) illustrates how the decision process is situated between two non-decisional processes (t_0): stimulus encoding and response output. The figure shows a possible path for a grammatical response. The bottom panel (B) shows the core decision process involving the drift rate (v), threshold separation (a) and starting point (z), which results in the characteristic right-tailed RT distributions.

upper boundary), with 0.5 being a starting point halfway between the two boundaries—a value that represents a lack of bias in the decision process. Shifts in bias lead to a higher proportion of responses, and faster responses, with respect to the relevant boundary. Taking the example shown in Fig. 1, if z_r had a value of 0.75, this would lead to significantly more and faster grammatical responses. Below we consider effects of bias, and how it interacts with other parameters, in more detail.

The final parameter, the drift rate (v), reflects the speed of evidence accumulation, and is modulated by the strength of the evidence provided by the stimulus. More informative or less equivocal information leads to higher drift rates, and less informative or more equivocal information leads to lower drift rates (e.g., Ratcliff & Smith, 2004). Changes to the drift rate have a number of important effects on speed and accuracy. Higher drift rates result in more consistent and faster responses towards the upper boundary, whereas drift rates approaching zero result in more erroneous responses towards the lower boundary, and slower response times overall (e.g. Ratcliff & McKoon, 2008).

5.1. Application to agreement attraction

Following Staub (2009), we advance the linking hypothesis that the continuous number value in the Marking and Morphing model can be mapped to the drift rate of a diffusion process. Here we extend this idea to model grammaticality judgments, rather than selection of the appropriate verb form. We hypothesize that in the attractor mismatch conditions, where number marking of the subject is somewhat ambiguous between singular and plural, the drift rate toward a decision boundary is reduced. In production, this plausibly leads to production errors in the number marking on the verb. In grammaticality judgments, this should lead to errors in computing the grammaticality of the input in both grammatical and ungrammatical sentences (i.e. grammatically symmetrical agreement attraction).

To confirm the basic accuracy predictions of the drift diffusion model where *S(r)* shifts the drift rate parameter, we recovered response data using the *construct-samples* function of *fast*-dm (Voss & Voss, 2007). The code to run this model can be found on the OSF page for the study. We left all parameters at default values, except the drift rate, which we lowered in the mismatch conditions ($v = 1.5$) compared to the match conditions ($v = 3$) to simulate the effect of ambiguous number marking predicted by a continuous valuation model.[2] The results are shown in Table 2 in the "No Bias" rows. As expected, the model shows fully symmetrical effects of

[2] Positive drift rates were used to simulate grammatical sentences, and negative drift rates for ungrammatical sentences, as the model coded grammatical responses on the upper boundary and ungrammatical responses on the lower boundary. We also simulated models where the difference

Table 2
Mean accuracy and correct RT by condition and bias for the drift diffusion model simulation. The difference in drift rate was held constant across the different bias conditions. No Bias had a relative starting point of 0.5, Low "Yes" Bias had a relative starting point of 0.6 towards the yes/grammatical boundary, and High "Yes" Bias had a relative starting point of 0.7 towards the yes/grammatical boundary.

		Grammatical		Ungrammatical	
		Match	Mismatch	Match	Mismatch
No Bias	Accuracy	0.95	0.82	0.95	0.82
	RT (ms)	451	512	451	512
Low "Yes" bias	Accuracy	0.98	0.88	0.91	0.74
	RT (ms)	423	478	475	540
High "Yes" bias	Accuracy	0.99	0.92	0.84	0.62
	RT (ms)	395	439	495	563

attractor number in grammatical and ungrammatical sentences, with an estimated effect of 0.13 in each. The relative values of RT show that responses to mismatch stimuli are slower than responses to match stimuli, and that this difference is the same in grammatical and ungrammatical conditions.

We then simulated two models with an identical shift in drift rate between the match and mismatch conditions, but with bias towards a grammatical response—one with a more modest shift where $z_r = 0.6$, and one with a larger shift where $z_r = 0.7$. Again, the results are shown in Table 2. This stepwise change in bias is associated with an increase in the size of an interactive pattern in accuracy and RT. Taking the High "Yes" Bias case as an example, with no change to the assumptions about how continuous number valuation affects drift rate, we see that there is a small effect of attractor number in the grammatical conditions ($=0.07$), while there is a large effect in the ungrammatical conditions ($=0.22$), and an overall decrease in accuracy in the ungrammatical match condition ($=0.84$) compared to the grammatical match ($=0.98$). This interactive pattern is due to the fact that a change in drift rate has less of an effect on accuracy when the starting point is shifted towards one boundary or another, as there is less of a chance that the evidence accumulator will be able to reach the opposite threshold. Similarly, the shift away from the ungrammatical boundary makes it less likely overall to reach an ungrammatical response, even in the match conditions.

In RT, the expected pattern of match being faster than mismatch is again observed. However, it is now also the case that grammatical responses are faster than ungrammatical ones. This is due to the geometric consequences of the shift in bias: starting closer to the correct boundary leads to overall faster responses as less evidence (and therefore less time) is needed to reach that threshold. This change in the geometry is also responsible for the magnitude of the difference between the match and mismatch conditions being larger in ungrammatical sentences (68 ms) than grammatical ones (44 ms) when bias is present (compared to the symmetrical 61 ms difference in the No Bias condition).

6. The present study

Our simulation reveals that the grammaticality asymmetry can in fact be predicted by representational models of attraction when they are situated in an explicit decision model. However, on this view, the asymmetry will only arise in the presence of response bias. Building on this observation, the present study examines the effect of response bias on both response accuracy and latency in a binary decision task, in order to adjudicate between cue-based models of agreement attraction and continuous valuation models. The cue-based model predicts that asymmetrical agreement attraction should arise even in the absence of response bias. While response bias, if present, could in principle have an effect on response accuracy and RT under a cue-based model, when effects of bias are neutralized or partialed out, asymmetrical attraction effects should remain. On the other hand, a continuous number valuation model of agreement predicts the asymmetry only in the presence of a bias towards grammatical responses, with the underlying pattern of symmetrical attraction emerging when bias is neutralized. The drift diffusion implementation of the continuous valuation model further predicts that RT with mismatching attractors should be slower than with matching attractors, due to a relative decrease in the rate of evidence accumulation; this main effect should be present regardless of bias, but should interact with bias causing a more pronounced effect of mismatch on ungrammatical sentences when bias is present. Finally, the model also predicts that in the presence of bias towards grammatical responses, RT in the grammatical condition should be faster than in the ungrammatical condition. When bias is neutralized, RT in the grammatical and ungrammatical conditions should be identical.

In three experiments, we use filler item composition and instruction manipulation to shift response bias. In Experiment 1, we provide a conceptual replication of the binary judgment task in Experiment 7 of Wagers et al. (2009), finding the same bias towards grammatical responses and the expected asymmetry. In Experiments 2 and 3, we manipulate response bias by increasing the number of ungrammatical fillers, and informing participants during the task instructions to expect either 2/3 ungrammatical sentences (Experiment 2) or a majority ungrammatical sentences (Experiment 3), finding a stepwise decrease in bias corresponding to the dissipation of the grammaticality asymmetry. In all experiments, we find that RT is slower in the mismatch compared to the match

(*footnote continued*)
in drift rate between the match and mismatch conditions was smaller or larger. We found the same qualitative patterns reported here, with the only difference being in the magnitude of the effect.

conditions, and shifts in the difference in RT between grammatical and ungrammatical conditions as a function of bias. Using the *fast-dm* procedure (Voss & Voss, 2007), we then fit a diffusion model to the data from these experiments. The results support the view that an interaction between drift rate and starting point is indeed responsible for the differences in the grammaticality asymmetry observed across the three experiments.

7. Experiment 1

Previous studies of agreement attraction in comprehension have observed the grammaticality asymmetry in forced-choice acceptability judgment tasks (e.g. Wagers et al., 2009). The goal of Experiment 1 is to provide a conceptual replication of these studies to form a baseline for the subsequent experiments, which explicitly attempt to manipulate participants' response bias. The experiment has the further goal of measuring reaction times for speeded judgments in order to test two predictions of the drift diffusion model. First, RT in mismatch conditions should be slower than match. Second, in the presence of asymmetrical attraction, RT should be faster in grammatical conditions than ungrammatical. The preregistration of hypotheses, methods, and analysis can be found here: https://osf.io/q8anc/register/565fb3678c5e4a66b5582f67.

7.1. Participants

Data from 43 participants were collected, with three being excluded due to reporting a language other than English to be their native language. Therefore 40 participants were included in the analysis. In all experiments reported here, the participants were undergraduates in either Linguistics or Psychological and Brain Sciences at the University of Massachusetts Amherst, who were compensated with course credit and provided informed consent. Of the included participants, 26 identified as female, and 14 as male. The average age was 19.8 years, and ages ranged between 18 and 30 years. No participant took part in more than one of the experiments.

7.2. Materials and design

All three experiments used the same set of 60 experimental items (excepting changes to the items due to a verb-type manipulation in Experiment 1 discussed further below), and the same set of fillers (modulo modification of the fillers to change the number of subject-verb agreement errors in Experiments 2 and 3). Abstracting away from a between-subjects manipulation of verb-type in Experiment 1, the experimental items were arranged in a 2 × 2 design, with grammaticality (grammatical/ungrammatical) and attractor number (match/mismatch) as factors. An example set is given in Table 3. All experimental items follow the same form of *determiner-noun-preposition-determiner-noun-adverb-verb*. The first noun was always a definite singular subject. The second noun was the attractor, which was embedded in a prepositional phrase, and alternated between a singular and plural form according to the factorial design. An adverb was used to create a buffer between the complex subject and the critical verb to mitigate against known processing effects that occur following plural nouns (see Wagers et al., 2009).

The critical verb was always the final word of the sentence, and provided either a grammatical or ungrammatical continuation. In Experiment 1, an additional between-subjects manipulation of verb type was employed such that half of the participants (n = 20) saw only lexical verbs (e.g. *rusts/rust*) in which the stimulus could form a complete sentence, and the other half (n = 20) saw incomplete sentences that ended in one of three counterbalanced auxiliary verbs (*has/have, was/were, is/are*). The manipulation was included in Experiment 1 to provide a comparison to Staub (2009), where the forced-choice completion task utilized auxiliary verbs. Past work examining differences in verb type in comprehension has not recovered any apparent differences in the profile of agreement attraction (Lago et al., 2015), therefore we did not expect any differences in the current study. This expectation was largely confirmed, as detailed in the results section below. Therefore in Experiments 2 and 3, the between-subjects manipulation was dropped, and only lexical verbs were used.

The 60 experimental stimuli were combined with 80 fillers, resulting in 140 items in total. The full list of fillers can be found in Appendix B. In Experiment 1, 50% of the fillers were ungrammatical, resulting in 50% of verbs in the experiment as a whole being ungrammatical. All filler items ended in an agreeing verb that resulted in either a grammatical or ungrammatical dependency. Three

Table 3
Example experimental stimuli in each condition of the grammaticality by attractor number manipulation. Experiment 1 included a by-subjects manipulation of verb type (lexical vs. auxiliary). Experiments 2 and 3 included only lexical verb items.

Experiments 1–3: Lexical verbs		
Grammatical	Match	The friend of the nurse frequently visits
	Mismatch	The friend of the nurses frequently visits
Ungrammatical	Match	The friend of the nurse frequently visit
	Mismatch	The friend of the nurses frequently visit
Experiment 1: Auxiliary verbs		
Grammatical	Match	The friend of the nurse frequently is/was/has
	Mismatch	The friend of the nurses frequently is/was/has
Ungrammatical	Match	The friend of the nurse frequently are/were/have
	Mismatch	The friend of the nurses frequently are/were/have

types of structures, counterbalanced for grammaticality, were used in the fillers: coordinated DP subjects (e.g. *the dog and the cat always play*), object relative clauses (e.g. *the camera filmed the line that the runners quickly cross*), and complement clauses (e.g. *the pilot thinks that the passengers usually sleep*). These structures provided two crucial controls. First, they ensured that the first noun was not uniformly the controller of agreement for the final verb. Second, they ensured that the experiment included grammatical plural verbs, and ungrammatical singular verbs. This is important, as the experimental stimuli alone confound ungrammaticality with verb plurality—a confound that is present in much of the literature, as attraction effects are most often tested with singular head nouns due to the mismatch asymmetry. Given these controls, participants could not pay attention to only the first noun and the final verb and reach the correct response, nor could they adopt a strategy of responding "ungrammatical" to plural verbs and "grammatical" to singular. Participants had to actively parse the sentences in order to arrive at the correct response.

7.3. Procedure

The experiments were conducted in a lab with up to four participants at a time, where each participant was comfortably seated at their own keyboard and monitor. Participants were arranged such that they could not see the screens or keyboards of the other participants. The experiment was presented to each participant on an iMac desktop running the PsychoPy experimental software (Peirce, 2007).

Sentences were presented using word-by-word RSVP on a grey background. Each trial consisted of the following sequence. The trial was initiated with a fixation cross that appeared on the screen for 1.5 s. Participants were instructed to look at the cross to prepare for the oncoming sentence. Following the fixation cross, a 100 ms blank screen was presented, followed by the presentation of the sentence. The sentence was presented one word at a time at a rate of 225 ms per word with a 100 ms interstimulus interval (ISI), for a total stimulus onset asynchrony (SOA) of 325 ms. Each word was displayed at the center of the screen in white font, except for the final word of the sentence (in all cases the critical verb), which was green. Participants were instructed to make a yes/no grammaticality judgment using the F and J keys when they saw the green word. F was uniformly used to indicate the sentence was grammatical, and J to indicate the sentence was ungrammatical. The verb remained on the screen until participants gave a response. If participants failed to respond within 3 s of the presentation of the green word, the words "TOO SLOW" appeared on the screen in bold red letters. No feedback about the accuracy of responses was given.

Prior to beginning the experiment, participants were given oral instructions by the experimenter. Participants were instructed that their task was to read and judge sentences of English. The instructions informed participants of the limited response deadline, and emphasized using gut reactions rather than deliberated responses based on mental rehearsal of the stimulus. Following the instructions, participants completed 6 practice trials that matched the proportion of ungrammatical sentences present in the experiment. After the experimental trials were completed, participants completed a survey that probed for response strategies. In Experiments 2 and 3, we additionally asked participants to report what percentage of the sentences they thought were ungrammatical. In total, the session lasted no more than 30 min, including the obtaining of consent, instructions, practice, the experimental task, the post-experiment survey, and debriefing.

7.4. Results

Trials where participants failed to respond within the 3 s response deadline were excluded from the analysis. Participants had little trouble responding before the deadline, as only 0.33% of trials were excluded for this reason. All other trials were included in the analysis of both accuracy and response time.

For both accuracy and response time data, we calculated likelihood ratio tests to compare the fits of the logistic mixed effects models (in the case of accuracy) and linear mixed effects models (in the case of raw RT and log-transformed RT on correct responses, both of which were specified in the preregistration) that either included or excluded the fixed effect of verb type, i.e. auxiliary vs. lexical. In the models of accuracy and raw RT, the likelihood ratio tests were not significant ($p > 0.05$). However the test on log-transformed RT was significant ($p = 0.008$). Inspection of the verb-type model revealed that while the qualitative patterns of RT were the same across verb-type in the four conditions, the effect of attractor mismatch in the ungrammatical conditions was greater in the auxiliary verbs compared to the lexical verbs. This interaction does not contradict any of the core patterns expected on the current hypotheses, namely that the mismatch conditions should show slower RT than the match. This was true in both the lexical and auxiliary conditions. Given this fact, and the finding of non-significant differences in the accuracy and the raw RT tests, we only report results for the models that exclude verb type. Results for the models including verb type, and the code to calculate the likelihood ratio test, are included in the analysis script in the supplementary materials.

For all three experiments, mixed effects models including both grammaticality and attractor number as fixed effects, and subject and item random slopes and intercepts corresponding to each of these fixed effects, were separately fit to the accuracy and raw RT data. A logistic model was fit to the accuracy data, and a linear model to the raw RT data for correct responses only. In both cases, effects coding was used for the fixed effects (mismatch = 0.5, match = −0.5; grammatical = 0.5, ungrammatical = −0.5). For the logistic model, an inference criterion of $p < 0.05$ was used; for the linear model, an inference criterion of $t > 2.0$ was preregistered. We additionally include a *p*-value derived from the *lmerTest* package in R (Kuznetsova, Brockhoff, & Christensen, 2017), based on the Satterthwaite approximation for the denominator degrees of freedom, though this test was not preregistered. The inferences from the $t > 2.0$ and $p < 0.05$ criteria are identical in all cases.

Table 4
Mean and by-subjects standard error for accuracy and RT in Experiment 1.

	Grammatical Match	Mismatch	Ungrammatical Match	Mismatch
Accuracy	0.89 (0.02)	0.86 (0.02)	0.80 (0.03)	0.60 (0.04)
RT (ms)	926 (33)	970 (36)	1022 (34)	1140 (54)

7.4.1. Accuracy

Mean accuracy and by-subjects SEM were calculated for each of the four experimental conditions, and are shown in Table 4, and in the left-most panel of Fig. 2. As expected, the means reveal an interactive effect characteristic of the grammaticality asymmetry: The effect of attractor number on accuracy is larger in the ungrammatical conditions than the grammatical conditions. An additional effect is apparent by comparing the grammatical and ungrammatical match conditions: Independent of the effect of a mismatching attractor in the ungrammatical conditions, the ungrammatical conditions show decreased accuracy compared to the grammatical ones. In other words, participants are more likely overall to respond "grammatical" than "ungrammatical"; this difference is one of the indicators that bias towards grammatical responses is present.

Results of the logistic mixed effects model are given in Table 5. Both main effects of attractor number and grammaticality were significant, as was the interaction between these two factors ($p < 0.001$ in all cases). The interaction provides support for the hypothesis that the effect of attractor number is asymmetrical: A mismatching attractor has a bigger effect in the ungrammatical conditions than in the grammatical ones.

To support this interpretation of the interaction, we ran *post hoc* (i.e. not preregistered) traditional *t*-tests, as well as Bayesian *t*-tests (Rouder, Speckman, Sun, Morey, & Iverson, 2009) using the Bayes Factor package in R (Morey & Rouder, 2015). In Experiments 2 and 3, these tests were preregistered. In contrast to *p*-values in null hypothesis significant testing, which cannot distinguish whether the failure to observe an effect is due to factors such as lack of power or if it is due to actual invariance, Bayes Factors provide odds that can allow us to evaluate the strength of evidence either in favor of a difference or against one. For example, in a one-sample Bayesian *t*-test, a Bayes Factor of 0.1 can be interpreted as ten-to-one odds in favor the null hypothesis (i.e. the mean is *not* different from zero). Conversely, a Bayes Factor of 10 is interpreted as ten-to-one odds in favor the alternative hypothesis (i.e. that the mean *is* different from zero). Unlike *p*-values, there is no significance cutoff, however there are values that are conventionally related to degrees of certainty. Intuitively, 1:1 odds are considered entirely equivocal. Odds of < 3:1 provide weak evidence, between 3:1 and 10:1 substantial evidence, between 10:1 and 100:1 strong evidence, and over 100:1 decisive evidence (Jeffreys, 1961). We transform all Bayes Factors so that they are greater than or equal to 1, in each case explaining which hypothesis is favored.

Tests on the difference score between the mismatch and match conditions in ungrammatical sentences revealed a significant effect ($t(39) = 6.88$, $p < 0.001$) with a Bayes Factor of 439,965 in favor of the hypothesis that there is an effect. In contrast, the test of the grammatical sentences was not significant ($t(39) = 1.56$, $p = 0.127$) with a Bayes Factor of 1.93 in favor of the null hypothesis. The tests therefore reveal the expectedly decisive evidence in favor of an effect of attractor number in the ungrammatical conditions, with weak evidence in favor of a null effect in grammatical conditions.

7.4.2. Reaction time

Mean raw RT and by-subjects SEM for correct responses were calculated for each of the four experimental conditions, and are shown in Table 4. As specified in the preregistration, parallel tests were run on log-transformed RT. The results did not differ, in terms of patterns of significance, so we only report the findings for raw RT. The results are shown in the left-most panel of Fig. 3. The means reveal an overall slowdown for ungrammatical compared to grammatical responses, and a slowdown in the mismatch compared to

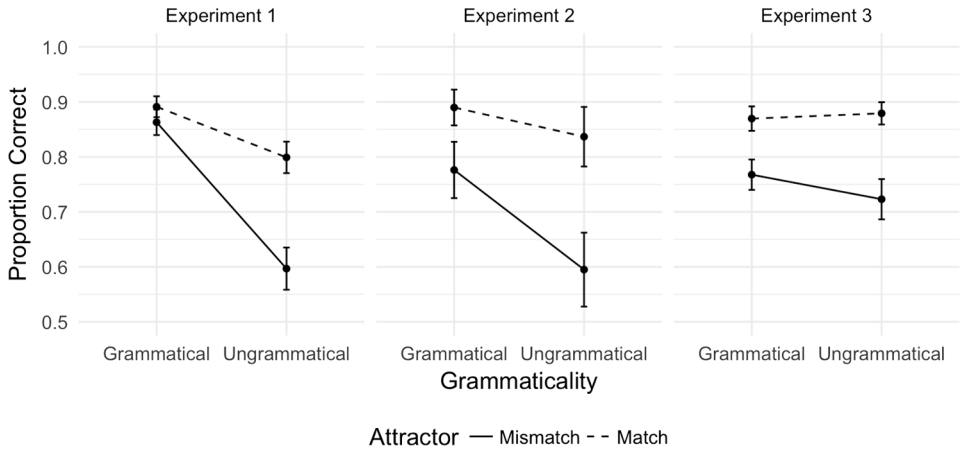

Fig. 2. Interaction plot of mean and by-subjects SEM for each condition in Experiments 1–3.

Table 5
Results of logistic mixed effects model on accuracy (top) and linear mixed effects models on RT (bottom) for Experiment 1.

Parameter	Estimate	Standard error	z-value	p-value
Grammaticality	1.28	0.20	6.48	< 0.001
Attractor	0.14	0.14	−5.31	< 0.001
Grammaticality × Attractor	0.24	0.24	3.86	< 0.001
Parameter	**Estimate**	**Standard error**	**t-value**	**p-value**
Grammaticality	−143.38	21.40	−6.70	< 0.001
Attractor	88.15	23.05	3.82	< 0.001
Grammaticality × Attractor	−91.30	35.25	−2.59	0.013

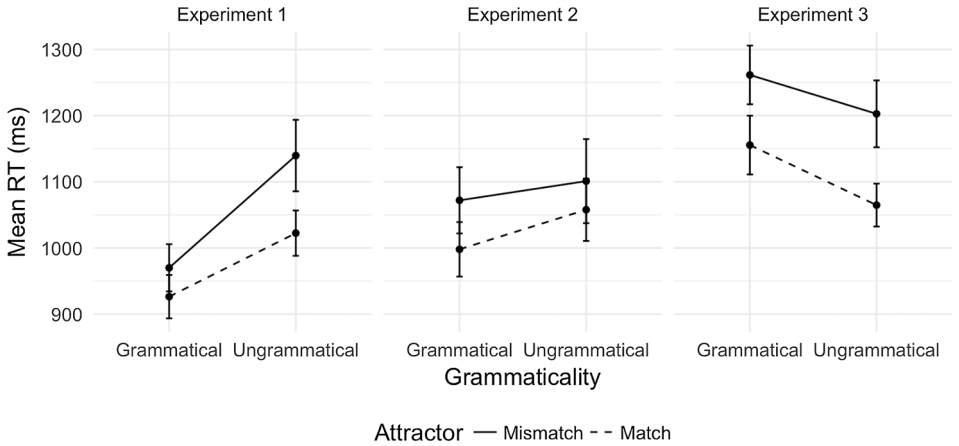

Fig. 3. Interaction plots of mean and by-subjects standard error for response time in Experiments 1–3.

the match conditions. Results of the linear mixed effects model are given in Table 5. Both main effects of grammaticality ($t = -6.70$; $p < 0.001$) and attractor number ($t = 3.82$; $p < 0.001$) were significant, as was the interaction between these two factors ($t = -2.59$; $p = 0.013$), with the effect of a mismatching attractor on correct RT being larger for ungrammatical sentences than grammatical ones.

As with accuracy results, *post hoc* traditional and Bayesian *t*-tests on the difference between the mismatch and match conditions were run to clarify the interpretation of these results. The test revealed a significant effect in grammatical sentences ($t(39) = -2.42$, $p = 0.020$) with a Bayes Factor of 2.24 in favor of the hypothesis that there is an effect. In ungrammatical sentences, the traditional *t*-test was also significant ($t(39) = -2.89$, $p = 0.006$) with a Bayes Factor of 6.09 in favor of the hypothesis that there is a difference.

7.4.3. Response bias

We calculated the signal detection bias measure *c* (Macmillan & Creelman, 1991) for each of the participants, and found the mean bias and by-subjects standard error for the experiment as a whole. (Note that this analysis was not included in our preregistration for Experiment 1.) The formula for *c* is given in Eq. (3). For each participant, we took the mean accuracy in the grammatical condition as the hit rate (a correct grammatical response), and one minus the mean accuracy in the ungrammatical condition as the false alarm rate (an incorrect grammatical response). When $c = 0$, the responder is said to be unbiased. More negative values of *c* indicate greater bias towards grammatical responses, and more positive values of *c* indicate greater bias towards ungrammatical responses.

$$c = -\frac{Z(Hit\ \ Rate) + Z(False\ \ Alarm\ \ Rate)}{2} \tag{3}$$

As shown in Fig. 4, participants as a whole showed significant bias towards grammatical responses (i.e. a negative value of *c*). A *t*-test revealed that bias in Experiment 1 differed significantly from zero ($t(39) = -7.24, p < 0.001$), where zero indicates that no bias was present. A Bayesian *t*-test on the same comparison results in a Bayes Factor of 1,262,208 in favor of the hypothesis that bias differs significantly from zero. In the results of the subsequent experiments, we provide preregistered statistical tests that reveal how bias shifted across the three experiments.

7.5. Discussion

The accuracy results replicate the general finding in the literature: The significant interaction between grammaticality and attractor number indicates an asymmetry in the degree of attraction in grammatical and ungrammatical sentences, such that

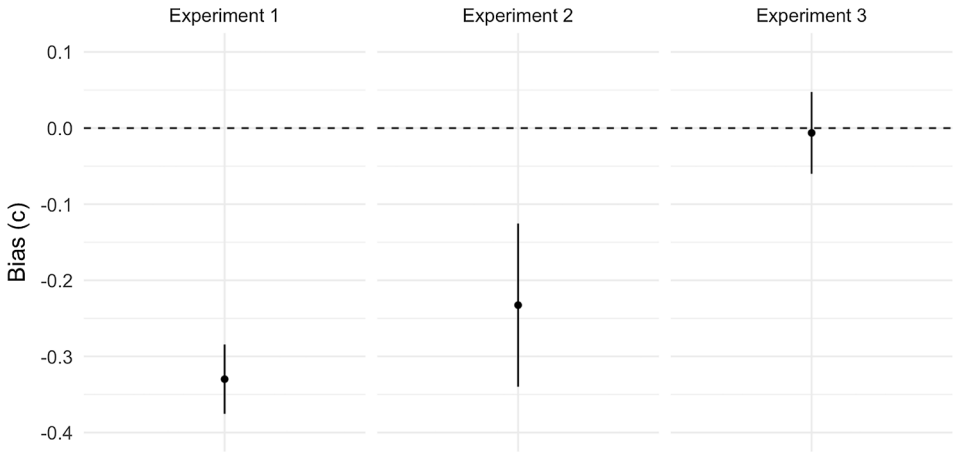

Fig. 4. Mean and by-subject standard error for bias in Experiments 1–3. Bias was calculated for each participant using the signal detection measure *c*, shown in Eq. (3).

ungrammatical sentences show significantly more attraction than grammatical ones. Furthermore, the analysis of reaction time reveals that judgments in the mismatch conditions were slower than the match. This extends the finding of Staub (2009), who showed that verb-choice decisions in a forced-choice sentence continuation task were slowed in mismatch compared to the match conditions. Additionally, the models revealed a main effect of grammaticality, such that judgments in ungrammatical conditions were made more slowly than in grammatical conditions. Lastly, we observed an interaction between grammaticality and attractor number, suggesting that the slowing in the mismatch condition compared to the match condition was more pronounced in ungrammatical sentences. An interaction with this shape is expected based on the diffusion model simulations presented in the introduction, where the High "Yes" Bias condition showed a larger mismatch effect on RT in ungrammatical conditions.

Overall, the results are consistent with a decision process where mismatching attractors decrease the rate of evidence accumulation, and participants have a bias towards grammatical responses. The previously discussed diffusion model simulations show that decreased drift rate in the mismatch conditions predicts slowed RT in comparison to match conditions, and that adding a bias towards grammatical responses can create an empirical interaction between mismatch and grammaticality in the accuracy measure, and can increase RT in the ungrammatical conditions and produce an interactive effect. These effects emerge due to the geometry produced by shifts in the decisional starting point within the drift diffusion model: Starting closer to the grammatical boundary means that it takes longer, on average, to reach the ungrammatical boundary, and that a change in the drift rate that is similar in size in both the grammatical and ungrammatical conditions will have a reduced effect on accuracy in the grammatical conditions. As expected on this view, a direct test of response bias using the signal detection measure *c* showed a pronounced bias towards grammatical responses.

The drift diffusion model account of these results predicts a dependency between response bias and the grammaticality asymmetry. This raises the question of whether reducing the grammatical response bias of participants would allow symmetrical effects of agreement attraction to emerge in the response accuracy measure, as expected under the Marking and Morphing model. In Experiments 2 and 3, we manipulate response bias in order to test this prediction.

8. Experiment 2

The goal of Experiment 2 was to change the overall response bias in the experiment to determine if symmetrical effects of agreement attraction emerge. Our link between the Marking and Morphing and drift diffusion models predicts that symmetrical agreement should appear when response bias is neutralized, or weaken as response bias shifts towards zero. On the other hand, an account that attributes agreement attraction to cue-based interference predicts that asymmetrical agreement attraction should remain when bias is neutralized. Our account also predicts that independently of bias, RT in the mismatch conditions should continue to be slower than the match conditions, on the hypothesis that equivocal number marking leads to a decrease in the rate of evidence accumulation. However, as bias shifts the difference between grammatical and ungrammatical RT should disappear, as the starting point of the decision process becomes equidistant from the decision boundaries.

There are many ways to manipulate bias in decision tasks, including payoff schemes (e.g. increasing reward for correctly making certain responses, and disincentivizing certain erroneous responses), or base rate variation, where the experimenter manipulates items to make a certain type of response more common overall. To shift the response bias of participants, two changes were made compared to the design of Experiment 1. First, we modified the base-rate of ungrammatical responses by adapting fillers so that approximately 2/3 of the total items were ungrammatical, as compared to 50% in Experiment 1. Second, participants were explicitly told to expect 2/3 of the sentences to be ungrammatical during the instruction period. That is, participants were informed of the base-

rate of responses—a base rate that was veridically reflected in the materials (see Bröder & Malejka, 2017, for explanation of why giving veridical information about base rates is important[3]). Manipulations of this type are common in the field of decision-making, and have been previously shown to significantly shift response bias in a wide variety of psychological domains (e.g. Healy & Jones, 1975; Ratcliff, Sheu, & Gronlund, 1992; Rhodes & Jacoby, 2007). The preregistration of the hypotheses, method, and analyses is available here: https://osf.io/fbt4m/register/565fb3678c5e4a66b5582f67.

8.1. Participants

Data from 22 participants were collected. Two participants were excluded due to reporting a language other than English as their native language, leaving 20 participants included in the analysis. Of the included participants, 13 identified as female, and 7 as male. The average age was 19.9 years, and ages ranged from 18 to 24 years.[4]

8.2. Materials and design

The experimental items were identical to the lexical verb condition of Experiment 1. In addition, the fillers were modified such that 64% of the overall items in the experiment were ungrammatical. To accomplish this, 60 of the 80 fillers were made ungrammatical. This was done in a way that maintained the counterbalancing described in the procedure for Experiment 1.

8.3. Procedure

The procedure was identical to that of Experiment 1, except for changes to the instructions and the composition of the practice and filler trials. In addition to the instructions previously described, participants were told to expect 2/3 of the sentences in the experiment to be ungrammatical, and the practice items and fillers were adapted to result in 2/3 of the sentences being ungrammatical.

8.4. Results

Trials where participants failed to respond within the 3 s response deadline were excluded from the analysis. Again, participants had no trouble responding before the deadline, as only 0.25% of trials were excluded for this reason. All other trials were included in the analysis of both accuracy and response time.

8.4.1. Accuracy

Mean accuracy and by-subjects SEM were calculated for each of the four experimental conditions, and are shown in Table 6, and in the center panel of Fig. 2. The patterns reveal a marked shift from the results obtained in Experiment 1, as there is an increase in size of the effect of attractor number in grammatical sentences. Notably, however, the effect in ungrammatical sentences still appears to be larger. In addition, there is an attenuation of the overall decrease in accuracy in grammatical conditions compared to ungrammatical conditions, suggesting that bias may have been partially mitigated by the instruction manipulation.

Results of the logistic mixed effects model are given in Table 7. The analysis reveals a significant main effect of attractor number ($p < 0.001$) and a marginally significant effect of grammaticality ($p = 0.068$), qualified by a significant interaction between these two factors ($p = 0.033$). Planned contrasts using traditional and Bayesian t-tests on the match minus mismatch difference scores reveal a significant difference in both grammatical sentences ($t(19) = 3.51$, $p = 0.002$; BF = 17.50 in favor of the alternative hypothesis that the difference is not zero) and ungrammatical sentences ($t(19) = 5.56$, $p < 0.001$; BF = 1036.67 in favor of the alternative). Thus while the effect of attractor number was larger in the ungrammatical conditions, as evidenced by the significant interaction, an effect of mismatch was reliably present in the grammatical conditions as well.

8.4.2. Reaction time

Mean accuracy and by-subjects SEM were calculated for each of the four experimental conditions raw RT for correct responses only, and are shown in Table 6 and in the center panel of Fig. 3. Again, parallel tests were run on log-transformed RT. Since the results did not differ, we only report the findings for raw RT. The results show that mismatching attractors slowed reaction time compared to

[3] We note here an additional preregistered experiment within the OSF project page, not presented here, where we manipulated only the base-rate of ungrammatical sentences by making 100% of the fillers ungrammatical. The preregistration can be found here: https://osf.io/h2qru/register/565fb3678c5e4a66b5582f67, and the data can also be found on the OSF page for the study. Overall, the results of the experiment are consistent with smaller shifts in bias decreasing the grammaticality asymmetry, and is thus redundant with the results presented here. However, the manipulation lent itself to independent response strategies, as only the critical PP modifier construction was grammatical. Therefore we do not report the results in the present article.

[4] We acknowledge here that the logic motivating our N across the three experiments was not clear in the preregistrations, and would like to explain our decisions in more detail for the purpose of clarity. After running E1, we matched the N of E2 to the sample size of the between-subjects group (N = 20) rather than the full N. Prior to running E3 (presented below) we decided to increase our N to 40—the overall N of Experiment 1. This was done to match power and to make cross-experimental comparisons easier. That said, we would like to underscore that all sample sizes were established prior to collection of data and analysis—we take this to be the primary function of preregistering sample size.

Table 6
Mean and by-subjects standard error for accuracy and RT in Experiment 2.

	Grammatical Match	Mismatch	Ungrammatical Match	Mismatch
Accuracy	0.89 (0.03)	0.78 (0.05)	0.84 (0.05)	0.60 (0.07)
RT (ms)	998 (41)	1072 (50)	1058 (47)	1101 (64)

Table 7
Results of logistic mixed effects model on accuracy (top) and linear mixed effects models on RT (bottom) for Experiment 2.

Parameter	Estimate	Standard error	z-value	p-value
Grammaticality	0.94	0.52	1.83	0.068
Attractor	−1.56	0.24	−6.57	< 0.001
Grammaticality × Attractor	0.84	0.39	2.14	0.034
Parameter	Estimate	Standard error	t-value	p-value
Grammaticality	−46.82	37.44	−1.25	0.228
Attractor	61.18	19.94	3.07	0.002
Grammaticality × Attractor	13.53	43.84	0.31	0.759

matching ones, independently of grammaticality. The linear mixed effects model, given in Table 7, confirms that only the main effect of attractor number reached significance in the model ($t = 3.07$; $p = 0.002$).

Post hoc traditional and Bayesian *t*-tests on the difference between the mismatch and match conditions were run to support the interpretation of the linear model. The tests revealed a significant effect in grammatical sentences ($t(19) = -2.64$, $p = 0.016$) with a Bayes Factor of 3.54 in favor of the hypothesis that there is an effect. In ungrammatical sentences, the traditional *t*-test reached failed to reach significance ($t(19) = -1.63$, $p = 0.119$), but with a Bayes Factor of 1.39 in favor of the hypothesis that there is an effect the result is equivocal with respect to whether an effect is present or absent.

8.4.3. Response bias

As in Experiment 1, the response bias measure *c* was calculated for each participant. The mean and by-subjects error are shown in the center panel of Fig. 4. While the mean bias appears to have shifted towards zero in comparison to Experiment 1, the shift is not drastic, and the standard error is high. A traditional and Bayesian *t*-test analysis reveals weak evidence to support the hypothesis that bias is significantly different from zero ($t(19) = -2.17$, $p = 0.043$; BF = 1.57 in favor of the alternative). A further Welch's *t*-test comparing the bias of Experiments 1 and 2 did not reach significance ($t(26.075) = 0.83$, $p = 0.41$), and the Bayesian analysis provides weak evidence in favor of there being no difference in bias between the two experiments (BF = 2.44 in favor of the null).

8.5. Discussion

The results of Experiment 2 suggest the bias manipulation was partially successful, as a small shift in bias did occur, and the expected effects on accuracy and RT were obtained. In contrast to Experiment 1, there was a significant effect of attractor number in both ungrammatical and grammatical sentences. While the effect was still shown to be bigger in ungrammatical sentences, the effect of attractor number on grammatical sentences is uniformly unexpected on current cue-based accounts. On the other hand, an account such as Marking and Morphing, which attributes agreement attraction to ambiguous number marking on the subject, can capture this effect. The modest shift in bias was enough to allow the effect of attractor number on grammatical sentences to emerge, but was not enough to result in fully symmetrical agreement attraction.

The RT data further support an account where continuous number marking is responsible for attraction, as the reduced rate of evidence accumulation in the mismatch compared to the match conditions results in increased RT. Furthermore, the shift in bias led to more comparable RTs in grammatical and ungrammatical conditions overall, resulting in a lack of main effect of grammaticality in Experiment 2. The interaction between grammaticality and attractor number that reached significance in Experiment 1 was not present in Experiment 2, which is also expected as bias is neutralized.

The results of Experiment 2 are promising, but fall short of confirming the critical prediction that fully symmetrical agreement attraction effects should be possible. This is not unexpected, however, given that while a shift in bias did occur in Experiment 2, bias was not entirely neutralized. One reason for this was that our instructional manipulation was not entirely successful. In particular, two of the participants reported in the post-experiment survey that they believed 1/3 of the sentences were ungrammatical, rather than 2/3. We believe this exposes a misunderstanding of the instructions. An exploratory analysis excluding these two participants suggests this misinterpretation may be responsible for the high standard error and the overall small shift in bias. Excluding these participants from the analysis results in an average *c* that is closer to zero (a shift from −0.23 to −0.14), and a smaller standard error (a shift from 0.107 to 0.094), despite the fact that this decreases the total number of participants (which otherwise would be expected to lead to an increased standard error, given its dependence on sample size). To address this, in Experiment 3, we changed the

instructions to be more comprehensible, and added an exclusion criterion based on the responses to the post-experiment survey, in order to see if it is possible to fully neutralize response bias and to reveal the symmetrical agreement attraction effects expected under Marking and Morphing.

9. Experiment 3

The goal of Experiment 3 was to make a second attempt at manipulating response bias to further test the hypothesis that the grammaticality asymmetry is due to bias towards grammatical responses. Our linkage of Marking and Morphing with the drift diffusion decision model predicts symmetrical agreement attraction only when response bias is fully neutralized, while current implementations of the cue-based retrieval models that rely on interference effects to drive agreement attraction do not predict symmetrical attraction.

To accomplish this shift in bias, we minimally changed the methods of Experiment 2. First, we preregistered exclusion criteria that allowed us to exclude participants whose responses on the post-experiment survey indicated that they failed to understand the instruction describing the grammaticality proportions. Second, we decided (post pre-registration, but prior to running the experiment itself), that instead of providing a quantitative instruction to expect 2/3 ungrammatical sentences, participants would be instructed simply to expect a majority of the sentences to be ungrammatical. The preregistration of all hypotheses, methods, and analyses can be found here: https://osf.io/92m6u/register/565fb3678c5e4a66b5582f67.

9.1. Participants

Data from 52 participants were collected for Experiment 3, with 40 participants being included in the final analysis. 10 participants were excluded for reporting a language other than English as their native language. One participant was excluded for reporting in the post-experiment survey that they thought less than 50% of the items were ungrammatical, indicating they did not understand the instructions. Finally, one participant was excluded due to filler accuracy that was below 2.5 standard deviations from mean filler accuracy. Of the included participants, 14 identified as female and 26 as male. The average age was 20.0 years, and the ages ranged from 18 to 22 years.

9.2. Materials and design

All materials and design choices were identical to those in Experiment 2.

9.3. Procedure

The procedure was identical to Experiment 2, except participants were told to expect *the majority* of sentences to be ungrammatical, rather than 2/3 of sentences. This was done to ensure more straightforward comprehension of the instructions.

9.4. Results

Trials where participants failed to respond within the 3 s response deadline were excluded from the analysis. As in Experiments 1 and 2, participants had little trouble responding before the deadline: In total, only 0.83% of trials were excluded for this reason. All other trials were included in the analysis of both accuracy and response time.

9.4.1. Accuracy

Mean accuracy and by-subjects SEM were calculated for each of the four experimental conditions, and are shown in Table 8 and in the right panel of Fig. 2. There are roughly equal effects of attractor number in both the grammatical and ungrammatical conditions, and no general decrease in accuracy in the ungrammatical conditions compared to the grammatical conditions.

Results of the logistic mixed effects model are given in Table 9. Only the main effect of attractor number reached significance ($p < 0.001$). In contrast to Experiments 1 and 2, the interaction between grammaticality and attractor number characteristic of the grammaticality asymmetry was not significant ($p = 0.111$). Comparisons using traditional and Bayesian *t*-tests on the match minus mismatch difference scores were significant for both grammatical sentences ($t(39) = 4.46$, $p < 0.001$; BF = 346) and ungrammatical sentences ($t(39) = 5.77$, $p < 0.001$; BF = 16,053), providing further evidence that effects of attractor number occurred in both grammatical and ungrammatical conditions.

Table 8
Mean and by-subjects standard error for accuracy and RT in Experiment 3.

	Grammatical		Ungrammatical	
	Match	Mismatch	Match	Mismatch
Accuracy	0.87 (0.02)	0.77 (0.03)	0.88 (0.02)	0.72 (0.04)
RT (ms)	1156 (44)	1262 (44)	1065 (32)	1202 (50)

Table 9
Results of logistic mixed effects model on accuracy (top) and linear mixed effects models on RT (bottom) for Experiment 3.

Parameter	Estimate	Standard error	z-value	p-value
Grammaticality	0.05	0.26	0.19	0.849
Attractor	−1.06	0.15	−7.30	< 0.001
Grammaticality × Attractor	0.40	0.25	1.59	0.111
Parameter	**Estimate**	**Standard error**	**t-value**	**p-value**
Grammaticality	78.30	31.59	2.48	0.018
Attractor	122.34	18.06	6.77	< 0.001
Grammaticality × Attractor	−18.75	43.57	−0.43	0.670

9.4.2. Reaction time

Mean accuracy and by-subjects SEM were calculated for each of the four experimental conditions on raw RT for correct responses only, and are shown in Table 8 and the right panel of Fig. 3. The parallel tests on log-transformed RT produced the same results. As in the previous two experiments, there is a slowdown in RT in the mismatch conditions compared to the match conditions in both grammatical and ungrammatical sentences. Additionally, there is a slowdown in the grammatical conditions compared to the ungrammatical conditions, which is the opposite effect to that observed in Experiment 1. Results of the linear mixed effects model are given in Table 9. The analysis revealed significant effects of attractor number ($t = 6.77$; $p < 0.001$) and grammaticality ($t = 2.48$; $p = 0.018$), and no interaction between these factors.

Post hoc traditional and Bayesian *t*-tests on the difference between the mismatch and match conditions revealed a significant effect in grammatical sentences ($t(39) = -3.87$, $p < 0.001$) with a Bayes Factor of 68 in favor of the hypothesis that there is an effect. In ungrammatical sentences, the traditional *t*-test also reached significance ($t(39) = -4.27$, $p < 0.001$), with a Bayes Factor of 205 in favor of the hypothesis that there is an effect.

9.4.3. Response bias

The mean and by-subjects standard error for response bias is shown in the right panel of Fig. 4. The mean shows that bias is now essentially neutralized. Traditional and Bayesian *t*-tests provide evidence to support this conclusion. The traditional *t*-test did not reveal a significant effect ($t(39) = -0.12$, $p = 0.91$) and the Bayesian analysis suggests that there are 5.82:1 odds in favor of the null hypothesis (i.e. that there was no response bias present).

Pairwise two-sample *t*-tests comparing the bias in Experiment 3 to Experiments 1 and 2 were significant for the comparison of Experiments 1 and 3 ($t(75.97) = 4.59$, $p < 0.001$; BF = 1086 in favor of a difference), and marginally significant for the comparison of Experiments 2 and 3 ($t(28.87) = 1.89$, $p = 0.069$; BF = 1.69 in favor of a difference). Experiment 3 therefore shows a marked shift in bias, particularly in comparison to Experiment 1. This shift occurred to the point that bias was statistically undetectable.

9.5. Discussion

Experiment 3 provided evidence that neutralizing response bias allows an underlying pattern of symmetrical agreement attraction to emerge. The relationship between bias and the appearance of the grammaticality asymmetry is predicted under the Marking and Morphing model, on the linking hypothesis that the indeterminate number marking in the mismatch conditions leads to a lower rate of evidence accumulation (i.e. lowers the drift rate). Support for the hypothesis is further bolstered by the finding that the mismatch conditions had uniformly slower RT than the match conditions—an effect predicted by lower drift rate. One surprising effect in Experiment 3 was the presence of slowed RT in the grammatical compared to the ungrammatical conditions. While such an effect is broadly predicted when response bias shifts towards *ungrammatical* responses, all else being held equal, it is not predicted if bias is simply neutralized, as appeared to be the case in the current experiment. We do not advance an interpretation of this effect at this time, though we do note that the drift diffusion model fits, to be presented in subsequent sections, can capture this effect.

Cue-based retrieval models predict that the underlyingly asymmetrical profile of agreement attraction should arise even when bias is neutralized. It further predicts that no agreement attraction should appear in grammatical sentences, as attraction effects are thought to arise due to cue-dependent interference effects. The appearance of symmetrical agreement attraction calls into question an account of agreement processing that relies on cue-based interference effects in the mismatch conditions, rather than representational changes to number marking on the subject phrase.

Before moving to the drift diffusion analysis, where we directly test the hypotheses that evidence accumulation rates and starting point bias can account for the accuracy and RT effects observed in the three experiments, we provide an omnibus analysis that includes all 100 participants. This analysis tests the relationship between response bias and the magnitude of the grammaticality asymmetry.

10. Omnibus analysis

We found that overall shifts in bias across the three experiments led to the neutralization of the grammaticality asymmetry in judgment accuracy. To provide a final test of the hypothesis that neutralizing bias is responsible for the emergence of symmetrical

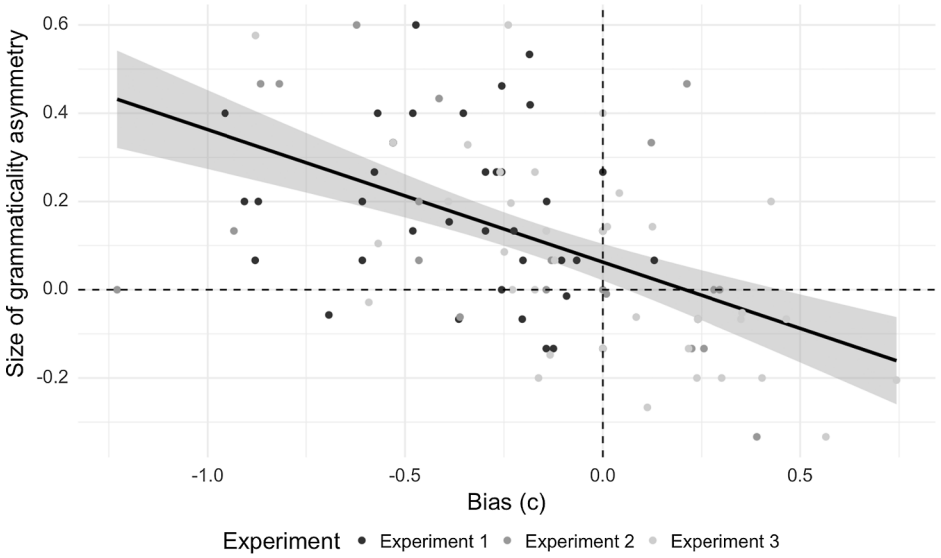

Fig. 5. Scatter plot and linear regression of the magnitude of the grammaticality asymmetry (the difference between the size of the attractor effect in ungrammatical versus grammatical sentences) against the signal detection bias measure *c*, for each participant. Shading of points indicates experimental groupings.

agreement attraction effects, we ran a logistic mixed-effects model with the data from all three experiments (N = 100). This analysis included bias as a continuous predictor, in addition to the fixed effects of grammaticality and attractor number and the random effects of item and subject. The analysis was preregistered as part of the additional planned analyses for Experiment 3. We predicted a three-way interaction between attractor number, grammaticality, and bias such that the size of the two-way interaction between attractor number and grammaticality decreases as bias approaches zero.

A visualization of the relationship between the magnitude of the grammaticality asymmetry and response bias is given in Fig. 5. Each point represents an individual subject. As response bias towards grammatical sentences increases (a more negative value on the x-axis), the grammaticality asymmetry in the canonical direction grows (a more positive value on the y-axis). As expected, the y-intercept of the linear regression line is near the origin: When there is no response bias, there is little or no grammaticality asymmetry. A number of participants who showed a bias towards *ungrammatical* responses showed an asymmetry in the opposite direction—that is, a larger effect of attraction in grammatical sentences than ungrammatical sentences. This is expected given a drift diffusion model of the decision making process, on the same logic that predicts an asymmetry in the presence of grammatical response bias.

The result of the logistic mixed effects model is given in Table 10. The model revealed a significant effect of attractor number ($p < 0.001$) and interactions between bias and grammaticality ($p < 0.001$) and grammaticality and attractor number ($p = 0.006$). These effects are qualified by a significant three-way interaction ($p = 0.031$) such that the interaction of grammaticality and attractor changes with bias. The results are consistent with the claim that the grammaticality asymmetry is an artifact of response bias: The interaction indicative of the grammaticality asymmetry and the current bias measure travel together. Only in the presence of response bias do asymmetrical attraction effects emerge. In the absence of bias, symmetrical agreement attraction appears.

11. Diffusion model analysis

The results of the three experiments suggest that a combination of response bias, as manipulated by the experimental instructions and filler items, and differences in the evidence strength between the match and mismatch conditions, are responsible for the accuracy and response time profile of agreement attraction illusions. In short, both decisional and language-related processes

Table 10
Results of logistic mixed effects models for omnibus analysis.

Parameter	Estimate	Standard error	z-value	p-value
Grammaticality	−0.04	0.15	−0.30	0.761
Attractor	−0.98	0.11	−8.63	< 0.001
Bias	0.44	0.28	1.57	0.116
Grammaticality × Attractor	0.48	0.18	2.76	0.006
Bias × Attractor	0.13	0.24	0.54	0.588
Bias × Grammaticality	−4.11	0.23	−17.52	< 0.001
Bias × Gramm. × Attract.	−0.95	0.44	−2.16	0.031

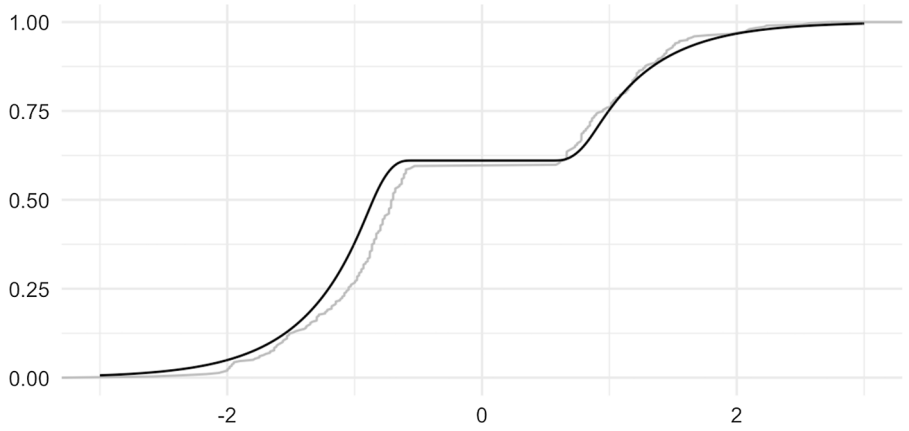

Fig. 6. Example of Kolmogorov-Smirnov approach used to optimize parameter fit in *fast-dm* (taken from the ungrammatical mismatch condition of Experiment 2). The otherwise separate grammatical and ungrammatical raw RT distributions are combined in a single CDF by multiplying RT in the ungrammatical threshold (i.e. the lower threshold) by −1. The figure shows the model estimation in the black line and the experimental data in the gray line. In the KS method, the parameter space is searched in order to minimize the maximal vertical distance between the two CDFs.

contribute to the observed results in a judgment experiment. Diffusion model analyses allow for the recovery of parameters that correspond to the cognitive variables of interest that underlie the decision process, allowing the underlying language-related effects to emerge with more clarity. In this section we fit the diffusion model to the results of the three experiments, recovering estimates of parameters that allow us to more directly test these hypotheses about the source of agreement attraction.

There are two critical hypotheses that stem from the current account. First, the drift rate is a function of the equivocality of number marking: In the attractor mismatch conditions number marking is ambiguous, which lowers the drift rate in comparison to the match conditions. This should occur across all experiments, and independently of the effect of grammaticality. Second, the instruction manipulation is hypothesized to affect the position of the decisional starting point. In particular, we expect a significant shift in the starting point away from the grammatical response threshold as we move from Experiment 1, where we observed a strong bias towards grammatical responses, to Experiment 3, where bias appeared to be neutralized.

11.1. The fast-dm procedure

To fit a diffusion model to the results of the three experiments, we used the Kolmogorov-Smirnov (KS) optimization procedure as implemented in *fast-dm*, originally described in detail by Voss and Voss (2007), with more recent updates by Voss, Voss, and Lerche (2015). An overview of the procedure is provided here, and the reader is referred to these papers for full details. One of the core issues in fitting diffusion models is obtaining an estimate of the RT distributions for both the lower and upper response thresholds (in the current case, the grammatical and ungrammatical response thresholds) for each of the conditions of interest. Rather than estimating these distributions independently, the *fast-dm* procedure instead forms a single distribution of RT that combines the distributions from both the upper and lower response thresholds. To do this, all of the RTs from the lower threshold are multiplied by −1, and a single Cumulative Distribution Function (CDF) is formed. An example of a CDF is shown in Fig. 6. CDFs can be understood in the following way: For each value of a variable x, the CDF gives the probability (plotted on the y-axis) that the variable x will have a value less than or equal to a particular value of x. For *fast-dm*, the variable x is raw RT. It is essential that raw RT is used, as the model explicitly predicts the heavy right tail that generally characterizes RT distributions in chronometric experiments, and which is eliminated by, e.g., the log transform. The CDF captures both the lower and upper RT distribution given the negative transformation of the RTs from the lower threshold.

The procedure then compares the CDF produced from the observed data with a CDF formed by the estimated parameters of the diffusion model, and attempts to minimize the vertical distance between these two functions, as diagnosed by the KS statistic, by searching the parameter space using the Nelder-Mead Simplex algorithm (Nelder & Mead, 1965). Again, an example of what the model CDF might look like in comparison to the experimental CDF is shown in Fig. 6. The model generates a single p-value associated with the final KS statistic, which indicates the probability that the observed data would be sampled from the distribution generated by the best-fitting model parameters. A low p-value indicates that the observed data are *unlikely* to be sampled from this distribution—that is, poor model fit. In the case that parameters are estimated with respect to different conditions, the overall p-value represents the product of the p-values generated within each condition. Therefore models where parameters are fit to multiple conditions for each participant tend to be overly conservative in indicating a poor fit. Models with a final overall probability of $p < 0.05$ are conventionally considered to be poor fits of the data.

11.2. Analysis of the current experiments

The present models were coded such that the upper threshold was a “grammatical” response, and the lower threshold an

Table 11
Summary of drift diffusion model parameters, the cognitive interpretation of each, and the setting for the current model.

Parameter		Interpretation	Setting
Drift Rate	v	average rate of evidence accumulation	Estimated by subject and condition
Threshold separation	a	response caution	Estimated by subject
Relative starting point	z_r	decision bias	Estimated by subject
Non-decisional constant	t_0	duration of non-decisional processes	Estimated by subject
Difference in non-decisional constant	d	response preparation/response inhibition	Set to default value (0)
Inter-trial drift rate variability	s_v	differences in stimulus properties, fluctuations in attention	Set to default value (0)
Inter-trial starting point variability	s_{zr}	differences in expectations	Set to default value (0)
Inter-trial non-decisional constant variability	s_{t0}	differences in speed of response execution	Estimated by subject

"ungrammatical" response, as opposed to the conventional "correct" and "incorrect" coding. This ensures that the bias parameter z can be interpreted as reflecting grammatical or ungrammatical bias. However, it results in the sign of the drift rate (v) being positive for grammatical sentences, and negative for ungrammatical sentences. This is discussed further below.

While *fast-dm* makes it possible to estimate every parameter of the model, recent work has shown that it is not always desirable to do so (Lerche & Voss, 2017; Lerche, Voss, & Nagler, 2017). Due to the relatively low number of trials in each of the present experiments (60 trials per subject with 15 in each condition), the most parsimonious model possible that still allows for the testing of the critical hypotheses was chosen. For each participant, the relative position of the starting point (z_r), the decision threshold separation (a), the non-decision constant (t_0), and the inter-trial variability of the non-decisional constant (s_{t0}) were estimated. The estimation of z_r and a allows for a between-subjects analysis of whether experimental instructions shifted bias and response caution. Both t_0 and s_{t0} were estimated, as failing to do so can have a significant effect on the accuracy of the predicted shape of the RT distribution (Voss et al., 2015). The drift rate (v) was estimated within each participant and experimental condition to test the hypothesis that mismatching attractors decrease drift rate. All of the other inter-trial variability parameters were set to default values, as the number of trials was deemed too small to allow for a robust estimation of these parameters (Voss et al., 2015). The difference in the non-decisional constant (d) was set to the default value 0, as doing so ensures an accurate estimation of response bias (Voss, Voss, & Klauer, 2010). A summary of the model parameters, their interpretations, and their settings for the present study is given in Table 11.

11.3. Model fit

A crucial step is to assess the fit of the model to experimental data, as it is the fit of the model that justifies the assumption that the cognitive processes underlying behavior in the experiment are similar to those driving the model. We present an assessment of the model fit first, then discuss the parameter values recovered by the model.

Ideally, the p-value produced by the KS statistic would allow for the model to be validated. However, as alluded to above, the p-value can be artificially lowered as a result of being the product of the p-values produced for each experimental condition, or can be artificially high due to low power (Voss et al., 2015). In the present results, none of the p-values were lower than the 0.05 threshold, giving initial evidence that the model provides a good fit for the data. However, we further validated the results by comparing the accuracy and RT values observed in the experimental data with those produced by the model.

The model RT and accuracy predictions were determined using the *plot-cdf* function within *fast-dm*, which provides a closed-form solution that can derive a CDF given a set of parameter values. The output produced by *plot-cdf* was used to calculate the predicted mean accuracy and mean correct response latency. We then compared these values to those observed in each experiment. Both the model and experimental values are shown in Fig. 7. The full CDFs are given in Appendix C.

We consider first the fit with response accuracy. While the model overestimates accuracy in all conditions, the model captures all of the qualitative patterns in the data, most notably the asymmetrical pattern of agreement attraction in Experiment 1, the increased illusion of ungrammaticality in Experiment 2, and finally symmetrical attraction in Experiment 3. Thus, the current model can capture both the asymmetrical and symmetrical patterns of agreement attraction.

Turning now to the RT fits, overall the results match the qualitative patterns. In general, fit is better in the match compared to the mismatch conditions, and the model overestimates RT across the board. However, the model captures the core patterns that reached statistical significance in the analysis of the three experiments: Mismatch is always slower than match, and in Experiment 1 the ungrammatical conditions are slower than the grammatical conditions—an effect that dissipates over the course of Experiments 2 and 3.

The imperfect fits to both RT and accuracy are most likely due to the low number of trials that served as input to the model, which results in a very low number of errors in some conditions, and a less stable estimation of the distribution of RT. Although the KS optimization method of *fast-dm* is known to provide robust estimates of parameters even with a relatively small number of trials and errors, it does perform better with larger samples (Lerche et al., 2017). However, the present results show that the model captures the general shape of the data, and changes in the qualitative patterns across experiments. Future work should fit the model with a larger sample (at least 50 trials per condition) to further validate the fit of a diffusion model.

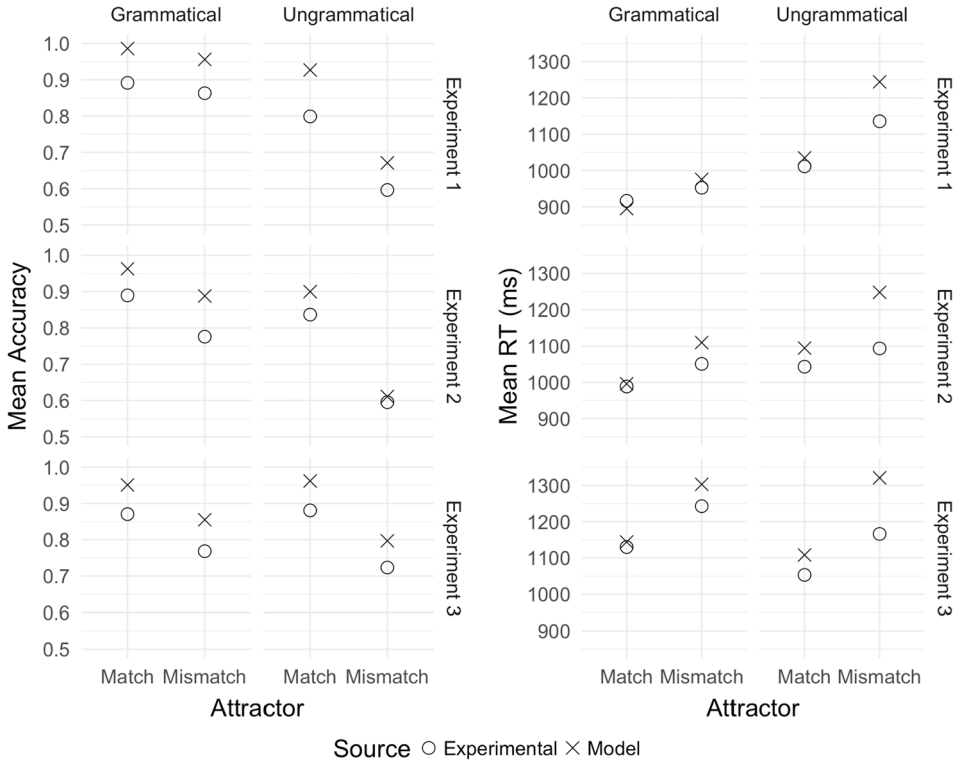

Fig. 7. Model fits for accuracy (left) and RT (right) for each condition of each experiment in the *fast-dm* drift diffusion analysis.

Table 12
Mean recovered relative starting point, threshold separation, and non-decisional time for *fast-dm* diffusion analysis of Experiments 1–3.

	Experiment 1	Experiment 2	Experiment 3
z_r	0.570	0.493	0.447
a	1.584	1.571	1.765
t_0	0.613	0.646	0.671

Table 13
Mean recovered drift rate for each condition in Experiments 1–3 for *fast-dm* diffusion analysis. The coding of the upper and lower thresholds as grammatical and ungrammatical leads drift rates towards correct responses to be positive in the grammatical conditions, but negative in the ungrammatical condition.

	Grammatical		Ungrammatical	
	Match	Mismatch	Match	Mismatch
Experiment 1	2.328	1.666	−1.897	−0.641
Experiment 2	2.112	1.338	−1.374	−0.270
Experiment 3	1.894	1.163	−1.643	−0.643

11.4. Results

The best fitting parameter values are given in Tables 12 and 13. Only those parameters allowed to vary are reported, as the other parameters were constant for all participants and conditions (see Table 11 for the settings of the fixed parameters). Note that the calculation of the parameters in *fast-dm* uses a diffusion constant (a scaling parameter) of 1, where other researchers use a constant of 0.1 (Voss et al., 2015). In order to transform these estimates for comparison to other solutions, one must multiply drift rate, starting point (non-relativized), and threshold separation by 0.1.

11.4.1. Drift rate

In order to assess the hypothesis that mismatching attractors decrease the drift rate, for each experiment the estimated drift rates for each subject and condition were entered into a 2 × 2 ANOVA with grammaticality and attractor number as factors. To allow for a

straightforward comparison between the grammatical and ungrammatical conditions, in the ungrammatical conditions the sign of the drift rate was reversed. This ensured that positive drift rates correspond to correct responses, and negative drift rates towards incorrect responses, for both grammatical and ungrammatical sentences.

For all three experiments, there were significant main effects of grammaticality and attractor number on drift rate: Experiment 1 grammaticality ($F(1, 39) = 7.12$, $p = 0.011$), attractor number ($F(1, 39) = 16.36$, $p < 0.001$); Experiment 2 grammaticality ($F(1, 19) = 6.95$, $p = 0.016$), attractor number ($F(1, 19) = 10.98$, $p = 0.004$); Experiment 3 grammaticality ($F(1, 39) = 5.53$, $p = 0.024$); attractor number ($F(1, 39) = 31.19$, $p < 0.001$). Thus, evidence strength was relatively weak both when there was a mismatching attractor and when the sentence was ungrammatical. The interaction of these factors was marginal in the first experiment ($F(1,39) = 3.38$, $p = 0.074$), and not significant in the latter two (Experiment 2: $F(1, 19) = 0.88$, $p = 0.359$; Experiment 3: $F(1, 39) = 1.39$, $p = 0.245$). We return to our interpretation of this pattern in the general discussion.

11.4.2. Relative starting point

To test the hypothesis that the instruction manipulation affected the decision bias across the experiments, a two-sample *t*-test compared the relative starting point of Experiment 1 to that of Experiment 3. The choice was made to compare only Experiments 1 and 3 as they are equal in the number of subjects (N = 40 each), and they represent the two extremes in the manipulation of the instructions. The test revealed a significant effect ($t(78) = 4.30$, $p < 0.001$), suggesting that the changes to the instructions and fillers succeed in manipulating the starting point in Experiment 3 compared to Experiment 1. One-sample t-tests against a mean of 0.5, which represents a lack of bias, revealed that in Experiment 1 there was a significant bias towards grammatical responses ($t(39) = 3.48$, $p = 0.001$), and in Experiment 3 a significant bias towards ungrammatical responses ($t(39) = -2.61$, $p = 0.013$). The finding of significant ungrammatical bias in Experiment 3 appears inconsistent with the finding from our signal detection bias measure that no bias was present. We discuss this discrepancy below.

11.4.3. Threshold separation and non-decisional constant

To test the possibility that the instruction manipulation affected response caution or the duration of non-decisional processes, two-sample *t*-tests were conducted comparing the recovered threshold separation and non-decisional constants of Experiments 1 and 3. Both tests revealed marginally significant results (a: $t(78) = -1.83$, $p = 0.070$; t_0: $t(78) = -1.80$, $p = 0.076$), and Bayes Factor analyses on each test were equivocal (a: BF = 1.01; t_0: BF = 1.07 in favor of the presence of a difference), and thus failed to show particular support for either the null or alternative hypotheses. Therefore while there is no clear evidence that the instruction manipulation also affected response caution or non-decisional processes, we are also unable to rule out this possibility.

11.5. Discussion

Both the hypothesis that mismatching attractors reduce evidence strength compared to matching attractors regardless of grammaticality, and the hypothesis that bias shifted away from the grammatical threshold from Experiment 1 to Experiment 3, were supported by the recovered parameters from the diffusion model analysis of the three experiments. Statistical tests on the recovered parameters found the expected main effect of attractor number on drift rate, a significant shift in the starting point parameter from Experiment 1 to 3, and no significant effect in either threshold separation or the non-decisional constant. The model captures the qualitative patterns in the accuracy and response latency data for all three experiments, providing further validation of the claim that a constant evidence strength effect, together with shifts in response bias, are responsible for the shifts in the grammaticality asymmetry observed in the three experiments. There was no clear evidence of a grammaticality asymmetry in the evidence strength parameter itself. Instead, the grammatical asymmetry seen in the judgment responses was primarily captured by means of a combination of an effect of mismatch that does not differ between grammatical and ungrammatical sentences, and a bias towards grammatical responses. Overall, the model fits the data well, and can capture both asymmetrical and symmetrical patterns of agreement attraction, and the slowdown in RT in the mismatch conditions compared to the match conditions.

The model did show a numerical increase in threshold separation (a) in Experiment 3 compared to Experiment 1, although this effect did not reach significance, and the Bayes Factor analysis was equivocal with respect to its interpretation. To the extent that this reflects a true difference, it may show that the instruction manipulation had the additional effect of increasing response caution. This could provide an explanation for why RT was slower overall in Experiment 3 compared to Experiments 1 and 2: As detailed in the introduction, higher response caution is generally associated with an overall slowdown in RT.

The model did reveal two somewhat unexpected results. First, we observed a main effect of grammaticality on the drift rate, with higher drift rates for grammatical sentences. Second, the starting point parameter in all three experiments seems lower than we would expect given our measure of bias using c. In particular, we found that the bias in Experiment 3 was significantly skewed towards ungrammatical responses, rather than simply being neutralized. In fact, these two issues may reflect the fact that response bias manipulations may be realized in the diffusion model in either the drift rate or starting point parameters (e.g. Ratcliff & McKoon, 2008; Starns, Ratcliff, & White, 2012; White & Poldrack, 2014). We found faster drift rates for grammatical compared to ungrammatical conditions, and what appears to be an underestimation of bias towards grammatical responses. These two phenomena may be seen as offsetting, despite the fact that changes in drift rate and response bias have subtly different effects of the shape of the distribution of RT (Ratcliff & McKoon, 2008). Again, it is likely that fitting the model to a larger sample will allow for more definitive model fits.

12. General discussion

We investigated the role of response bias in generating the grammaticality asymmetry in agreement attraction, finding that grammatically asymmetrical agreement attraction arises when response bias is present (Experiment 1) and that symmetrical attraction emerges as response bias decreases (Experiments 2 and 3). The results of all three experiments extended the finding from forced-choice production that mismatching attractors slow RT in decision tasks compared to matching attractors (Staub, 2009, 2010); this occurred whether the sentence was grammatical or ungrammatical. Furthermore, the experiments showed that grammatical bias is associated with slower RT in ungrammatical compared to grammatical conditions, while neutralizing bias leads to increased similarity in RT between these two conditions (modulo the unexpected main effect of grammaticality on RT in Experiment 3). Fitting the diffusion model to the results of each experiment provided further support for the linking hypothesis between continuous number marking and the rate of evidence accumulation, and the role of bias in shifting the starting point of the decision-making process and driving asymmetrical attraction.

Our results support the claim that agreement attraction is due to a continuous number representation on the subject, as captured under a spreading-activation account such as Marking and Morphing, as well as the claim that equivocal number marking lowers the rate of evidence accumulation in the decision making process. Rather than indicating a fundamental role for retrieval in agreement processing, where attraction arises due to the cue interference dynamics of the memory retrieval mechanism, grammatically asymmetrical agreement attraction was shown to be largely an artifact of response bias in our experiments. In short, our findings suggest that agreement attraction with PP-modifiers is well modeled by a representational account that allows for a continuous or graded representation of number of the agreement controller. Current implementations of the alternative view—that attraction results from categorical number marking of the subject embedded in a noise-prone memory architecture—fail to account for the full range of effects presented here.

A representational account of agreement attraction has a number of theoretical benefits. Chief among those is that such an account can unify attraction effects in comprehension and production by locating the source of the error in the representation of number on the subject noun phrase. Assuming that the format of the syntactic representations recruited in both comprehension and production is the same (e.g. *representational identity*; Momma & Phillips, 2018), nothing further needs to be said about why attraction occurs in both comprehension and production. This “single mechanism” approach has a conceptual advantage over accounts that posit two separate mechanisms to capture attraction in production and comprehension (e.g. Tanner et al., 2014).

Despite the empirical and conceptual advantages of our account, our conclusions and results nonetheless do raise questions about how to account for asymmetrical attraction effects in more implicit measures such as eye tracking and EEG. These claims and issues, as well as additional implications of the present findings, are discussed below.

12.1. Bias in psycholinguistic measures

We first address the methodological implications of our work. The finding that symmetrical effects of agreement attraction can be masked by the presence of response bias presents a serious methodological lesson for psycholinguistics. Rotello, Heit, and Dubé (2015) showed that highly replicable and theoretically important patterns from a variety of domains are in fact artifacts of response bias. Our findings lead us to ask whether an important psycholinguistic result—the grammaticality asymmetry—is in this class.

The solution to the methodological challenge raised by our work is not to simply collect more data. In fact, Rotello and colleagues point out that a central conundrum is that interactive effects due to bias actually *strengthen* as statistical power increases; thus, the role of bias evades discovery when experiments are replicated in the usual manner. The suggested antidote is to employ analyses where bias is accounted for within the model, and to understand the signs that bias is present within a pattern of responses so that the analyst can monitor for them. In the present study, we manipulated response bias explicitly, and we also used the drift diffusion analysis, which accounts for bias by relegating these effects to the starting point parameter.

To our knowledge, we are among the first to apply diffusion models to grammaticality judgment tasks (though see Chen & Husband, 2018; and for a discussion of linguistic acceptability as psychological evidence see Bard, Robertson, & Sorace, 1996). An upshot of the present result is a proof of concept that the diffusion model can be used as a powerful tool to understand the cognitive process that drive behavior in linguistic processing. Besides providing a way to factor out effects of bias in judgment tasks, the model provides a unified analysis of both latency and accuracy data, providing a path for accounting for effects of the speed-accuracy tradeoff in a less demanding way than methods such as SAT. The present paper is further aligned with recent work that applies the Receiver Operating Characteristic (ROC) curve (Green & Swets, 1969; Macmillan & Creelman, 1991) to model scaled acceptability judgments (Dillon, Andrews, Rotello, & Wagers, 2017). Like our approach, ROC analysis provides a means of controlling for response bias, and as such is an alternative methodological approach to address the issue raised by response bias (Rotello et al., 2015).

These considerations are of vital importance to the field of psycholinguistics, as decision-making processes underlie a wide variety of tasks used to probe psycholinguistic questions. In binary judgment tasks such as that employed in the current study, the relationship between decision-making and our response measures is relatively straightforward: Participants are required to choose between two options in judging whether a stimulus is grammatical or ungrammatical. A similar logic applies to scaled judgments, where participants must decide where a given stimulus falls on the presented scale. Any task that involves judgment necessarily involves decision-making, and decision-making involves a variety of components that can obscure direct measurement of the cognitive variables related to language processing.

Despite these issues, the present study provides reasons for optimism with respect to explicit judgment tasks. While the dependent measures used with judgments are highly susceptible to bias effects, we can use models of the decision process, such as the diffusion

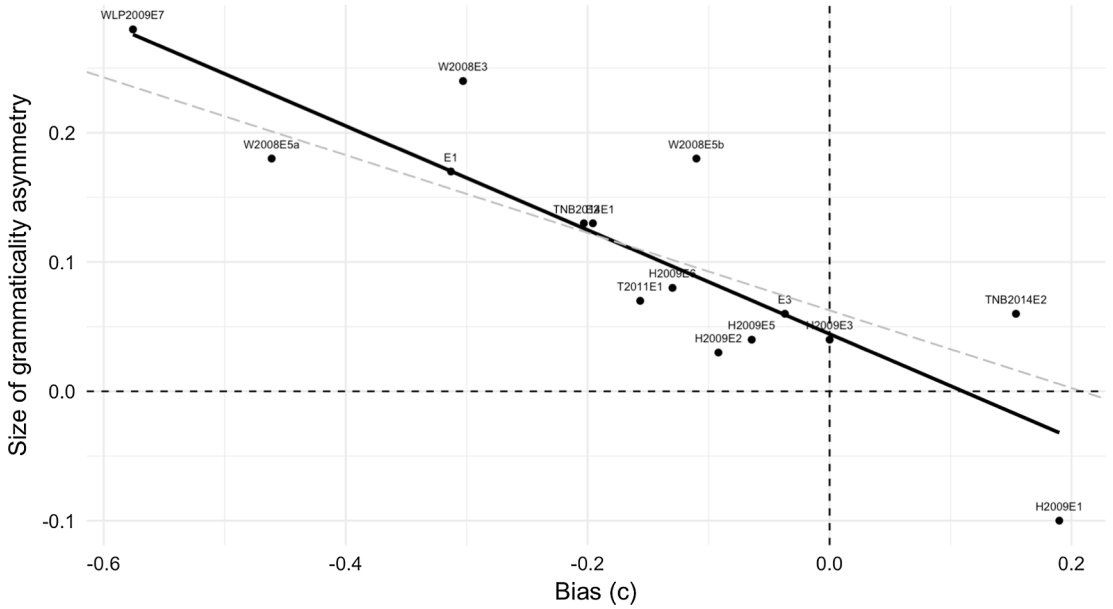

Fig. 8. Size of grammaticality asymmetry versus signal detection bias in all binary judgment experiments in the literature where mean accuracy was reported, including the current experiments. The solid black regression line is that derived from the literature, and the dotted gray regression line is derived from the by-subjects regression in Fig. 5. The labels on the points indicate the source. The label was created by taking the first letter of each author's last name, followed by the year of publication, followed by the experiment number as reported in the original paper.

model, to disentangle the effects of decision-making from those that arise from language processing. In this sense, judgment tasks are extremely well suited to reveal the processes that underlie language comprehension: The task-specific effects that would otherwise obscure effects from language processing itself are well understood, and can be accounted for in the analysis. As a result, we can draw conclusions that are more likely to be independent of the task itself, and speak to the mental states that underlie sentence processing.

While not frequently considered, there are also decisional components to what are usually considered implicit dependent measures, such as reading time in self-paced reading, where participants read through a sentence one word or chunk at a time by pressing a button to proceed from one region to the next. Even if participants are reading in order to accurately answer a comprehension question at the end of the sentence, they are required to make a decision in each region of the sentence about whether to continue reading that region, or to move forward to the next region by making a button press. The processes that underlie the decision to press a button are assumed to reflect the ease or difficulty with which the different chunks of the sentence are comprehended. However, button press latency in self-paced reading likely reflects a combination of non-decisional processes such as word recognition, the certainty with which the stimulus present in a given region can be integrated with the information processed within the previous regions, and most pertinently for the results of the present paper, the degree of bias a participant has towards a decision that culminates in pressing a button to move forward in the sentence. As has been demonstrated in forced-choice decision tasks, these factors are not likely to be additive in their effect on response time measures, but rather stand in interactive relationships that can obscure otherwise cognitively independent effects.

This finding also raises the question of whether differences in bias are responsible for the relative instability of the grammaticality asymmetry observed in the literature review summarized above and shown in Table 1. To examine this question, we calculated the overall bias in each of the binary forced-choice judgment studies that reported the full set of accuracy data, and compared this to the size of the interaction effect observed in each experiment.

The results are shown in Fig. 8. We observe the expected effect of bias on the size of the grammaticality asymmetry, given the results of the current experiments. The more bias an experiment shows towards grammatical responses (i.e. the more negative c becomes), the more pronounced the asymmetry becomes. Importantly, interpolation from the regression line reveals that a lack of bias predicts a lack of grammaticality asymmetry, as the y-intercept is near the origin. We further observe that experiments with bias in the opposite direction reveal a trend towards the opposite asymmetry—a result that is expected given the continuous valuation model combined with the diffusion analysis of the decision task. We also note the remarkable similarity between the regression line relating individual subjects' bias and asymmetry in the current experiments, and the regression line relating bias and asymmetry across experiments in the literature as a whole.

The question remains, though, of how more implicit measures of processing such as EEG and eye tracking are affected by bias, where it is not clear that an explicit decision-making process is occurring at the verb. The small literature on agreement attraction using these methods has been inconclusive with respect to the grammaticality asymmetry, or faces methodological challenges. In eye tracking, two out of three of the reviewed studies failed to find a significant asymmetry. It is therefore possible that eye tracking provides a measure that is more resistant to effects of bias, and naturally shows symmetrical effects. However, it is also the case that

the Pearlmutter et al. (1999) study, which provides the strongest evidence for symmetrical attraction in eye tracking and the only reported study to test the PP modifier construction, may be confounded by the plural processing penalty, which can independently create longer reading times in the grammatical mismatch condition (Wagers et al., 2009). In any case, more work is needed to determine the profile of agreement attraction in eye tracking measures.

In EEG, all known studies have used judgment tasks while recording ERPs, and the results of the majority of these studies show a significant grammaticality bias (see Fig. 8). These results rely on the failure to find a significant P600 in grammatical mismatch sentences compared to match sentences. In general, the P600 is thought to index violations of morphosyntactic constraints, however it is possible that the judgment task, and bias in that task, may also modulate this effect. We plan to make this hypothesis the object of future work.

12.2. The grammaticality asymmetry

Our experimental results and our literature review together suggest that the grammaticality asymmetry is both weaker than previously thought, and largely an artifact of response bias. We have argued that this challenges the prevailing interpretation of this finding. Overall, our results favor a representational account of agreement attraction over accounts that attribute this effect to a memory retrieval error. However, it is important to note that our modeling and literature review reveal that a modest grammatical asymmetry in acceptability judgments may still exist, even after partialing out response bias. Consider the drift rate parameters we recovered from our data in Table 13. It can be seen that number mismatch has a more modest impact on drift rate for grammatical sentences than it does for ungrammatical sentences. On average, mismatch reduced the drift rate by 0.72 in grammatical sentences, but 1.12 in ungrammatical sentences. In other words, a number mismatching distractor reduces the rate of evidence accumulation more in ungrammatical than grammatical sentences. This may reflect a real difference. Even though the critical interaction was not significant for any one of the three experiments, a post-hoc omnibus ANOVA over by-participant drift rate estimates for all three experiments, with the additional power that this analysis provides, does yield a significant interaction of grammaticality and attractor mismatch ($F(1,99) = 5.71, p = 0.019$). In addition, we note that both regression lines in Fig. 8 have slightly positive *y*-intercepts. This means that a perfectly unbiased responder is predicted to show a very modest grammaticality asymmetry. If we take these regressions at face value, we would predict approximately 5% more errors in ungrammatical than grammatical conditions for a perfectly unbiased responder. In this way, our results indicate that there may still be a small, but non-zero difference in the magnitude of the mismatch effect in grammatical and ungrammatical differences.

We do not have firm statistical evidence to support this limited grammaticality asymmetry effect, and therefore we refrain from too strongly interpreting this finding. At a minimum, however, it suggests that the agreement computation process in comprehension may indeed be influenced by the number marking on the verb itself, to a limited degree. Still, it is clear that verb number does not interact with the agreement computation to the degree predicted by current implementations of cue-based retrieval models. The effect size of the asymmetry that remains after the effect of bias is accounted for is far smaller than what our simulations from cue-based retrieval predict (see Appendix A). Furthermore, we emphasize that the current implementations of the cue-based model predicts not only an asymmetrical effect such that there is greater attraction in ungrammatical than grammatical sentences, but a lack of an effect in grammatical sentences. Further work is necessary to understand how, and to what degree, the verb number marking itself interacts with the judgment process under investigation here. However, it is clear that the verb's effect on the processing of number agreement is less salient than is predicted on a cue-based model alone.

12.3. Cue-based retrieval and RT

We have primarily emphasized the ability of a continuous number valuation account of agreement attraction, embedded within a diffusion decision model, to capture the patterns of judgment accuracy across conditions and across experiments. But we have also pointed out that this model naturally predicts the observed RT patterns. In particular, it predicts the reliable finding – emerging across both grammatical and ungrammatical sentences in all three experiments – that responding is slowed in the presence of a mismatching attractor. Staub (2009) found similar RT effects when the subject's task was to choose a verb form, and found a clear correspondence, across conditions, between the size of the RT effect and the effect of a mismatching attractor on accuracy (see also Keung & Staub, 2018). The architecture of the diffusion model ensures that changes in evidence strength, across conditions, result in effects on both accuracy and RT: As accuracy decreases, RT increases.

Is this correct prediction of the RT patterns another feature that distinguishes our preferred account from a cue-based retrieval account? Or, does a cue-based retrieval account also predict that the presence of a mismatching attractor should slow retrieval? In activation-based architectures such as ACT-R, activation determines both the speed and accuracy with which retrieval occurs (Lewis & Vasishth, 2005). Activation levels are a function of cue match with the retrieval trigger, similarity-based interference with other potential antecedents, time-based decay, and random noise. The retrieval process is instantiated as a race, where the encoding with the highest activation wins. When multiple elements share the same cues, retrieval is predicted to be slowed due to competition between these elements—an effect known as inhibitory interference. But when multiple elements share distinct cues with the retrieval trigger, then retrieval is predicted to be faster due to facilitatory interference (see Jäger et al., 2017 for more detail, and Appendix A for an R-based simulation). Set in terms of the PP modifier agreement attraction configuration, the account predicts that retrieval in attractor match conditions should be *slower* than in the attractor mismatch conditions. In grammatical sentences, this arises due to inhibitory interference in the match condition (Nicenboim et al., 2018). In ungrammatical sentences, this occurs due to facilitatory interference in the mismatch condition. In sum, this implementation of the cue-based retrieval process not only fails to

predict the increased RT in mismatch conditions that is reliably observed across experiments, but actually predicts the opposite pattern of faster retrieval in the presence of a mismatching attractor.

Several caveats may be in order. First, Vasishth et al. (2008) have explicitly cautioned against assuming a transparent mapping between retrieval times derived from a model like ACT-R and judgment latencies, as a complex decision process presumably intervenes between retrieval and judgment. We note, however, that it is not obvious how a decision process would result in the pattern of predicted retrieval latencies – longer times for match than mismatch – actually being reversed in judgment RT.

Second, we have explored the retrieval latency predictions of one specific version of cue-based retrieval. It is not clear whether other variants of cue-based retrieval could capture the critical pattern. For example, McElree, Foraker, and Dyer (2003) assume that retrieval speed is constant, with only retrieval accuracy being affected by the presence of interference. This model, like the diffusion model presented here, holds that differences in RT may be a function of the quality of the representation retrieved. In order to account for our RT findings on this model, one would need to assume that the number mismatching configurations diminish the quality of the memory encodings of the target and the distractor, which would negatively impact judgment processes that rely on retrieving these noun encodings. It is not clear that this assumption is independently warranted, however. A priori, we might expect the number mismatching configurations to permit a more stable encoding of each noun, because the number mismatch makes them more distinct in memory. But again, this view makes the wrong RT prediction, as it would predict faster judgment times for mismatching configurations than matching configurations, owing to the higher quality of the memory encodings of target and distractor in mismatch configurations.

In short, we do not see a clear path to reconcile cue-based retrieval models with the RT findings presented here. A representational account, however, immediately predicts the observed RT effects. The consistent slowdown in judgment times we observed for mismatching configurations provides an additional data point in favor of a representational account of agreement attraction effects.

12.4. Agreement attraction across constructions

We now turn to an additional caveat in interpreting the present results. The three experiments presented in the current paper were restricted to the PP modifier construction. This constitutes a case of *intervening* attraction, where the attractor linearly intervenes between the subject head noun and the critical verb. Other cases of linearly intervening attraction include the subject relative clause modifier (Bock & Cutting, 1992; Hammerly & Dillon, 2017), the possessive relative clause (Häussler, 2009), and the genitive possessive construction (Lago et al., 2018). In the literature, so-called *non-intervening* attractors, where the attractor does not linearly intervene between the subject and the verb, have also been frequently investigated. This most prominently includes the object relative clause construction (e.g. *the runner(s) that the driver wave(s) to…*), where the critical verb is embedded in the relative clause (*wave*), and the attractor is the head of the relative clause (i.e. *the runner*). The object relative clause construction was the focus of the original investigation of the grammaticality asymmetry in Wagers et al. (2009), and therefore warrants consideration in light of the results reported here.

At present, we do not have direct evidence to support the extension of a representational account such as Marking and Morphing to the non-intervening attraction cases, though it is possible in principle. In contrast to the percolation-style accounts discussed in the introduction, where features can only percolate *up* through the syntactic structure (Bock & Eberhard, 1993; Eberhard, 1997; Franck et al., 2002; Vigliocco et al., 1995), the spreading activation function of Marking and Morphing can spread the feature value bidirectionally in the syntactic representation (Eberhard et al., 2005, p. 544). Therefore one could appeal to representational issues with number marking on the subject to account for attraction in both intervening and non-intervening constructions. This would predict that the grammaticality asymmetry in object relative clause constructions should also be shifted by bias in a similar fashion to that observed in the three experiments presented here. Given the review in Table 1, the results in the literature are inconclusive, as 8 of the 12 studies on the object relative clause constructions found a significant grammaticality asymmetry.

However, there is existing evidence that attraction in intervening and non-intervening constructions may arise from distinct sources, which may give reason to withhold direct application of the representational account to the non-intervening cases. As early as Bock and Miller (1991), it was noted that effects of animacy mismatch between the head noun and attractor differ in the intervening and non-intervening constructions—a result that has been interpreted to support the idea that attraction in the non-intervening construction results from true subject misidentification (Eberhard et al., 2005). We take up the issue of subject misidentification in more detail below. Furthermore, the RT distributions in forced-choice production tasks indicate that a mismatching attractor shifts the entire distribution to the right in the intervening construction, but only affects the right tail of the distribution in the non-intervening construction (Staub, 2010). This is expected on a theory where representational effects of number on a correctly identified subject influence every trial in the intervening condition, whereas in the non-intervening conditions, only in the subset of trials where the subject has been misidentified does number have an effect. More direct research on these effects in comprehension measures is needed to disentangle whether agreement attraction in the intervening and non-intervening constructions share the same source.

12.5. Agreement attraction and homuncularity

Finally, we reflect on what is arguably an intuitive appeal of cue-based retrieval as an explanation of agreement attraction effects, and a corresponding intuitive drawback of our preferred account. On the cue-based retrieval account, the reason that comprehenders judge *the key to the cabinets are rusty* to be grammatical on some trials is that, on those specific trials, they retrieved the wrong noun, *cabinets*, as the agreement controller (i.e. subject misidentification occurs). Their overt decision therefore reflects a specific cognitive

event, which is different on those trials than on the trials on which the comprehender correctly judges that fragment to be ungrammatical; on the latter trials, they have retrieved *key* as the controller. Percolation accounts of agreement attraction (e.g., Bock & Eberhard, 1993), while differing from cue-based retrieval accounts in other respects, share the commitment to the idea that a cognitive process has gone wrong on specifically those trials on which attraction is in evidence; the plural feature on *cabinets* has been assigned to the subject phrase as a whole.

The Marking and Morphing model (Eberhard et al., 2005), on the other hand, makes a radically different assumption about the difference between instances on which attraction occurs and instances on which it does not. On this model, *cabinets* exerts its influence by making the subject phrase somewhat plural, or somewhat less clearly singular, on all trials. The trial-to-trial variability in the verb form that a speaker actually uses, or in the present experiments, the variability in whether a verb with a given number is judged grammatical or ungrammatical, is the result of a decision process whose output probabilistically reflects the underlying continuous number valuation, which is itself invariable from trial to trial. In other words, the model assumes that there is no cognitively interpretable difference – neither a difference in which noun is taken to be the agreement controller, nor a difference in the number assigned to the subject phrase – between trials on which attraction does happen, and trials on which it does not.

We may think of the two approaches as differing along the dimension of *homuncularity* (cf. Ryle, 1949). A homuncular theory is one that presumes that corresponding to each overt behavior, there is a cognitively interpretable inner event or state that causes that behavior. In the context of agreement, one such inner event might be described as 'agreement checking,' an explicit evaluation of whether a retrieved noun and the verbal agreement morphology are in agreement. A non-homuncular theory, on the other hand, allows that at least some variability in overt agreement behavior may not be traceable to variability in a cognitively interpretable inner event or state. In the context of agreement, variability in overt behavior may be thought to arise simply because a continuously-valued representation must be converted into a binary response, due to the formal requirements of the language or the demands of the task.

We acknowledge that homuncular theories correspond to intuition. But despite their appeal, we think there are no good reasons to prefer them in this context. Non-homuncular theories are theoretically parsimonious, and as already discussed at length, evidence accumulation models such as the diffusion model, which embody non-homuncularity, have been extremely successful in explaining decision processes in a wide variety of domains (Ratcliff & McKoon, 2008). In addition, the kinds of evidence that would directly support a homuncular model of agreement attraction are conspicuously absent or weak. One kind of evidence would consist of the finding that an attractor sometimes induces a speaker or comprehender to treat the wrong noun as the thematic subject; that is, true subject misidentification of the type discussed above. But in fact, there is fairly clear production evidence that this does not happen, at least in the standard intervening attraction configuration (Antón-Méndez, Nicol, & Garrett, 2002; Barker, Nicol, & Garrett, 2001; Bock & Miller, 1991; Eberhard et al., 2005), though we leave open the possibility that such a process occurs in the non-intervening configurations. Another kind of evidence would consist of the finding that an attractor can induce a speaker or comprehender to encode the subject as genuinely plural, in a manner that has downstream ramifications, for example in a post-sentence task that probes the final number representation of the head noun. Existing evidence for this claim is mixed (Brehm, Jackson, & Miller, 2018; Patson & Husband, 2016; Schlueter, Parker, & Lau, 2017; Tanner, Dempsey, & Christianson, 2018).

13. Conclusion

We have argued that the grammaticality asymmetry in agreement attraction arises due to response bias, rather than the dynamics of cue-based memory retrieval, supporting a representational account of agreement attraction where mismatch effects arise due to a noisy or equivocal representation of the subject's number. In three experiments, we found a clear relationship between the degree of bias and the magnitude of the grammaticality asymmetry, and showed that an identical relationship between bias on the grammaticality asymmetry is present in the literature, which may account for the unstable nature of this effect. We adopted a linking hypothesis that ties the equivocal representation of number within the Marking and Morphing model to the rate of evidence accumulation in a decision-making model, in order to account for both accuracy and response time in our binary forced-choice acceptability judgment task. The analysis revealed that, contrary to claims in the literature, asymmetrical attraction can arise on a representational account, but only in the presence of response bias. The underlyingly symmetrical pattern of agreement attraction reveals itself as bias is neutralized. The diffusion model captured both response time and accuracy through the interaction of the starting point parameter, which changes with shifts in bias, and drift rate, which is uniformly lower in the attractor mismatch conditions compared to the match conditions, independently of bias. Together, the data highlight the importance of using explicit models of decision making in analyzing data from psycholinguistic tasks, and support the view that a representational change in number marking, rather than retrieval interference, is responsible for agreement attraction effects in comprehension.

Appendix A. ACT-R simulation of cue-based retrieval

In order to confirm the accuracy predictions, and derive the retrieval latency, of the ACT-R implementation of cue-based retrieval for the PP modifier construction, we simulated retrieval using the ACT-R-in-R model first developed by Rick Lewis and William Badecker, and extended by Felix Engelmann. The script for the model can be found here: https://github.com/felixengelmann/ACTR-in-R. We note that similar results can be obtained through the Shiny App implementation, which can be found here: https://engelmann.shinyapps.io/inter-act/.

The latencies of the onsets of each word in the model mirrored the SOA of the RSVP presentation used in the experiment. Three features were used for retrieval: syntactic category (e.g. DP, NP, PP), number (PL/NULL), and case (NOM/NULL). Default values for

Table A1
Retrieval accuracy and correct retrieval latency for the ACT-R-in-R simulation of retrieval for the PP modifier construction.

	Grammatical		Ungrammatical	
	Match	Mismatch	Match	Mismatch
Accuracy	0.99	0.99	0.99	0.71
Latency (ms)	150	132	770	642

all parameters (as described in Lewis & Vasishth, 2005) were used, except for the latency scaling parameter *F* (a constant), which was increased from 0.14 to 0.5 to exaggerate the scale RT differences (this has no effect of the qualitative patterns). 1000 Monte Carlo trials of the simulation were run. Retrieval accuracy and latency are reported only for the critical verb. The results are given in Table A1.

The accuracy results show the expected asymmetry between agreement attraction in grammatical and ungrammatical attraction. While there is an attraction effect of 0.28 in the ungrammatical conditions, there is no hint of attraction in the grammatical sentences. This matches the qualitative predictions of Wagers et al. (2009). In latency, the mismatch conditions are predicted to be faster than the match condition. This is due to the presence of inhibitory interference in the match conditions compared to the mismatch conditions (Jäger et al., 2017). Note this is the opposite of the predictions of the drift diffusion model, which predicts mismatch to be slower than match due to a decrease in the drift rate as a function of inconclusive evidence about the number marking of the subject.

Appendix B. Stimuli

Lexical version of experimental stimuli adapted from Staub (2009). In Auxiliary condition of Experiment 1, items 1–20 were *was/were*, items 21–40 *is/are*, and items 41–60 *has/have*.

1. The slogan on the poster(s) really surprise(s)
2. The label on the bottle(s) usually rip(s)
3. The wall with the advertisement(s) often change(s)
4. The picture on the postcard(s) frequently amaze(s)
5. The problem with the school(s) unfortunately remain(s)
6. The car near the garage(s) frequently stall(s)
7. The typo in the book(s) always confuse(s)
8. The neighborhood with the museum(s) always recover(s)
9. The memo from the accountant(s) usually bore(s)
10. The letter from the lawyer(s) supposedly intimidate(s)
11. The warning from the expert(s) never change(s)
12. The check from the stockbroker(s) always clear(s)
13. The door to the office(s) unfortunately stick(s)
14. The entrance to the laborator(y/ies) unexpectedly close(s)
15. The apartment with the leak(s) allegedly stink(s)
16. The student with the backpack(s) probably stud(y/ies)
17. The paycheck for the maid(s) always arrive(s)
18. The paint on the fence(s) often change(s)
19. The argument about the bill(s) never end(s)
20. The map of the creek(s) always confuse(s)
21. The record from the singer(s) frequently play(s)
22. The manager of the archive(s) probably remember(s)
23. The pamphlet from the agenc(y/ies) foolishly plagiarize(s)
24. The service by the preacher(s) potentially help(s)
25. The leader of the riot(s) frequently resist(s)
26. The light in the hallway(s) unexpectedly change(s)
27. The worker at the factory(s) never steal(s)
28. The advisor of the student(s) carefully read(s)
29. The principal of the school(s) rarely snap(s)
30. The memo for the executive(s) suddenly end(s)
31. The soldier by the tank(s) impatiently wait(s)
32. The office of the accountant(s) apparently stink(s)
33. The actor in the film(s) sometimes cr(y/ies)
34. The input from the consultant(s) always interfere(s)
35. The assistant for the lab(s) rarely sleep(s)
36. The courier with the message(s) usually run(s)

37. The star of the musical(s) always sing(s)
38. The picture of the politician(s) apparently offend(s)
39. The composer of the opera(s) never speak(s)
40. The teacher with the certificate(s) frequently rant(s)
41. The editor of the book(s) hardly work(s)
42. The applicant for the award(s) supposedly astonish(es)
43. The office with the computer(s) still smell(s)
44. The demonstrator at the rall(y/ies) repeatedly yell(s)
45. The volunteer in the village(s) frequently dream(s)
46. The assistant to the politician(s) sometimes interject(s)
47. The dream about the castle(s) frequently repeat(s)
48. The story about the goat(s) always delight(s)
49. The lane for the bus(es) apparently help(s)
50. The announcement about the game(s) never play(s)
51. The proposal about the park(s) supposedly work(s)
52. The claim by the judge(s) needlessly provoke(s)
53. The paper from the student(s) clearly succeed(s)
54. The envelope for the application(s) always rip(s)
55. The advertisement for the club(s) always work(s)
56. The helicopter of the executive(s) never crash(es)
57. The contract for the actor(s) rarely satisfy(y/ies)
58. The bill from the accountant(s) always arrive(s)
59. The friend of the nurse(s) frequently visit(s)
60. The museum with the sculpture(s) always amaze(s)

Fillers consisted of three types of structures: coordinated NP subjects (1–20), Object Relative Clauses (21–50), and complement clauses (51–80). The lexical versions are shown below. For the between-subjects manipulation in Experiment 1, *is/are* was used for items 1–4, 11–14, 21–23, 32, 33, 36–40, 51–55, 66–70, *was/were* for items 5–7, 15–17, 14–27, 35, 41–45, 56–60, and 71–75, and *has/have* for items 8–10, 18–20, 28–31, 36, 46–50, 61–65, and 76–80. In Experiment 1 half were grammatical (items 11–20, 36–50, and 66–80). This is reflected in the items below. In Experiments 2 and 3, only 25% were grammatical (items 1–6, 36–42, 66–72).

1. The dog and the cat always plays
2. The sailor and the captain usually fights
3. The chair and the desk frequently breaks
4. The computer and the printer often restarts
5. The tyrant and the queen suddenly agrees
6. The tailor and the customer cheerily talks
7. The guitar and the amplifier really rocks
8. The book and the magazine never sells
9. The thesaurus and the dictionary frequently changes
10. The jersey and the glove probably stinks
11. The fireman and the cop usually chat
12. The executioner and the victim never speak
13. The chainsaw and the lawnmower frequently break
14. The coffee and the muffin never cool
15. The camera and the lens definitely amaze
16. The picture and the frame apparently match
17. The microphone and the guitar loudly buzz
18. The cash and the check likely suffice
19. The folder and the file apparently explain
20. The desk and the bookshelf always squeak
21. The teacher supports the student that the principal usually punish
22. The raccoon dug up the rose bush that the gardener usually trim
23. The girl rode the elephant that the zookeeper carefully train
24. The crowd likes the movie that the theater always play
25. The dad hated the story that the child really love
26. The man fixed the leash that the dog aggressively chew
27. The mother sewed the shirt that the boy deeply treasure
28. The cop sips the coffee that the waitress carefully pour
29. The cat lost the toy that the kid always buy
30. The magazine publishes the essays that the editor wrongly reject

31. The committee awards the building that the architect truly love
32. The teenager hates the tourists that the city usually attract
33. The landlord fixes the leaks that the pipe repeatedly produce
34. The worker replaces the light that the tenant always break
35. The critic praised the restaurant that the celebrity really like
36. The camera filmed the line that the runners quickly cross
37. The clock shows the time that the athletes amazingly run
38. The waiter brought the meal that the boys apparently want
39. The shelf holds the phone book that the secretaries often use
40. The crew cleans the plane that the pilots expertly fly
41. The book describes the war that the citizens sadly remember
42. The photographer captured the tree that the flowers beautifully adorn
43. The butler checked the window that the maids always clean
44. The crowd cheered the player that the coaches apparently support
45. The electrician lost the hammer that the carpenters always use
46. The billionaire tips the apprentice that the tailors usually train
47. The editor rejects the titles that the journalists regularly propose
48. The captain releases the whales that the fishermen sometimes catch
49. The doctor caught the flu that the shots supposedly prevent
50. The nurse loves the decorations that the residents always make
51. The owner claimed that the restaurant often fill
52. The teacher said that the student tentatively pass
53. The diplomat believes that the guard probably suffice
54. The repairman said that the refrigerator still work
55. The nurse knows that the treatment probably work
56. The thief thinks that the window likely unlock
57. The chef knows that the dish probably amaze
58. The mechanic insists that the brake still work
59. The lumberjack loves that the chainsaw never break
60. The camper knows that the cabin usually charm
61. The mayor knows that the sheriff often complain
62. The astronaut thinks that the engine never fail
63. The salesman knows that the customer usually decline
64. The technician found that the laboratory never quiet
65. The boss insists that the secretary always smile
66. The pilot thinks that the passengers usually sleep
67. The stewardess knows that the pilots always gossip
68. The bus driver worries that the brakes frequently squeak
69. The chimney sweep found that the bricks often break
70. The model accepted that the shoes never fit
71. The novelist wrote that the characters suddenly meet
72. The captain knows that the sailors often smoke
73. The professor reported that the students always complain
74. The principal discovered that the teachers never sleep
75. The nurse saw that the soldiers never rest
76. The reporter knows that the spies often vanish
77. The analyst discovered that the accountants sometimes miscalculate
78. The manager reported that the contracts probably expire
79. The janitor knows that the offices quickly dirty
80. The lawyer said that the clients frequently lie

Appendix C. Comparison of experimental and model CDFs for *fast-dm* analysis of Experiments 1–3

The *fast-dm* procedure compares the CDF predicted by the model to the empirical CDF produced by the experimental data, and attempts to minimize the distance between the two functions. To provide an additional assessment of model fit, in Fig. C1 we provide both the model and experimental CDFs for all conditions of the three experiments. As in the assessment provided in the body of the paper, the model shows qualitatively good fit to the experimental data, capturing the shape of the distribution in all conditions of all three experiments. As the CDF also shows information about the fit to the RT distribution of the error response threshold (the lower boundary in grammatical conditions and the upper boundary in ungrammatical conditions), this further supports the adoption of the assumption that the recovered parameters reflect the cognitive processes underpinning the behaviors observed in the task.

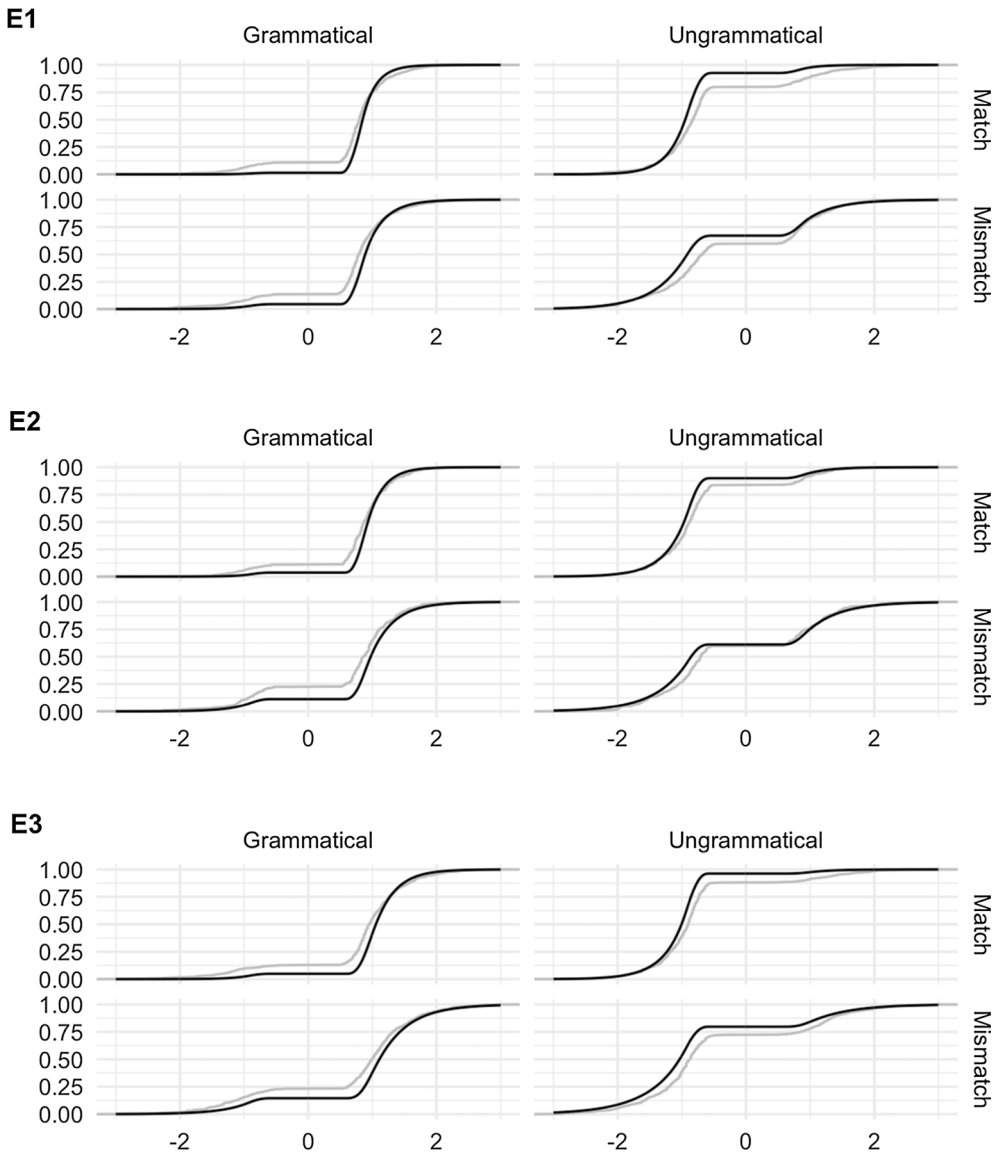

Fig. C1. CDFs produced by the experimental data (gray) and the best-fitting model parameters (black) for all four conditions in each of the experiments.

Appendix D. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cogpsych.2019.01.001.

References

Antón-Méndez, I., Nicol, J. L., & Garrett, M. F. (2002). The relation between gender and number agreement processing. *Syntax, 5*(1), 1–25.
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.
Bard, E. G., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language,* 32–68.
Barker, J., Nicol, J., & Garrett, M. (2001). Semantic factors in the production of number agreement. *Journal of Psycholinguistic Research, 30*(1), 91–114.
Bock, K., & Cutting, J. C. (1992). Regulating mental energy: Performance units in language production. *Journal of Memory and Language, 31*(1), 99–127.
Bock, K., & Eberhard, K. M. (1993). Meaning, sound and syntax in English number agreement. *Language and Cognitive Processes, 8*(1), 57–99.
Bock, K., Eberhard, K. M., Cutting, J. C., Meyer, A. S., & Schriefers, H. (2001). Some attractions of verb agreement. *Cognitive Psychology, 43*(2), 83–128.
Bock, K., & Miller, C. A. (1991). Broken agreement. *Cognitive Psychology, 23*(1), 45–93.
Brehm, L. E. (2014). *Speed limits and red flags: Why number agreement accidents happen* (Unpublished doctoral dissertation)University of Illinois at Urbana-Champaign.
Brehm, L., Jackson, C. N., & Miller, K. L. (2018). Speaker-specific processing of anomalous utterances. *Quarterly Journal of Experimental Psychology,* 1–15.

Bröder, A., & Malejka, S. (2017). On a problematic procedure to manipulate response biases in recognition experiments: The case of "implied" base rates. *Memory, 25*(6), 736–743.
Chen, S. Y., & Husband, E. M. (2018). Comprehending anaphoric presuppositions involves memory retrieval too. *Proceedings of the Linguistic Society of America, 3*(1), 44–51.
Clifton, C., Frazier, L., & Deevy, P. (1999). Feature manipulation in sentence comprehension. *Rivisti di Linguistica, 11*(1), 11–39.
Deevy, P. L. (1999). *The comprehension of English subject-verb agreement* (Unpublished doctoral dissertation)University of Massachusetts Amherst.
Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93*(3), 283.
Dillon, B., Andrews, C., Rotello, C. M., & Wagers, M. (2017). *A new argument for co-active parses during language comprehension.* Open Science Framework.
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.
Eberhard, K. M. (1997). The marked effect of number on subject–verb agreement. *Journal of Memory and language, 36*(2), 147–164.
Eberhard, K. M., Cutting, J. C., & Bock, K. (2005). Making syntax of sense: Number agreement in sentence production. *Psychological Review, 112*(3), 531.
Engelmann, F., Jäger, L.A., Vasishth, S., 2018. The effect of prominence and cue association in retrieval processes: A computational account. https://doi.org/10.31234/osf.io/w2ckt.
Enochson, K., & Culbertson, J. (2015). Collecting psycholinguistic response time data using Amazon Mechanical Turk. *PloS one, 10*(3).
Foraker, S., & McElree, B. (2011). Comprehension of linguistic dependencies: Speed-accuracy tradeoff evidence for direct-access retrieval from memory. *Language and linguistics compass, 5*(11), 764–783.
Franck, J., Colonna, S., & Rizzi, L. (2015). Task-dependency and structure-dependency in number interference effects in sentence comprehension. *Frontiers in Psychology, 6*, 349.
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.
Franck, J., & Wagers, M. W. (2015). *Hierarchical structure and memory retrieval mechanisms in attraction: An SAT study. 28th annual CUNY conference on human sentence processing.*
Gillespie, M., & Pearlmutter, N. J. (2011). Hierarchy and scope of planning in subject–verb agreement production. *Cognition, 118*(3), 377–397.
Green, D. M., & Swets, J. A. (1969). *Signal detection theory and psychophysics.* New York: Wiley.
Hammerly, C., & Dillon, B. (2017). *Restricting domains of retrieval: Evidence for clause-bound processing from agreement attraction. 30th CUNY sentence processing conference.*
Harrison, A. J. (2009). *Production of subject-verb agreement in Slovene and English* (Unpublished Doctoral Dissertation)University of Edinburgh.
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.
Haskell, T. R., & MacDonald, M. C. (2003). Conflicting cues and competition in subject–verb agreement. *Journal of Memory and Language, 48*(4), 760–778.
Häussler, J. (2009). *The emergence of attraction errors during sentence comprehension* (Unpublished doctoral dissertation)University of Konstanz.
Healy, A. F., & Jones, C. (1975). Can subjects maintain a constant criterion in a memory task? *Memory & Cognition, 3*(3), 233–238.
Jäger, L. A., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language, 94*, 316–339.
Jeffreys, H. (1961). *Theory of probability.* Oxford: Oxford University Press.
Keung, L., & Staub, A. (2018). Variable agreement with coordinate subjects is not a form of agreement attraction. *Journal of Memory and Language, 103*, 1–18.
Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13.
Lago, S., Gracanin-Yuksek, M., Safak, D. F., Demir, O., Kırkıcı, B., & Felser, C. (2018). Straight from the horse's mouth: agreement attraction effects with Turkish possessors. *Linguistic Approaches to Bilingualism.*
Lago, S., Shalom, D., Sigman, M., Lau, E., & Phillips, C. (2015). Agreement attraction in Spanish comprehension. *Journal of Memory and Language, 82*, 133–149.
Lerche, V., & Voss, A. (2017). Retest reliability of the parameters of the Ratcliff diffusion model. *Psychological Research, 81*(3), 629–652.
Lerche, V., Voss, A., & Nagler, M. (2017). How many trials are required for parameter estimation in diffusion modeling? A comparison of different optimization criteria. *Behavior Research Methods, 49*(2), 513–537.
Lewis, S., & Phillips, C. (2015). Aligning grammatical theories and language processing models. *Journal of Psycholinguistic Research, 44*(1), 27–46.
Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive science, 29*(3), 375–419.
Macmillan, N., & Creelman, C. (1991). *Detection theory: A user's guide.* Cambridge University Press.
McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language, 48*(1), 67–91.
Momma, S., & Phillips, C. (2018). The relationship between parsing and generation. *Annual Review of Linguistics,* (4), 233–254.
Morey, R., & Rouder, J. (2015). Package 'bayesfactor'.
Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal, 7*(4), 308–313.
Nicenboim, B., Vasishth, S., Engelmann, F., & Suckow, K. (2018). Exploratory and confirmatory analyses in sentence processing: A case study of number interference in German. *Cognitive Science,* (42), 1075–1100.
Nicol, J. L. (1995). Effects of clausal structure on subject-verb agreement errors. *Journal of Psycholinguistic Research, 24*(6), 507–516.
Nicol, J. L., Forster, K. I., & Veres, C. (1997). Subject–verb agreement processes in comprehension. *Journal of Memory and Language, 36*(4), 569–587.
Parker, D., Lago, S., & Phillips, C. (2015). Interference in the processing of adjunct control. *Frontiers in Psychology, 6*, 1346.
Parker, D., & Phillips, C. (2017). Reflexive attraction in comprehension is selective. *Journal of Memory and Language, 94*, 272–290.
Parker, D., Shvartsman, M., & Van Dyke, J. A. (2017). The cue-based retrieval theory of sentence comprehension: New findings and new challenges. In L. Escobar, V. Torrens, & T. Parodi (Eds.). *Language processing and disorders.* Newcastle: Cambridge Scholars Publishing.
Patson, N. D., & Husband, E. M. (2016). Misinterpretations in agreement and agreement attraction. *The Quarterly Journal of Experimental Psychology, 69*(5), 950–971.
Pearlmutter, N. J., Garnsey, S. M., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and language, 41*(3), 427–456.
Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *Journal of Neuroscience Methods, 162*(1–2), 8–13.
Phillips, C., Wagers, M. W., & Lau, E. F. (2011). Grammatical illusions and selective fallibility in real-time language comprehension. *Experiments at the Interfaces, 37*, 147–180.
Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59.
Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*(4), 873–922.
Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99*(3), 518.
Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review, 111*(2), 333.
Rhodes, M. G., & Jacoby, L. L. (2007). On the dynamic nature of response criterion in recognition memory: Effects of base rate, awareness, and feedback. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(2), 305.
Ristic, B., Molinaro, N., & Mancini, S. (2016). Agreement attraction in Serbian. *The Mental Lexicon, 11*(2), 242–276.
Rotello, C. M., Heit, E., & Dubé, C. (2015). When more data steer us wrong: Replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin & Review, 22*(4), 944–954.
Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t*–tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237.
Ryle, G. (1949). *The concept of mind.* London: Hutchinson.
Schlueter, Z., Parker, D., & Lau, E. (2017). *(Mis)interpreting agreement attraction: Evidence from a novel dual-task paradigm. Talk given at the 30th CUNY Conference on Human Sentence Processing.* MIT.

Schlueter, Z., Williams, A., & Lau, E. (2018). Exploring the abstractness of number retrieval cues in the computation of subject-verb agreement in comprehension. *Journal of Memory and Language, 99*, 74–89.
Shen, E. Y., Staub, A., & Sanders, L. D. (2013). Event-related brain potential evidence that local nouns affect subject–verb agreement processing. *Language and Cognitive Processes, 28*(4), 498–524.
Slioussar, N. (2018). Forms and features: The role of syncretism in number agreement attraction. *Journal of Memory and Language, 101*, 51–63.
Slioussar, N., & Malko, A. (2016). Gender agreement attraction in Russian: Production and comprehension evidence. *Frontiers in Psychology, 7*, 1651.
Solomon, E. S., & Pearlmutter, N. J. (2004). Semantic integration and syntactic planning in language production. *Cognitive Psychology, 49*(1), 1–46.
Starns, J. J., Ratcliff, R., & White, C. N. (2012). Diffusion model drift rates can be influenced by decision processes: An analysis of the strength-based mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(5), 1137.
Staub, A. (2008). *The computation of subject-verb number agreement: Response time studies* (Unpublished doctoral dissertation)University of Massachusetts Amherst.
Staub, A. (2009). On the interpretation of the number attraction effect: Response time evidence. *Journal of Memory and Language, 60*(2), 308–327.
Staub, A. (2010). Response time distributional evidence for distinct varieties of number attraction. *Cognition, 114*(3), 447–454.
Tanner, D. (2011). Agreement mechanisms in native and nonnative language processing: Electrophysiological correlates of complexity and interference. Unpublished Doctoral Dissertation. University of Washington.
Tanner, D., Dempsey, J., & Christianson, K. (2018). *Does attraction lead to systematic misinterpretation of NP number? Probably not. 31st CUNY sentence processing conference*.
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.
Tucker, M. A., Idrissi, A., & Almeida, D. (2015). Representing number in the real-time processing of agreement: Self-paced reading evidence from Arabic. *Frontiers in Psychology, 6*, 347.
Van Dyke, J. A., & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of Memory and Language, 65*(3), 247–263.
Vasishth, S., Brüssow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science, 32*(4), 685–712.
Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language, 103*, 151–175.
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.
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.
Vigliocco, G., Hartsuiker, R., Jarema, G., & Kolk, H. (1996). One or more labels on the bottles? Notional concord in Dutch and French. *Language and Cognitive Processes, 11*, 407–442.
Villata, S., Tabor, W., & Franck, J. (2018). Encoding and retrieval interference in sentence comprehension: Evidence from agreement. *Frontiers in Psychology, 9*, 2.
Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. *Experimental psychology, 60*(6), 385.
Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition, 32*(7), 1206–1220.
Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods, 39*(4), 767–775.
Voss, A., Voss, J., & Klauer, K. C. (2010). Separating response-execution bias from decision bias: Arguments for an additional parameter in Ratcliff's diffusion model. *British Journal of Mathematical and Statistical Psychology, 63*(3), 539–555.
Voss, A., Voss, J., & Lerche, V. (2015). Assessing cognitive processes with diffusion model analyses: A tutorial based on fast-dm-30. *Frontiers in Psychology, 6*, 336.
Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language, 61*(2), 206–237.
Wagers, M. (2008). *The structure of memory meets memory for structure in linguistic cognition* (Doctoral dissertation, University of Maryland College Park).
White, C. N., & Poldrack, R. A. (2014). Decomposing bias in different types of simple decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(2), 385.