SPEED LIMITS AND RED FLAGS:
WHY NUMBER AGREEMENT ACCIDENTS HAPPEN

BY

LAUREL ELLEN BREHM

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Doctoral Committee:
Professor Kathryn Bock, Chair
Professor Gary Dell
Professor Cynthia Fisher
Associate Professor Sarah Brown-Schmidt
Assistant Professor Darren Tanner

ABSTRACT

Trouble in language production sometimes surfaces in errors and sometimes surfaces in delays. Since these two symptoms of difficulty can trade off, theories may make predictions that are confirmed with measures of accuracy but disconfirmed with measures of speed, and vice-versa. In work on grammatical agreement in particular, there are accounts of variability in verb number production that emphasize the roles of lexical sources of number information and accounts that emphasize structural sources. Depending on whether speed or accuracy is measured these alternative views can differ in the success of their predictions. To evaluate the alternatives, we carried out six experiments gauging speed and accuracy together in producing agreement. The data were analyzed using a statistical method that integrates speed and accuracy into a coherent framework. The findings demonstrate that grammatical agreement mechanisms are substantially more sensitive to conceptual than to lexical forces, confirming a central hypothesis of a structural account of sentence production.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Speaking is a complex, multi-stage process that typically happens easily. However, mishaps can occur between thinking and speaking. By examining the ways in which typical speech breaks down, it is possible to examine which processes go in to fluent speech and what types of interactions occur between different cognitive systems. This research is concerned with deviations in number agreement, including those that create overt errors and those that simply disrupt fluency

Number agreement draws upon a variety of sources of information in language and cognition. Its implementation demands the integration and reconciliation of notional, conceptual, grammatical, lexical, morphological, and phonological information. This implementation also involves an implicit categorization of singular or plural and demands cognitive resources such as attention and working memory. The range of cognitive processes tapped allows grammatical agreement to serve as a tidy model for how speakers coordinate the production of entire utterances, providing insight on the interweaving of many cognitive processes. In addition, despite the typical success of its novel computations, the usage of agreement can and does depart from what a speaker intends in meaning, grammatical acceptability, and fluency. Set against successful agreement, these variations offer clues to the cognitive and linguistic culprits behind less-than-successful speaking in general.

Broadly speaking, the process of grammatical agreement involves linking an agreement *controller* (in the present sentence, the subject *the process of grammatical agreement* is a controller) and an agreement *target* (in the present sentence, the verb

involves is a target). The product of the linkage is that a controller and its target index the same value of an agreement *feature*, such as gender, person, or number (in the previous sentence, the subject and verb were both singular). Subject-verb number agreement is a familiar form of grammatical agreement in English, where typical controllers are nouns and typical targets include verbs, pronouns, and determiners.

Deviations from standard agreement patterns have fueled considerable psycholinguistic research. One frequently studied deviation is called *attraction,* in which the source of a verb's number is not the number of the subject per se, but the number of another part of the subject noun phrase (e.g. Bock & Miller, 1991). An illustration of attraction is a line from the talk-show host Conan O'Brien, who exclaimed "*The back of my pants* ***are*** *falling off*" (*Conan,* episode 201, January 23, 2012). Here, the subject is the entire phrase *the back of my pants*. Even though the phrase's referent is a notional singleton and the subject is grammatically singular in number, its singular number fails to materialize on the target verb. Instead, the verb is plural, seemingly reflecting the plural number of a subcomponent of the subject (*my pants)*. The position of *my pants* in the example typifies the erroneous number source in attraction errors, a noun phrase in the neighborhood of the verb (but not necessarily immediately preceding it). The attractor is often called the *local noun,* for the sake of simplicity.

Similar to a classic type of speech error, what has happened is that the production process failed to accurately encode the speaker's meaning because of a linguistic misfire (Dell, 1986; Garrett, 1980). Viewed as an error process, attraction is a product of interference between grammatically plural nouns and grammatically singular sentence subjects. There is little involvement of number meaning: Speakers ordinarily construe the

referent of *pants* as a notional singular (Bock, Eberhard, Cutting, Meyer, & Schriefers, 2001), yet its grammatical plurality causes attraction. In contrast, though *clothing* refers to a notional plural, it fails to attract (Bock & Eberhard, 1993; Bock, Eberhard, & Cutting, 2004) or attracts only weakly (Haskell & MacDonald, 2003).

Differences in number meaning due to notional ambiguity can nonetheless create variations in number agreement. With collective nouns, plural agreement changes in likelihood depending on the construal of the referent. For instance, a collective sentence subject like *The gang **on** the motorcycles* is easily construed in terms of distinct individuals, whereas *The gang **near** the motorcycles* is more likely to be construed as a single group. Consistent with this notional difference, the former is more likely than the latter to elicit plural verb agreement (Humphreys & Bock, 2005). Similar effects occur for distributivity. The phrase *The test for the students* can represent two construals, either a single test type (e.g. the abstract content of the exam) or several tokens of one type (e.g. a stack of exam printouts). The latter construal is distributive and elicits increased plural notional agreement (Eberhard, 1999; Vigliocco, Butterworth, & Semenza, 1995). Finally, the purest examples of notionally-sensitive agreement come from conjunctions of singular nouns. When conjoined noun phrases serve as sentence subjects, the verbs used with them can vary reliably between singular and plural, especially when the referents have a natural singleton construal (e.g. *sleet and freezing rain,* Lorimor, 2007).

Other agreement variations are connected with fluency. Even when implemented without error, some instances of agreement are accomplished faster than others (e.g., Brehm & Bock, 2013; Haskell & MacDonald, 2003; Staub, 2009). Even more noticeable are disfluencies such as *uh* or *um*, pauses, and restarts (e.g. Clark & Wasow, 1998; Clark

& Fox Tree, 2002). The (real) speakers who said

> "Any of these alternative classifications are..is..."
> "At least as far as this data...these data are concerned..."
> "The breaking of relations in themselves..in itself..."

appear to be wrestling with the links between number controllers and targets.

There are several accounts of agreement production that aim to explain these variations in terms of general language production mechanisms. These accounts can be classified as *lexical* approaches and *structural* approaches (Bock & Ferreira, 2014). From a lexical perspective, agreement variability stems from statistical information learned from specific lexical items. In contrast, from a structural perspective, agreement variability stems from information carried by abstract features used to compose structures. Though not mutually exclusive and clearly interdependent in processing, these perspectives do serve to define major lines of psycholinguistic debate and are thus useful as a starting point. The next section sketches how the two play out in current accounts of agreement.

Lexical and structural sources of agreement trouble

Lexical sources. Under lexical accounts, agreement attraction is explained in terms of difficulties in selecting and retrieving words. The architecture of the relevant processes is schematized on the left side of Figure 1. One lexical account supposes that in the course of retrieval, probabilistic morphological relationships between nouns and verbs support subject-verb number agreement (e.g. Haskell & MacDonald, 2003). Specifically, if agreement is driven by a speaker's experience in producing plural nouns with plural verbs, the retrieval of a plural noun can increase the probability of producing a plural verb regardless of the structural relationship between the noun and verb. The

result is that a verb's number may deviate from the value that would be expected in light of what the speaker intended to convey.

This is related to a classic view of attraction in which the speaker loses track of the intended agreement controller (Fowler, 1937). The target then takes its value from another element of the planned utterance, such as a noun or a phrase more accessible in memory. Accounts like this are captured in a variety of hypotheses about attraction (e.g. Badecker & Kuminiak, 2009; Haskell & MacDonald, 2003; Solomon & Pearlmutter, 2004; Thornton & MacDonald, 2003; Vigliocco, Butterworth, & Garrett, 1996; Vigliocco et. al., 1995; Vigliocco & Franck, 1999; Wagers, Lau, & Phillips, 2009). The supposition is that the agreement problem results from a breakdown during formulation between the speaker's intended idea and the selection of an agreement controller. Though Conan O'Brien may have intended *back of my pants* as the subject, the utterance he produced used *my pants* as the agreement controller.

The defining feature of a lexical version of controller confusion is that the confusion arises in the linkage between a verb and noun. An illustrative account was proposed by Solomon and Pearlmutter (2004). On this account, the core mechanism of attraction is competition between singular and plural nouns during lexical retrieval in sentence formulation. The likelihood of competition increases under circumstances that promote the parallel encoding of lexical information, allowing a noun other than the one that is conventionally associated with verb number to control agreement. A potential consequence of this is that the number of the local noun is reflected in the verb, rather than the number of the head noun of the subject phrase, as illustrated in *The phone with the missing buttons* ***were*** *black*.

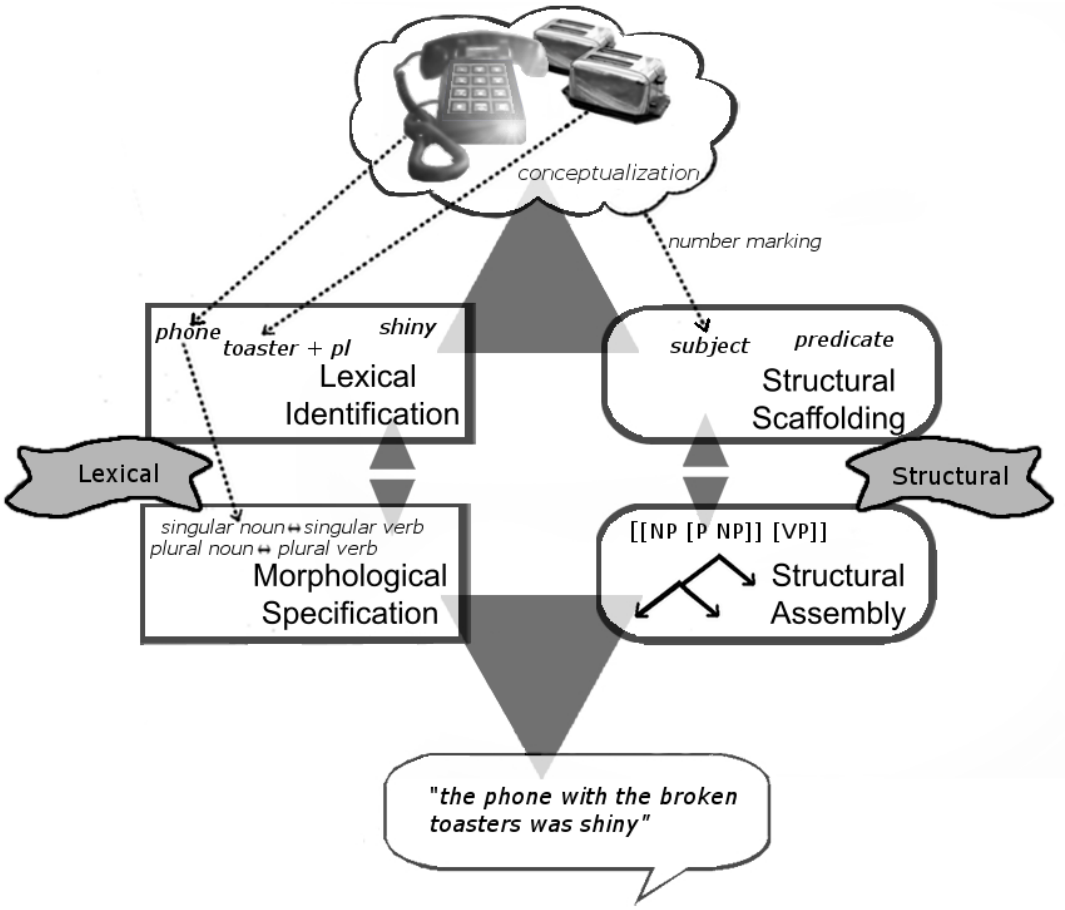


Figure 1. Hypothesized sources of lexical and structural difficulty in agreement. In this sketch, number information (dashed lines) passes from a conceptual level to sentence formulation processes. Plural notional number affects the abstract structural subject's number in parallel with lexical identification and morphological specification, which encode lexical concepts in words with appropriate grammatical number. Other processes are omitted for simplicity.

In several experiments, Solomon and Pearlmutter manipulated a notional property they expected to affect lexical retrieval, a property that we term *referential integration* (to emphasize its notional-number implications; note that Solomon and Pearlmutter's term was *semantic integration*). On the hypothesis that strong integration yields more overlap in the timing of retrieval for individual nouns, it can create a conflict that disrupts the usual relationship between noun and verb number. As a result, phrases constructed from

well-integrated mental models, such as *The phone with the missing buttons,* elicit more attraction than weakly-integrated phrases, such as *The phone with the broken toasters*.

Behind accounts like these there is a plausible mechanism called *content-addressable* memory retrieval. In a content-addressable system, retrieval cues serve to directly access relevant or activated information from memory (e.g. McElree, 1996; McElree, Foraker & Dyer 2003; Van Dyke & Lewis, 2003; see Dillon, Mishler, Slogget, & Phillips, 2013 for a discussion of memory retrieval in agreement comprehension). With prompting from a verb requiring number inflection (an agreement target), a controller may be retrieved from among the words being prepared for production. If the word is not the intended controller but nonetheless triggers the verb's inflection, attraction may ensue. In simplest terms, a high-probability lexical association between successive nouns (e.g. *pants*) and verb forms (e.g. *are*) masks the structural relationship between the intended subject and its predicate. The structural relationship loses its normal force.

Structural sources. In contrast, under a structural account of attraction, the normal structural relationship between subject controllers and verb targets is maintained, but the calculation of the controller's *number* goes awry and an abstract feature is mis-specified. Attraction arises in situations where the subject phrase remains a structured whole, with the intended structural relationship between subjects and predicates intact, but where the number feature of the subject is aberrant. The aberrant number feature is transmitted to the target verb in the ordinary way, with attraction as a possible result. The underlying problem isn't getting the subject (the controller) wrong, but getting the subject's number feature wrong.

Calculation of a number feature can vary because subject number is the product of

several different types of number information reconciled in terms of the subject as a structured whole. In principle, the reconciliation calls not only on the subject's lexical components, but on its notional, structural, and lexical number properties together. The crucial prediction is that the "wrong" verb number can result from a property associated with the subject's structure. Among these whole-structure properties are notional number (the construed number of subject noun phrase's referent; Bock et al. 2004; Humphreys & Bock, 2005; Lorimor, 2007) and lexical-grammatical number modulated by structural (rather than linear) distances among constituents of the phrase (Bock & Cutting, 1993; Franck, Lassi, Frauenfelder, & Rizzi, 2006; Franck, Vigliocco, Anton-Mendez, Collina, & Frauenfelder, 2008; Vigliocco & Nicol, 1998). These structural contributions to agreement are separable from lexical ones, as schematized on the right of Figure 1.

The structural approach to agreement and attraction is illustrated in the *Marking and Morphing* account (Eberhard, Cutting & Bock, 2005). In this model, subject number is a calculation that calls on notional and lexical-grammatical number, along with structural properties due to the subject's syntax. The product of the calculation is an abstract number feature, a *marking*. Preliminary to the process of marking, the referent of the subject noun phrase is evaluated for notional number as singleton or multiple. Marking occurs when notional number (a probabilistic, graded property of perceived and conceived sets) is interpreted as a discrete linguistic feature (singular or plural). The marked value combines with the grammatical number of words inserted into the subject phrase as they are retrieved, making a weighted contribution to subject number. These grammatical-number weights are determined in part by the subject's unfolding structural composition. The reconciliation of any conflicts between marked and lexical number

occurs during a *morphing* process in which morphological information is added to the planned structure, and the reconciled subject number then determines the grammatical number of the agreement target. Crucially, the marked value of the subject *as a whole* exerts its influence throughout morphing.

The competing predictions from lexical and structural accounts are nicely illustrated in how referential integration plays out in agreement. In terms of notional number, the referents of well-integrated sentence subjects (*The phone with the missing buttons*) are more notionally singular than weakly integrated ones (*The phone with the toasters*). Accordingly, weakly integrated subjects should be more likely to take plural verbs. Notice that this prediction is the opposite of the lexical prediction tested by Solomon and Pearlmutter (2004), and it finds support in experiments on both English (Brehm & Bock, 2013) and Dutch (Veenstra, Acheson, Bock, & Meyer, 2014). We return to this opposition later. The crucial point here is simply that these notional effects cannot be easily explained by properties of the words in the subject phrase alone.

Disentangling lexical and structural mechanisms

It is obvious that lexical and structural information are both necessary for the normal implementation of number agreement. What is less obvious is how these different sources of information have their impact. In order to separate and trace the two, these experiments were designed to assess covariations in the outcomes of agreement (singular or plural verb use) with latencies to produce number-agreeing verbs. We did this by manipulating factors attributable to whole-structure properties of sentence subjects (notional number) and factors associated with lexical properties of the same subjects (lexical-semantic relatedness between nouns and predicate adjectives in Experiment 1, 4,

5, and 6 and between subject nouns in Experiment 2). The grammatical number of local nouns was varied in order to create the conditions for attraction, conditions that are assumed in both lexical and structural accounts. This provides a tracer for the operation of factors that mediate the computation of agreement. To the degree that these factors are structural (characteristic of whole subjects), there is support for structural control of agreement; to the degree that these factors are lexical (characteristic of single words), there is support for lexical control of agreement.

Agreement as a cognitive process

In addition to requiring the choreography of structural and lexical information, agreement also counts upon a variety of other cognitive factors, many of which have been alluded to in earlier sections. These factors include number semantics, number cognition, memory, monitoring, and attention. The present experiments were designed to examine the way that these cognitive factors interact with the types of linguistic information described above.

Number and number agreement. There is a transparent link between number agreement and number. In particular, the categorization of nouns into singletons and multiples has a connection to counting number and number cognition: Things that come in units larger than one are plural, while things that come in units of one are singular. In order to assess the size of these sets, a speaker needs to implicitly categorize objects on 'how many'. For real-world objects, referents can be enumerated in two different ways. For small sets of objects (set sizes of approximately four for adults), we can subitize, or directly assess the number of things present. For larger sets of objects, subitization fails, and we need to enumerate, by either counting or approximating (e.g. Dehane, 1992).

Ability to do this singular/plural categorization and object enumeration are linked in development. The ability to deal with objects in sets larger than four and the use of number in language (e.g. use of quantification, ability to distinguish singular and plural) are developmentally associated (e.g Barner, Thalwitz, Wood, Yang & Carey, 2007; Li, Ogura, Barner, Yang & Carey, 2009).

Item grouping also affects both number agreement and number cognition, showing another link between the two. Research on distributivity (e.g. Humphreys & Bock, 2004) suggests that implicit mental representations of objects in space affects agreement: *The gang **on** the motorcycles* elicits more plural agreement than *The gang **near** the motorcycles*. Similarly, item grouping affects estimation of numerosity. Gestalt spatial properties seem to affect what objects go together (e.g. Wertheimer, 1923a,b), altering numerosity estimation. In particular, spreading items apart inflates numerosity judgments (e.g., Kruger, 1972), while perceptually joining them (e.g., with lines) makes numerosity estimation more accurate (e.g Franconeri, Bemis, & Alvarez, 2009).

Similar grouping abilities also occur directly within language, with the use of quantifiers to specify subsets of referents (e.g. Barwise & Cooper, 1981). Consider the differences between *the, one,* and *each*. All three agree with a singular verb (e.g. *the boy was, one boy was, each boy was)*. However, these three quantifiers have different assumptions of how many actors did how many actions, and the size of the comparison set. These properties of quantifiers have been demonstrated to affect number agreement (Eberhard, 1997), via grammatical and notional number. These parallels suggest links between the linguistic phenomenon of number and number cognition, and this was explored in Experiment 3.

Memory and comprehension. Current lexical accounts of agreement rely upon the cue-based retrieval of agreement controllers to explain patterns of agreement errors. This points to the importance of memory processes in number agreement and to the connection between memory and lexical factors. In particular, previous work suggests a strong role for memory retrieval in agreement comprehension (e.g. Dillon et al, 2013; Wagers et al, 2009), with more minimal support for the same in agreement production (e.g. Badecker & Kuminiak, 2009). Further evidence for a memory burden in comprehension relies upon misinterpretations in reading and listening. Previous work shows that comprehenders also make errors, and that these comprehension errors cause a listener to interpret an utterance as something more predictable, more common, or more correct (e.g Gibson, Bergen, & Piantadosi, 2013).

These lines of research point to important contrasts between comprehension and production, and to the importance of comprehension in production. Comprehension and production may not use lexical and structural information in the same way, and some of the errors attributed to production difficulty may instead be due to preamble comprehension difficulty. These questions are investigated in Experiments 4 and 5, which examine the role of agreement comprehension in the preamble completion task used in Experiment 1.

Monitoring for errors. Number agreement is a curious phenomenon because variation in plural verb production can come from either an error-based process (attraction, due to grammatical and lexical factors), or from a correct, non-prescriptive variation driven by the message (notional agreement, due to notional number). In order to address the differences between the two, we examine the relationship between agreement

and error monitoring. The idea is that the absence of errors is not necessarily indicative of the absence of difficulty: To the extent a speaker notices errors, they may edit errors out of their speech. Some previous research on speech errors invokes a monitor to describe patterns of phonological errors (e.g, Baars, Motley, & McKay, 1975; Motley, Camden, & Baars, 1981, 1982), though the exact role of this monitor is under debate (e.g Nozari & Dell, 2009). Investigating the role of monitoring in agreement therefore sheds light on the role of monitoring involving errors in speech (e.g. as discussed by Hartsuiker, 2006; Hartsuiker & Barkhuysen, 2006). Additionally, to the extent that speakers monitor for *errors* in particular dissociates notional agreement (not an error) and attraction (an error). This is the topic of Experiment 6.

Testing agreement production

In all the experiments in this document we used a standard paradigm for looking at agreement in language production. On each trial during the experiment, participants heard or saw a subject noun phrase (a preamble) that they had to complete into a full sentence. The preambles were designed to have notional, structural, and lexical properties that disrupt the normally reliable agreement process. We adapted the paradigm to control the final word of the completion produced, a predicate adjective, and to allow measurement of the latency to produce verbs. For example, at the beginning of a trial a participant might see the adjective *ringing* and then hear the preamble *The phone with the missing button*. The task was to add a completion to the preamble containing the adjective, beginning as rapidly as possible after the preamble's offset. This allowed the verb and its number to vary. Typically, participants used a completion containing a copula verb (*is, are, was, were*) and the designated adjective (e.g. *ringing).*

For purposes of measuring speech onset latencies, one drawback of completion paradigms is that they can be too successful at eliciting agreement errors (e.g., some conditions in experiments by Eberhard, 1999 and Thornton & MacDonald, 2003 yielded respective error rates of 31% and 23%). Response errors compromise the measurement of response latencies because error-prone responding contaminates reaction times with variability from the decision process (see Pachella, 1974). This clouds the interpretation of studies of agreement that have examined response latencies, which often exhibit the high error rates typical of the paradigm (e.g. Bock, Carrieras, & Meseguer, 2012; Brehm & Bock, 2013; Haskell & MacDonald, 2003). In some cases the response latencies track in the same direction as the error rates; in other cases they do not (e.g. Bock et al. 2012; Staub, 2009). But regardless of whether errors and latencies point to the same conclusions, the production mechanisms underlying responses may change depending on how sensitive speakers are to the likelihood of error. Accordingly, to interpret the implementation of agreement, we relied on a statistical method that combines measures of speed and accuracy in responding, the diffusion model (Ratcliff, 1978). This model and its application to agreement are discussed in the following chapter.

In Experiment 1, we examined the interaction of local-noun grammatical number (which triggers attraction) with structurally conveyed number (from referential integration, a source of notional number) and lexical-semantically conveyed number (mediated by semantic relatedness) in determining agreement attraction. The structural approach to agreement implies that whole-structure properties will be a strong force in attraction, exhibited in interactions between notional number and grammatical number. In contrast, the lexical approach to attraction predicts that variations in lexical processes will

be the chief determinant of attraction, exhibited in interactions between grammatical number and semantic relatedness.

In Experiment 2, we tested the same predictions using an analogous manipulation of local-noun grammatical number, along with structurally conveyed notional number (from abstract vs. concrete conceptual combinations) and lexical-semantic relatedness in a different syntactic structure (conjoined noun phrases). We used the same paradigm and analysis techniques as the first experiment. Notably, with conjoined subjects the conventionally correct verb number is a plural, with the consequence that local-noun plurality should promote correct responses rather than errors. The question again was whether this tendency would be modulated more by structural or lexical sources of number information.

In Experiment 3, we examined the way that two different sources of notional number, driven by visual and linguistic information, interact with grammatical number in agreement production. This investigates agreement interacts with numerosity information, and whether different sources of notional information influence agreement in the same fashion. In this experiment, spatial properties of object arrays were manipulated to convey an impression of larger or smaller object quantities and linguistic information was carried using quantifiers, which vary on notional and grammatical number. This experiment used a paradigm similar to the first two experiments, though with the addition of photos representing the preambles' referents.

In Experiments 4 and 5, we examined the influence of comprehension processes on the preamble completion task. This was to target the degree to which the agreement patterns observed in the previous experiments were due to issues in understanding

preambles and to examine differences between agreement comprehension and production. These experiments used the same stimuli as Experiment 1, preambles varying on notional, lexical, and grammatical number, but the paradigm was a button-press paradigm that does not involve overt speech. This paradigm is theorized to tap prediction in language. In Experiment 4, participants were forced to respond quickly, while in Experiment 5, they were allowed to take as much time as they wished to respond. Differences between these and Experiment 1 demonstrate how lexical and structural information use changes as the burden of difficulty depends more on comprehension.

In Experiment 6, we examined the role of monitoring in agreement. This was to assess the extent to which participants can notice and avoid their errors. In particular, this targets the separation of attraction and notional agreement. This experiment used the same paradigm as Experiments 4 and 5, and the same analysis technique as the other experiments, but with the addition of a grammar pretest for some subjects that made errors more salient. The idea was that this would encourage self-monitoring.

All experiments used a common analysis technique, diffusion modeling. Its methodology and the logic behind it are outlined in the following chapter.

CHAPTER 2: DIFFUSION MODELING

The projects discussed in this document revolve around combining speed and accuracy data in order to determine contributions of various sources of difficulty in agreement, including grammatical number, notional number, lexical factors, comprehension factors, and the presence of monitoring. The common thread is that in each of these, difficulty could be reflected in either response latencies or errors. To reconcile these measures and to examine how these manipulations affect the cognitive processes behind agreement, diffusion modeling will be used as an analysis technique.

Diffusion models use random drift processes to account for errors and reaction times simultaneously. By looking at the distributions of correct and error reaction times, as well as the proportion of errors produced in different experimental conditions, diffusion models can decompose the components of a decision into separate parameters. This allows the model to pull apart a general index of difficulty (*v)* and to separate the response properties that cause speed-accuracy tradeoffs (*a, z)*. In doing so, diffusion models segment reaction time into those aspects of response planning that apply to the decision process (*MDT)*, and those that are separate from it (*Ter)*. This makes diffusion modeling a robust, flexible method for examining response processes in two-choice decisions, such as those in memory retrieval (e.g. Ratcliff, 1978), lexical decision (e.g. Ratcliff, Gomez, & McKoon, 2004), response monitoring (e.g Nozari & Dell, 2009), perceptual learning (e.g Voss, Rothermund, & Brandtstädter, 2008), and semantic categorization (e.g. Vanderveckhove, Verheyen, & Tuerlinckx, 2010).

At its heart, number agreement in English is a two-choice decision. Agreement targets must take inflections in order for speakers to produce grammatical sentences, and

these inflections are either singular or plural. This means that when a speaker performs agreement, they must decide whether to use singular or plural number. Current psycholinguistic models of agreement production describe this process as a reconciliation of grammatical and notional number information (e.g., Marking and Morphing, Eberhard et al, 2004), or as a reconciliation of multiple constraints that suggest singular or plural number (e.g. Thornton and MacDonald, 2003, Haskell et. al., 2010). The common thread in both is that agreement is a reconciliation process occurring in real-time during running speech, like other types of tasks modeled with diffusion modeling.

As a process model, a diffusion model also has great appeal for agreement. This model was developed to explain memory retrieval phenomena, separating out processes of perception and articulation from the key retrieval components of a memory decision (Ratcliff, 1978). Number agreement production also represents a set of retrieval and composition processes, and these are embedded in another task, the task of planning and articulating speech. As such, it is incredibly useful to be able to partition reaction time into that which is consumed by agreement, and that which is consumed by other processes. A diffusion model does just this, separating out non-decision time (*Ter)* and providing a clean measure of the process of interest on its own.

The diffusion modeling technique separates forced-choice decisions into several components, all of which have a cognitively-feasible interpretation. For the purposes of modeling agreement, we assume that three parameters correspond to aspects of agreement itself and monitoring during agreement (see Figure 2). These allow the separate examination of different processes, as outlined below.

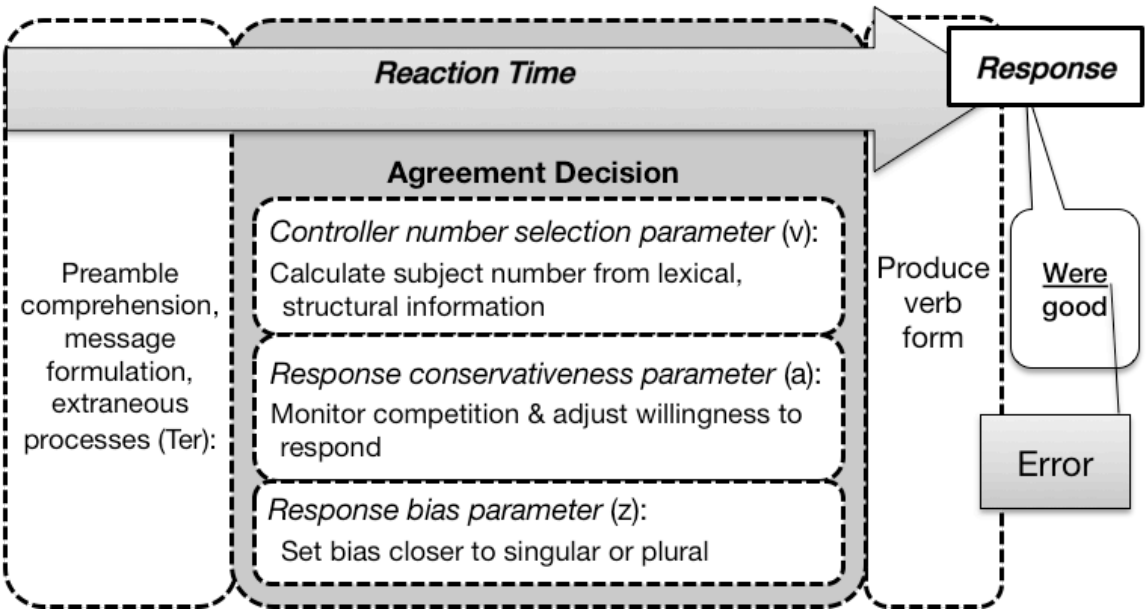


Figure 2. Schematic of evidence used in producing subject-verb number agreement and parameterization of diffusion model.

The parameter expected to track core agreement processes most closely is evidence strength *(v;* glossed as *controller number selection*). This parameter, a general index of experimental difficulty, reflects the accumulation of information towards a singular or plural response over time. Two other parameters account for monitoring and participant strategies. *Response conservativeness* (*a*) is typically indicative of participant-level differences, such as strategies related to task difficulty and particpant age (e.g. Ratcliff, Love, Thomspon, & Opfer, 2012). In this task, this may reflect strategy shifts in monitoring verb-number competition or grammatical correctness. The remaining decision-related parameter, *response bias* (*z)*, indexes the bias toward producing one of the two responses (e.g., correct singular or erroneous plural). What is critical is the relationship between *z* and *a,* or *relative response bias (z/a)*. Relative response bias of 0.5 indicates an unbiased starting point, and biases approaching 1 and 0, respectively, indicate a starting point biased towards corect singular or erroneous plural responses.

Previous research has shown this bias to be affected by the differing weights of positive and negative evidence (e.g. Voss et. al., 2008), and as such, it may reflect inherent asymmetries in the cognitive system, such as the markedness of grammatical plurals (e.g. Greenberg 1966).

The final parameter in the diffusion model allows for the removal of unwanted variance from the process of interest. This parameter is known as *non-decision time* (*Ter)* and captures the extraneous processes shown in Figure 2. For agreement, this corresponds to the segment of reaction time left after accounting for controller number selection. In our task, we assume that non-decision time will index a variety of processes, including preamble comprehension, message conceptualization and response articulation.

There are a variety of ways to perform diffusion modeling. In the current projects, diffusion modeling was done with the fast-dm program (Voss & Voss, 2007, 2008), a method that was selected rather than any alternative (e.g. EZ-diffusion, DMAT) as it can fit all of the relevant diffusion parameters (in contrast to EZ-diffusion; Wagenmakers, van der Maas, & Grasman, 2007) with no lower limit of the number of responses required per cell to fit the data (in contrast to the Ratcliff chi-square minimization method, e.g. DMAT; Ratcliff & Tuerlincx, 2002). This makes fast-dm more amenable than other methods to the sorts of sparse data inherent to sentence production, providing another advantage over more traditional analysis methods.

To compensate for the low error rates in certain conditions and the relatively few trials per participant, we estimated parameters separately for supersubjects in all experiments (Dell, Burger, & Svec, 1997; Ratcliff, Thapar, Gomez, & McKoon, 2004) rather than individual participants. Each supersubject was made up of four participants

who were run in the same experimental list (as in Konopka & Bock, 2009). Because the supersubjects were assembled from randomly grouped participants who saw identical materials, and all participants were sampled at random (or what passes for random in current practice) from the same pool, the inferences that follow from individual response patterns also hold for supersubjects.

For all experiments, several versions of the model were run in order to assess the contribution of variations in each of these four parameters. The model versions included a full model, a minimal model, and three models of intermediate complexity. The full model allowed *v, a, z*, and *Ter* to vary by condition and supersubject. The minimal model allowed only *v* to vary by condition and supersubject, allowing the other parameters to vary by just supersubject. The other models eliminated parameters that previous work has suggests are better left fixed by condition (e.g. Ratcliff & McKoon, 2008): One tested whether response criterion (*a)* needed to vary by condition (vs. fixed criterion), another tested whether response bias (*z)* needed to vary by condition (vs. fixed bias), and a third tested whether non-decision time (*Ter)* needed to vary by condition (vs. fixed *Ter*). In all models, there were scaling and error parameters. The error parameters (s_z, s_v, s_{Ter}) were allowed to vary by supersubject only and the scaling parameter *s* was set to 1. As well, the model's precision (an accuracy term for the model) was set to Voss and Voss's default of 3 (2007, 2008).

Model selection was based upon two curve-fitting standards. The first was calculated directly in the fast-dm program and used the Kolmogorov-Smirnov (K-S) test, a nonparametric test based upon the largest vertical difference between the cumulative density function (CDF) of the modeled (predicted) data and the empirical CDF provided

by the data. Larger K-S *p*-values signify a higher probability that the model precisely fits the empirical data. A K-S *p*-value was generated for each supersubject, for each version of the model that was run. See Appendix C for plots of K-S *p*-values by supersubject, by model, by experiment.

The other curve-fitting standard involved the comparison between the observed data and values predicted from running the model backwards. This was achieved by running the modeled parameters through the plot-CDF script in the fast-dm program. The plot-CDF script was used to output predicted curves for each supersubject, each condition, and each model that was run. Curve values were calculated at three critical points for the combined CDF created in the fashion of Voss and Voss (2007, 2008), which involves mirroring the error response latencies (treating them as negative), and then computing the cumulative density function over all responses combined. The three critical points corresponded to the median error RT (time at the median density value where time was negative), the median correct RT (time at the median density value where time was positive), and the proportion of error responses (the y-intercept of the combined CDF). See Appendix D for plots of these values by experiment.

After determining the best-fitting models, parameter values were assessed using mixed effect modeling, in order to determine the relative contributions of the fixed effects to the various parameters in the model. Reported results for all experiments will focus on the primary process of interest, controller number selection (*v*) and its relations to the other parameters.

CHAPTER 3

EXPERIMENT 1: SOURCES OF DIFFICULTY IN COMPLEX NOUN PHRASES

The first experiment was designed to examine variations in number agreement due to lexical and structural factors. The variables examined were the sentence subjects' local-noun number, referential integration, and lexical-semantic compatibility between the required predicate and the head or local noun. For example, with a referentially well-integrated subject like *The phone with the missing button[s]*, the required predicate could be either *ringing* (compatible with *phone* but not *ringing*) or *plastic* (more compatible with *button* than with *phone*), to elicit the completions in (a):

(a) *The phone with the missing button[s] was/were ringing.*

The phone with the missing button[s] was/were plastic.

Alternatively, with a referentially less integrated subject like *The phone with the broken toaster[s],* the required predicates could be either *ringing* or *shiny*, to elicit the completions in (b):

(b) *The phone with the broken toaster[s] was/were ringing*

The phone with the broken toaster[s] was/were shiny.

Use of the singular verb *was* is conventionally correct, whereas *were* indicates an attraction error. The measure of attraction is the difference between singular and plural local nouns in the frequency of using *were*.

The two accounts of agreement make differing predictions with regard to how often attraction occurs from variations in referential integration and predicate compatibility. The structural account predicts an interaction between local-noun number and the whole-subject property of integration, since subject number is more likely to be

singular with integrated than with less-integrated preambles. The lexical account instead predicts an interaction between local-noun number and predicate compatibility, since the differing strengths of the lexical-semantic relationship between the designated predicate and the head or local noun should create conflicts in controller selection. If both factors interact with local noun number, the implication is that both structural and lexical variations are important to the implementation of agreement.

Method

Participants. In exchange for course credit or $7.00 compensation, 165 undergraduates from the University of Illinois participated in the experiment. Participants were excluded if they had fewer than 66% usable experimental trials (N=21) or were non-native English speakers (N=3). An additional 13 participants were excluded due to technical difficulties (N=4) and counterbalancing errors (N=9). This left 128 participants.

Equipment. Stimuli were presented using PsyScope X B53 (Cohen, MacWhinney, Flatt, & Provost, 1993) on a Macintosh Mini computer with a 17-inch LCD flat-screen monitor. Audio was presented to participants over Koss headphones, and their speech was digitally recorded to a computer using a Sennheiser directional microphone run through a USB button box and Tube MP preamplifier. PsyScope recorded the latency of vocal responses through the button box.

Materials. There were 24 experimental items, all based on those with the highest integration differences in Solomon and Pearlmutter's ratings (2004, Experiment 4). These items were designed to serve as sentence subjects (*preambles)* and were made up of complex noun phrases. Preambles varied in integration (integrated, unintegrated). All had singular heads and local noun phrases that varied in grammatical number (singular,

plural). Preambles were paired with predicate adjectives that differed in their likelihood of modifying the head (head compatible) or local noun phrase (local compatible). This yielded eight versions of all 24 items varying in integration, compatibility, and local noun number. See Table 1 for an example and Appendix A for a full list of stimuli.

Table 1
Example stimuli from Experiment 1.

	Preamble	*Head-compatible*	*Local-compatible*
Integrated	The phone with the missing button(s) (head) (local)	ringing	plastic
Unintegrated	The phone with the broken toaster(s) (head) (local)	ringing	shiny

To construct the lexical-semantic compatibility manipulation for each item, candidate adjectives were rated for their fit with the head and local nouns. For head compatibility, selected adjectives were judged to be better modifiers of the head noun (e.g. *phone)* than either local noun (e.g. *button, toaster)*. For local compatibility, selected adjectives were judged to be better modifiers of the local noun (e.g. *button* or *toaster)* than the head noun (e.g. *phone)*. The selections were made on the basis of paper-and-pencil ratings in which judges (N=between 5 and 21 per item) assessed the likelihood of a given adjective modifying the accompanying noun. Statements were presented in the following format:

How likely is it that a phone is ringing? Not likely--1 2 3 4 5 --Very likely

Noun-adjective pairings were presented in lists in which no noun or predicate was repeated. Integration was fully balanced within these lists, but only the singular form of the local nouns was presented. Ratings were iterated until adjectives with the appropriate

biases (head > local or local > head) were identified.

For the final set of adjectives, the overall compatibility advantage in the head-compatible condition for head over local nouns was 1.08; in the local-compatible conditions the overall advantage for local over head nouns was 1.14. Mean ratings of compatibility in each condition are shown in Table 2. Any differences between means larger than 0.80 fall outside of the margin of error (the half-width of the 95% confidence interval for differences between condition means) calculated based on the mean-squared error of the highest-level interaction in a repeated measures ANOVA by items, using the Scheffé correction and type III sum of squares (more reliable for unbalanced designs).

Table 2
Mean lexical-semantic compatibility ratings for Experiment 1. Margin of error for differences between means = .80.

Predicate adjective	Preamble Nouns (e.g. "*The phone with the missing button/broken toaster"*)		
	Head noun *e.g."phone"*	Integrated local *e.g. "button"*	Unintegrated local *e.g. "toaster"*
Head-compatible "ringing"	3.81	2.91	2.55
Integrated local-compatible "plastic"	2.55	3.52	-
Unintegrated local-compatible "shiny"	2.47	-	3.77

In addition to the experimental items, there were 61 filler stimuli. These were designed to increase the variability of sentence types shown in the experiment and to balance the positions of singular and plural nouns across items. Filler preambles included prepositional phrases, conjoined noun phrases, and simple noun phrases. These were

paired with predicate adjectives different from those in the critical trials. Predicate adjectives for the filler items were determined using acceptability judgments from the author and a research assistant. Of the fillers, 14% took singular agreement, so that 42% of all stimuli required singular agreement.

Eight lists were created from the eight versions of each critical preamble and the fillers. Each list contained one version of each experimental item and all the filler items. List order was determined quasi-randomly, with fillers in fixed positions across lists and critical items assigned randomly to slots between fillers. Ordering was constrained so that no more than two experimental items appeared consecutively and no semantically similar items were adjacent. Experimental items were counterbalanced so that every item was represented once and only once on each list, with an equal number of item versions in each condition. Each list was also divided into two halves, with the order of the halves counterbalanced over participants, for a total of sixteen lists. Every list began with thirty of the filler items in order to form a covert practice block.

Integration and sensibility norming. Norms were collected for the integration and sensibility properties of the final set of experimental items in paper-and-pencil tasks. To establish that the integration difference remained for all adjective-head combinations, 16 participants who did not participate in the main part of the experiment were asked to rate the integration of the preambles combined with their predicates. For this task, participants saw the preambles for each of the 24 items in one of four versions, with integration and lexical-semantic compatibility fully crossed, and only the singular local nouns used. Items were presented in the following format:

The <u>phone</u> with the missing button was ringing Not linked--1 2 3 4 5 6 7--Very linked

Task instructions and examples were adapted from Solomon and Pearlmutter (2004). The instructions emphasized referential integration with examples in which the words *ketchup* and *mustard* in the phrase *The ketchup and the mustard* are described as only weakly linked, whereas the words *bracelet* and *silver* in the phrase *The bracelet made of silver* are described as strongly linked (see Appendix B for the complete instructions). Items were presented to participants in one of four lists, so that each participant received only one version of each item and an equal number of version-types in each condition.

To check variations in sensibility for each preamble-adjective combination, sensibility norms were collected from 32 additional participants. The same lists were used as in the integration norming, with participants asked to rate the likelihood of statements such as the following:

The phone with the missing button was ringing Not likely-- 1 2 3 4 5 6 7--Very likely

Full instructions are shown in Appendix B.

Ratings from both tasks were analyzed using a by-items repeated measures ANOVA, with the Scheffé correction for multiple comparisons and margin of error for differences between means calculated from the mean square error of the highest-level interaction. Table 3 shows the results. Integration ratings differed for complete sentences in the integrated and unintegrated item versions (integrated M=5.98, range 4.5 to 6.5; unintegrated M=2.42 range 1.13 to 4.00), but not for complete sentences in the head-compatible-predicate and local-compatible-predicate item versions (head-compatible M= 4.18, range 3.00 to 5.13; local-compatible M= 4.21, range 3.38 to 5.63). The integration ratings of items collapsed across lexical-semantic compatibility were highly correlated with the ratings of Solomon and Pearlmutter ($r(46)=.89$), and the ratings for the head-

and local-compatible item versions were highly correlated with each other ($r(46)=.90$). This suggests that for this set of materials, compatibility differences did not change the relative levels of integration of the completed sentences.

For sensibility norming, the completed sentences had an average rating of 4.14. Within levels of integration, differences in whole-sentence sensibility between head- and local-compatible predicates were roughly comparable, .88 for integrated and .33 for unintegrated, against a margin of error of 1.39. This indicates that the relative *sensibility* of the completed sentences was approximately the same at both levels of integration, meaning that whole-sentence sensibility differences did not compromise integration.

Sensibility did vary depending on combinations of compatibility and integration, but in an unsurprising way. Overall, the completed sentences for local-compatible integrated item-versions were rated as more sensible than those for head-compatible unintegrated item versions, as a logical consequence of a predicate's relationship to parts and wholes. For an integrated referent, a property of one of its parts (denoted in the local-noun phrase) must also be a property of the whole: The plastic button is an attribute of the phone. For an unintegrated referent, however, a property of a part is not necessarily a property of the whole: The shiny toaster is not an attribute of the phone.

*Preamble recording***.** Audio stimuli were recorded in a quiet room on a Sennheiser directional microphone run through a Tube MP preamplifier. The talker was a woman from northern Illinois. She produced the phrases in the carrier phrase *"The next sentence is X disappeared yesterday*" (e.g. *The next sentence is the phone with the missing button disappeared yesterday*). These carrier phrases were removed from the audio files. The

stimuli were edited using Audacity to shorten continuants and pauses between words, in order to increase the speech rate while keeping natural-sounding stimuli.

Table 3
Mean integration and sensibility norming ratings for Experiment 1. Right-most column contains the margin of error of differences between means for the rating.

		Predicate		
Norming task	Integration level	*Head-compatible ("ringing phone")*	*Local-compatible ("plastic button/shiny toaster")*	Margin of error
Integration	*Integrated "The phone with the missing button"*	5.90	6.05	1.10
	Unintegrated "The phone with the broken toaster"	2.45	2.38	
Sensibility	*Integrated "The phone with the missing button"*	4.18	5.06	1.39
	Unintegrated "The phone with the broken toaster"	3.50	3.83	

Procedure. The procedure was a modified version of a preamble completion paradigm. There were two types of trials, standard trials and catch trials (see Figure 3). The standard sequence (Figure 3a) was used for all experimental trials and a subset of the fillers. Trials began with a fixation cross presented for 500 ms, placed 10% of the screen width relative to the screen's left margin, midway between top and bottom. The predicate adjective then appeared in the same location for 200 ms in 36-point lowercase black Arial font. Next, the preamble was presented over headphones. Immediately after the offset of the preamble, the cue "!" appeared at the center of the screen for 500 ms, prompting the

participant to speak. A blank screen then appeared for 2 seconds, giving participants a total of 2.5 seconds to respond with a complete predicate (e.g., "*was ringing*").

Catch trials were included to encourage participants to attend carefully to preambles on every trial. The catch-trial sequence (Figure 3b) occurred on 31% of the filler trials. Events were the same as on standard trials up to the point at which the response was cued. At that point, participants were prompted to repeat the preamble before completing it with a predicate. This was signaled with the word *Repeat* in place of the exclamation point. After 500 ms a blank screen appeared and remained for 3.5 seconds, giving a total of 4 seconds for a response (e.g. "*The rubber ducky was cute*"). Catch trials were composed of simple noun phrases, conjoined noun phrases, and noun phrases modified by a prepositional phrase.

Participants were instructed to complete all preambles as quickly and accurately as possible. They received two explicit practice trials, one of each type, and were given the opportunity to adjust the volume of the audio to a comfortable level. Before starting the experiment, they were queried about their understanding of the procedure. The experimenter stayed in the room for the entire session.

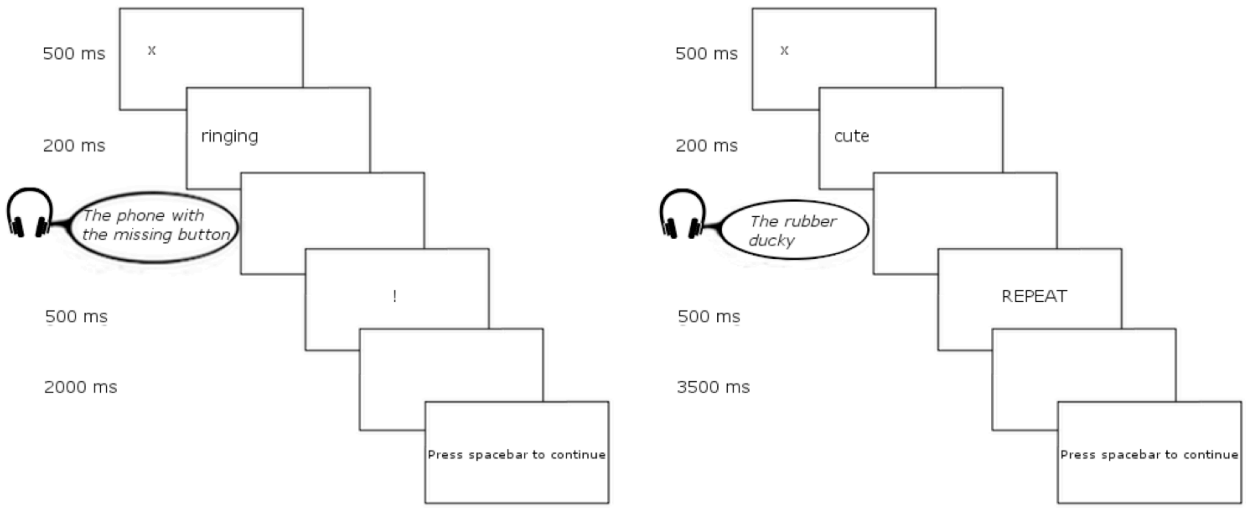


Figure 3. Trial sequences for standard trials (left) and catch trials (right) in Experiment 1.

Scoring. Responses on critical trials were scored as valid, miscellaneous, or missing. Valid responses consisted of an inflected form of the copula (*was, is, were, are)* and the correct predicate adjective for the trial, with no additional modifiers (e.g. *very* or *really)* or corrections to responses, and no disfluencies or non-speech noises before the verb. Valid responses were scored as singular or plural according to their verb number. Only valid trials were submitted to analyses.

Design. Each participant received one of the sixteen lists, each list containing one version of all 24 items, three in every condition, and counterbalanced by experiment half. Every list was presented to 8 participants, so that every item was tested on 32 participants. The fixed effects in the statistical analyses were integration (integrated-unintegrated), lexical-semantic compatibility (head-local), and local-noun number (singular-plural), all fully crossed.

Analysis. Diffusion models were fitted using the fast-dm program, as outlined above. Five candidate models were run, and the two model fitting procedures outlined in Chapter 2 were performed. The full model had the best fit by the K-S *p*-value maximal distance criterion (average K-S $p = .82$), followed by the fixed-criterion model (average K-S $p = .79$), the fixed *Ter* model (average K-S p-value = .75), and the fixed bias model (average K-S $p = .73$), with the minimal model trailing far behind (average K-S $p = .15$) (See Appendix C). Turning to the observed and fitted data curve fitting criterion, the full model and the three models eliminating only one parameter predicted fairly minimal differences between the empirical data and the results predicted from the model, with some small trade-offs in accuracy of estimating the proportion error and the two median RT values. See Appendix D for a plot of these values.

Based on these standards, we report results from further analyses of parameters from the fixed-message planning time model: Though the full model has the highest K-S *p*-value, the fixed *Ter* model fits with the previous literature on diffusion modeling and predicts median RTs and mean error rates with similar accuracy to the full model (.79 vs .82). Parameter values were analyzed statistically using multi-level linear models in R with the package lme4 (Bates, Maechler, Bolker, & Walker, 2014; R Core Team, 2013, with random intercepts for super-subjects and random slopes for all within-subject factors. No random effects were fitted for items, as items were aggregated over in the diffusion analysis. Effects coding was used for fixed effect contrasts.

Results

Figure 4 summarizes the raw accuracy and speed measures across conditions. Overall, integrated preambles elicited more accurate, faster responses than unintegrated preambles (Integrated: mean correct RT = 804 ms, 10% plural responses; Unintegrated: mean correct RT = 827 ms, 17% plural responses). Local plural nouns, relative to singulars, created the standard attraction effect, with slower, less accurate responses (Local plural: mean correct RT = 832 ms, 21% plural responses; Local singular: mean correct RT = 800 ms, 5% plural responses). Lexical-semantic compatibility did not affect latencies (Local-compatible: 817 ms; Head-compatible: 814 ms), but did affect error rates (Local-compatible: 15% plural responses; Head-compatible 11% plural responses). Error response latencies (Table 4) showed minimal differences between levels of lexical-semantic compatibility and local noun number, but plural responses to integrated preambles were considerably slower than responses to unintegrated preambles.

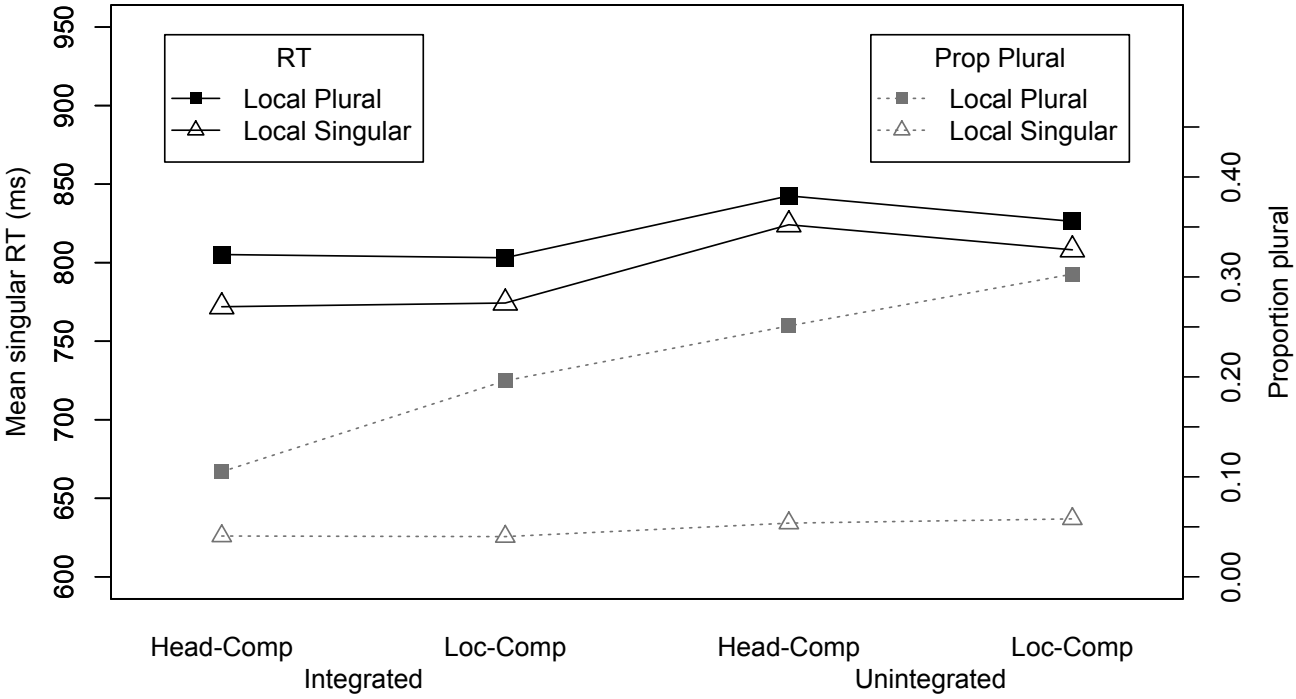


Figure 4. Experiment 1 singular verb production latencies (in ms; solid lines) and plural verb use (proportions; dotted lines) for preambles varying by integration, local noun number, and predicate compatibility

Table 4
Response latencies (in milliseconds) from Experiment 1.

Integration	*Predicate compatibility*	*Singular (Correct) Response*		*Plural (Error) Response*	
		Local Singular	*Local Plural*	*Local Singular*	*Local Plural*
Integrated	*Head-compatible*	770	789	893	890
	Local-compatible	781	788	776	928
Unintegrated	*Head-compatible*	819	829	927	828
	Local-compatible	808	817	833	852

The diffusion-model analysis of these data yielded the results shown in Figure 5. This figure shows effects of local number, integration and lexical-semantic compatibility on controller number selection (*v*, which we gloss as difficulty in computation of controller number). There was a general attraction effect, captured in the difference in difficulty between local singulars and plurals. This effect was systematically modulated by integration: In preambles with local plural nouns, integrated subjects drew less attraction than unintegrated (*v* = 1.42 and 1.88 respectively), for an attraction differential of .46. Lexical-semantic relatedness had a smaller impact (*v* = .88 vs. .57 respectively, for local vs. head compatibility), producing a nonsignificant attraction differential of .31.

These patterns were confirmed statistically with multi-level modeling (Table 5). The only significant interaction occurred between local-noun number and integration, while none of the interactions involving predicate compatibility reached significance. There was a marginal effect of compatibility alone (*p* = .09), consistent with a lexical-semantic effect that tended to occur after both singular and plural local nouns (i.e., not involving attraction).

Results for the remaining decision parameters (*a* and *z)* are not easily interpretable within existing views of agreement, and discussion of these will be postponed until Chapter 6. Previewing this, the primary outcomes indicate more conservative responding (*a)* for local-plural preambles and a bias (*z)* toward singular (correct) verb use. This bias was strongest in the integrated-local singular and unintegrated-local plural conditions.

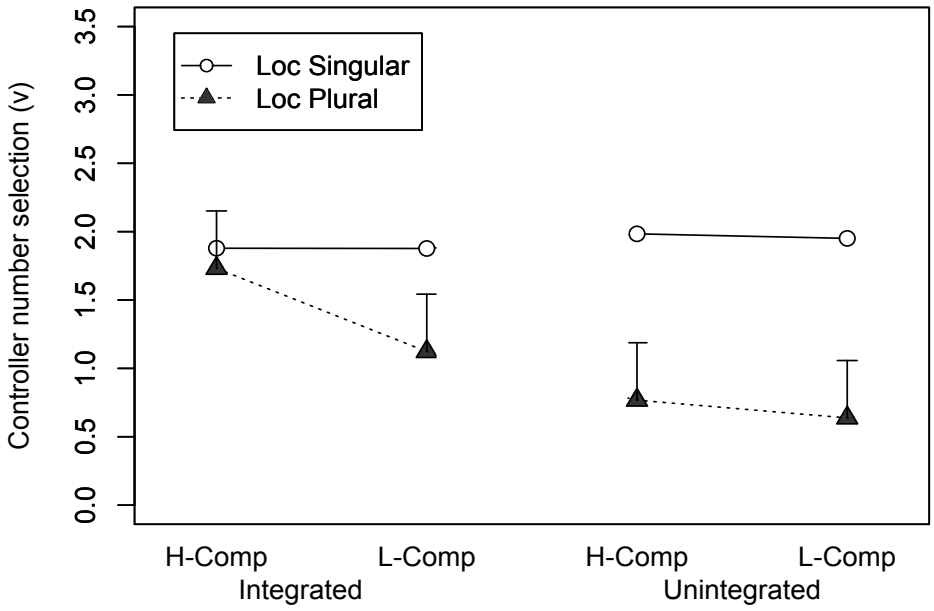


Figure 5. Fast-dm controller number selection parameter (v) from fixed *Ter* model for Experiment 1. Larger numbers reflect ease of agreement decision. Error bars represent margin of error for differences between means (0.59), calculated from the MSE of the highest-level interaction of a repeated-measures ANOVA by supersubjects.

Table 5
Experiment 1 parameter estimates for controller number selection (v) from fixed Ter diffusion model. P-values are approximated from a standard normal distribution.

	Estimate	*S.E.*	*t-value*	*p(z)*
Intercept	1.49	0.06	26.21	< 0.001
Local number	0.86	0.10	8.73	< 0.001
Integration	0.32	0.10	3.32	< 0.001
Predicate compatibility	0.19	0.11	1.70	0.09
Local number x integration	-0.81	0.22	-3.77	< 0.001
Local number x predicate compatibility	-0.35	0.23	-1.55	0.12
Integration x predicate compatibility	0.22	0.20	1.09	0.28
Local number x integration x predicate compatibility	-0.51	0.44	-1.15	0.25

Discussion

As expected, both local-noun number and integration affected the production of number agreement. This was reflected in raw measures (proportion plural responses, verb onset latencies) and in a measure derived to capture controller number selection. With local singular nouns, the calculation of singular (correct) verb number was uniformly easy. With local plural nouns, however, low levels of integration systematically disrupted the calculation of singular verb number. Lexical-semantic compatibility was less consequential, with a nonsignificant change in attraction due to differences between the head and local nouns in the compatibility of the predicate. Thus, attraction increased more with changes in integration than with changes in lexical-semantic compatibility.

In the raw measures of responding, inferences about underlying processes were complicated by the presence of a tradeoff between speed and accuracy (see Figure 4). Specifically, plural local nouns dramatically increased the use of grammatically incorrect plural verbs, but only slightly altered the latency of verb production over the singular-local noun baseline. To reconcile the two measures and arrive at a more coherent picture of how verb number is calculated, we used diffusion modeling (e.g., Ratcliff, 1978; Voss & Voss, 2007). This technique allowed us to more clearly define the contributions of lexical and structural information in agreement. The results of this analysis showed clear contributions of both notional number (in the form of referential integration) and grammatical number (in the form of local noun number) to the grammatical-number selection process. The interaction between notional number and local grammatical plurality points to a wholistic controller number that must be reconciled with the presence of plural grammatical number. This aligns with the structural account.

Although the influence of lexical-semantic compatibility was weak, the direction of the effect was similar to what Thornton and MacDonald (2003) found. Apart from its relative weakness, the only notable outcome of compatibility occurred in the raw error rates (Figure 4). Raw latencies were fairly immune to variations in compatibility. In contrast, variations in integration affected latencies as well as error rates.

In the second experiment we sought to generalize the implications of Experiment 1 by using different manipulations of notional and lexical properties. We did this with two well-studied variables, concreteness and semantic relatedness. In parallel to Experiment 1, these variables served to create contrasts in the effects of notional number and lexical processing on agreement.

CHAPTER 4

EXPERIMENT 2: SOURCES OF DIFFICULTY IN CONJOINED NOUN PHRASES

Experiment 2 was designed to further examine the contributions of structural and lexical components of processing to agreement, relying on the same method and analysis techniques as Experiment 1: We measured the speed and accuracy of producing number-marked verbs after preambles with varying notional and lexical properties.

The notional variable in Experiment 2 was concreteness and the lexical variable was semantic relatedness. These are factors that may play roles in agreement analogous to the constructs of integration and lexical-semantic compatibility in Experiment 1. High levels of concreteness, operationalized in terms of imageability, promote notional plurality due to the individuation of the referents, while low levels of concreteness do the opposite (e.g. Lorimor, 2007). Reducing concreteness is therefore expected to increase controller number selection difficulty due to notional number uncertainty. Relatedness is a strong promoter of lexical interference (e.g. Wheeldon & Monsell, 1994), eliciting competition due to shared semantic-category or associative features (Rahman & Melinger, 2007). Competition between concurrently processed nouns that differ in grammatical number has been argued to elicit attraction (e.g. Solomon & Pearlmutter, 2004). So, although Experiment 1 yielded only weak effects of lexical-semantic relatedness, the well-established impact of category relatedness on word production offers what could be a more potent and reliable source of disruption to agreement.

To enhance the potential for conflict between semantically related nouns, the experiment examined agreement with conjoined noun phrases. Conjunctions also allow variations in notional number depending on whether the referent is construed as a single

thing or event (e.g. Lorimor, 2007; Haskell & MacDonald, 2005), even though they are conventionally plural in English. This conventional plurality made it possible to explore notional and lexical effects with a different grammatically correct response than in Experiment 1-- plural verbs. The first noun in the experimental preambles was singular and the second noun varied between plural and singular, as in Experiment 1, but the plural local noun promoted the grammatically correct response.

Method

Participants. In exchange for course credit or $7.00 compensation, 113 undergraduates from the University of Illinois participated in the study. Of these participants, 13 were excluded from the study for having less than 84% unusable experimental trials (five or greater) and an additional four participants were excluded due to technical difficulties. This left 96 participants.

Equipment. Equipment was identical to Experiment 1, except that headphones were not needed.

Materials. There were 32 experimental items. These were conjunctions of nouns that varied in the imageability of their referents (concreteness), category relatedness (related or unrelated nouns), and local noun-number (singular or plural). See Table 6 for sample items. Stimuli were created using a free-association word database containing ratings for both nouns, including forward association (the number of times the second noun was generated when the first noun was prompted) and concreteness (Nelson, McEvoy, & Schreiber, 1998).

Abstract preambles contained nouns rated between one and three on a seven-point concreteness rating scale, while concrete preambles contained nouns rated between five

and seven on the same scale. Related and unrelated versions of these preambles were developed by changing the second noun, with related preambles having a forward association rating between 30% and 50% for the pair and unrelated preambles having a forward association rating between 0.01% and 10% for the pair. All but one of the nouns in the preambles had a regular plural form and a regular singular-plural alternation, regardless of their status as the head or local noun. The exception was the noun *tooth*, which occurred as a local noun. A full list of preambles is in Appendix A.

Table 6
Example stimuli from Experiment 2

	Related	*Unrelated*
Abstract	The hypothesis and the theory(ies)	The hypothesis and the thought(s)
Concrete	The dish and the plate(s)	The dish and the cat(s)

As in Experiment 1, filler stimuli were added to the critical preambles. The 192 filler preambles included a mixture of simple noun phrases and noun phrases modified by prepositional phrases. As in Experiment 1, some of the filler trials (18%) were catch trials. Among all fillers, 63% took singular agreement. This meant that the correct response was plural on 46% of all trials in the experiment.

The four different versions of each preamble were divided into four experimental lists. As in Experiment 1, they were counterbalanced such that all lists contained only one version of each item and an equal number of items of each type. The sequence of the preambles in the lists was determined in the same way as in Experiment 1, with the same

counterbalancing of presentation order, for a total of eight lists. All lists began with a fixed set of 12 fillers as a covert practice block.

Norming. Norming was carried out to establish the fit of the preamble nouns with a set of suitable adjectives to use as predicates. The adjectives were the ones used in Brehm and Bock (2013): *good, bad, ready,* and *true.* These adjectives were judged in similar fashion as the sensibility norming in Experiment 1, with the singular form of each noun presented with each possible adjective in the phrase *How likely is it that X is Y* (e.g. *How likely is it that a cat is bad?).* Items were divided into four lists, each with ten instances of all four adjectives counterbalanced across lists, and with each list containing one token of every noun. Items were presented in a fixed random order with no fillers.

Ratings were collected from 20 participants using paper and pencil surveys or a computer-presented Excel workbook. We calculated the average rating for the head and local nouns in a phrase across all adjectives, and for the highest-rated adjective for each of the nouns in an item's conjunction. Average ratings are displayed in Table 7. As in Experiment 1, margins of error were determined from the MSE of the highest-level interaction in an ANOVA by items with Scheffé corrections. The ratings suggest that as a whole, there was at least one well-fitting adjective for each isolated noun. The abstract pairs were rated generally as better fits with the adjectives than the concrete pairs on both metrics, but differences were small.

Norming was also carried out to establish the relationship between the dependent measures and integration. This used the same instructions and procedure as the integration rating in Experiment 1, though preambles were presented as noun phrases rather than as complete sentences. Both within-item variables (relatedness, local

plurality) were fully crossed, divided across four lists with items presented once per list. Ratings were collected from 20 participants and results are displayed in Table 7. High relatedness and low concreteness were associated with higher integration ratings, but ratings were equivalent for singular and plural local nouns. Importantly, the difference in integration ratings between levels of relatedness was similar for abstract and concrete items (abstract = 0.8; concrete = 0.9), suggesting that relatedness and concreteness contributed independently to integration judgments.

Table 7
Mean predicate sensibility and integration ratings for Experiment 2. Bottom row contains the margin of error for the rating.

		Sensibility		*Integration*	
Concreteness	Relatedness	*All adjectives*	*Best adjective*	*Local S*	*Local P*
Abstract "hypothesis"	*Related "theory"*	3.04	3.85	5.06	4.60
	Unrelated "thought"	3.03	3.92	4.05	3.95
Concrete "dish"	*Related "plate"*	2.82	3.68	3.99	4.44
	Unrelated "cat"	2.88	3.65	3.21	3.40
Margin of error		0.14	0.23	1.18	

Procedure. Preambles were presented visually for an interval equal to the longer of 1000 ms or 40 ms per character. Participants were instructed to read these silently and then complete them aloud with the best-fitting adjective from a memorized set of four adjectives, *good, bad, ready,* and, *true*. Previous work has shown a small memorized set

of adjectives to be effective for eliciting a relatively homogenous group of utterances suitable for use with a reaction time measure (e.g. Bock et al., 2012; Brehm & Bock, 2013). As in Experiment 1, there were standard (completion only) and catch (repeat and complete preamble) trials, with cuing, timing, and the sequence of events the same as in Experiment 1 (see Figure 3).

Scoring. Scoring was the same as in Experiment 1, except that plural verbs were scored as correct and singulars as incorrect.

Design. The variables of relatedness (related-unrelated), concreteness (abstract-concrete), and local-noun number (singular-plural) were fully crossed in a mixed design. Every participant received one version of each of the 32 experimental preambles, divided equally across the eight combinations of relatedness, concreteness and local–noun number. Every item in one of its versions was presented to 24 participants.

Analysis. Diffusion modeling was performed as in Experiment 1 using the fast-dm program with identical settings. Five candidate models were again fitted by supersubjects made up of four participants each, as outlined in Chapter 2. The full model had the largest average K-S *p*-value (.86), followed by the fixed criterion model (.80), the fixed *Ter* model (.78), the fixed bias model (.78), and the minimal model (.17). See Appendix C for distributions of these values. Again, comparison between observed and fitted data was done using the plot-CDF function. All models revealed small trade-offs in accuracy of estimating the proportion error and the two median RT values, with the minimal model performing the least well (See Appendix D). As in Experiment 1, we report the results from the fixed *Ter* model, as it had accuracy comparable to the full model and meshes with the previous diffusion modeling literature. Parameter values were

statistically evaluated with mixed-effect models containing random intercepts for supersubjects and random slopes for all fixed factors. Effects coding was used for fixed effect contrasts.

Results

Figure 6 shows mean accuracy and speed results across conditions. Overall, concrete preambles increased plural responding (concrete, 89% plural; abstract, 84% plural), as did the presence of local plural nouns (local singular, 79% plural; local plural, 95% plural). There were no accuracy differences due to semantic relatedness (related, 88% plural; unrelated, 88% plural). In response latencies, neither correct nor error responses revealed effects of any factor (Table 8).

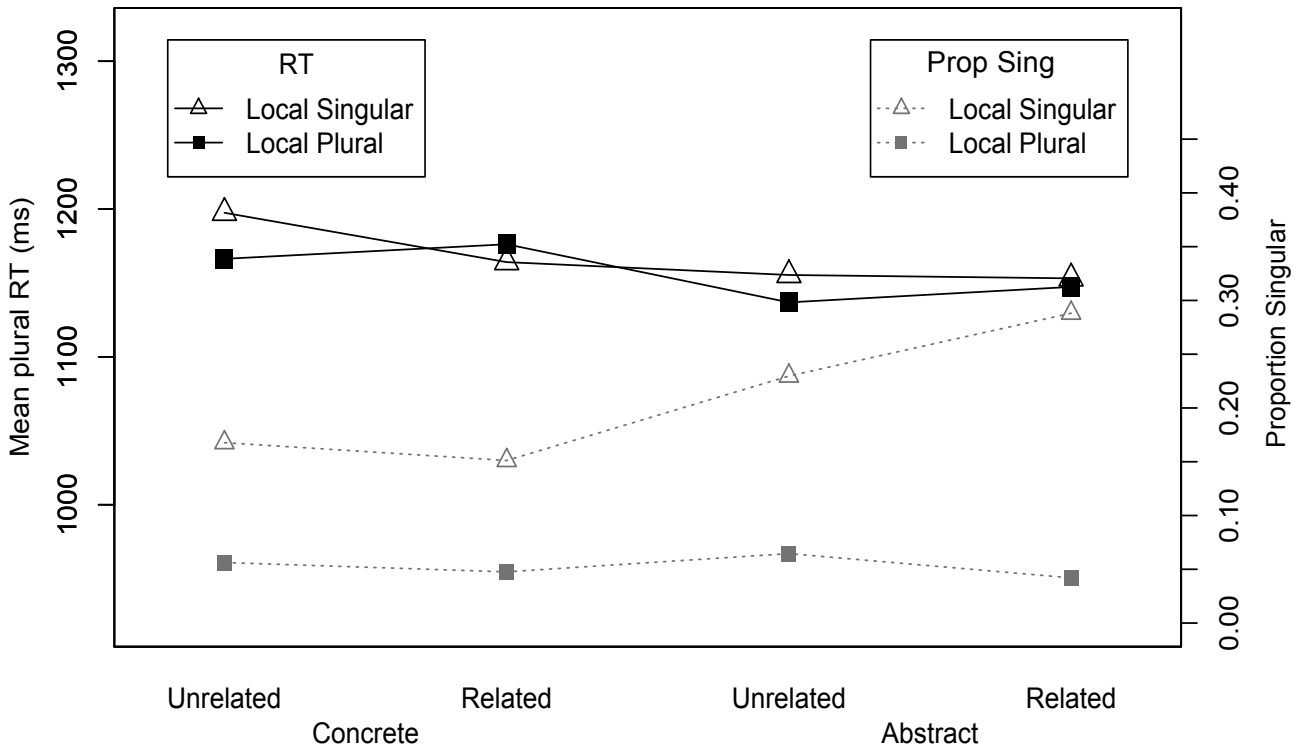


Figure 6. Experiment 2 plural verb production latencies (in ms) for preambles varying by concreteness, local noun number, and relatedness (solid lines); response tendencies for Experiment 2 by the same factors (dashed lines).

Table 8
Response latencies (in milliseconds) from Experiment 2.

Concreteness	*Relatedness*	*Singular (Error) Response*		*Plural (Correct) Response*	
		Local Singular	*Local Plural*	*Local Singular*	*Local Plural*
Abstract	*Related*	1243	1236	1150	1150
	Unrelated	1202	1315	1154	1137
Concrete	*Related*	1314	1203	1169	1176
	Unrelated	1264	1152	1202	1163

Figure 7 summarizes the results of the fixed *Ter* diffusion-model analysis for Experiment 2. The controller number selection parameter (*v*) showed main effects of concreteness and local number. Local nouns matching the expected verb response (local plural) were easier than verb mismatching local nouns (local singular), and concreteness also aided correct responding, with concrete pairs easier than abstract ones. The combination of notional singularity (abstractness) and singular local nouns increased difficulty slightly. Lexical relatedness had no evident effects, either alone or together with other factors. These effects were confirmed by multi-level modeling (Table 9).

Discussion of the results for the other two decision parameters, peripheral to the theoretical questions at issue here, are postponed until Chapter 6. In general, the response conservativeness parameter (*a)* showed that local singular nouns elicited less conservative responding, as did concrete pairs. The response bias (*z*) was neutral, unchanged by the experimental variables.

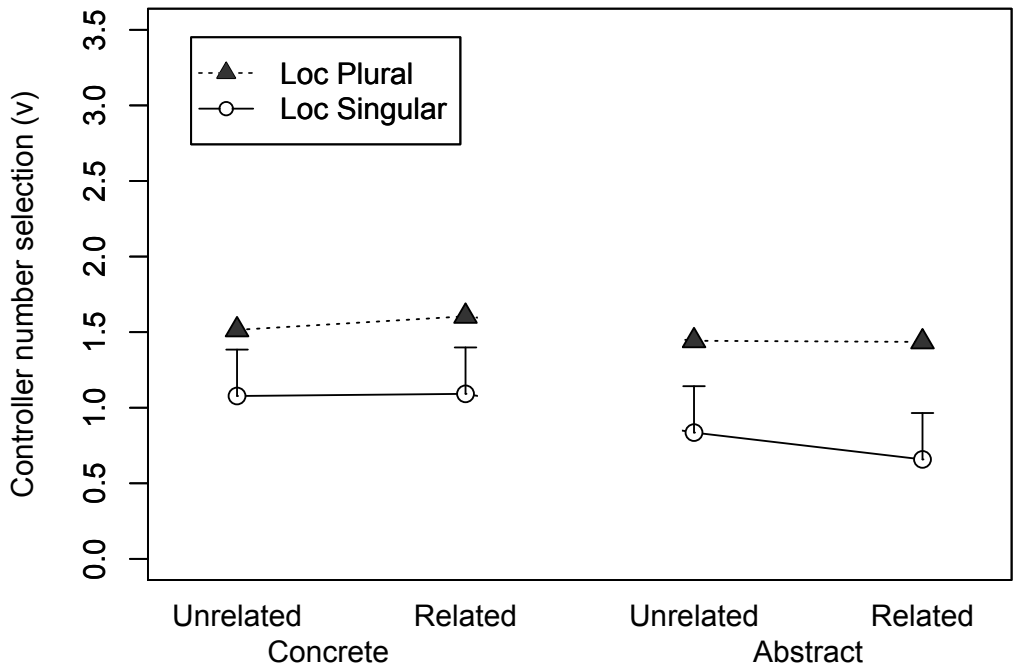


Figure 7. Fast-dm controller number selection parameter from fixed *Ter* model in Experiment 2. Larger numbers reflect ease of agreement decision. Error bars represent margin of error for differences between means (.43), calculated from the MSE of the highest-level interaction of a repeated-measures ANOVA by supersubjects.

Table 9
Experiment 2 parameter estimates for controller number selection (v) from fixed Ter diffusion model. P-values are approximated from a standard normal distribution.

	Estimate	*S.E.*	*t-value*	*p(z)*
Intercept	1.21	0.06	20.71	< 0.001
Local number	0.58	0.09	6.29	< 0.001
Concreteness	0.23	0.08	2.89	< 0.01
Relatedness	0.02	0.08	0.25	0.80
Local number x concreteness	-0.22	0.14	-1.57	0.12
Local number x relatedness	-0.12	0.14	-0.89	0.37
Concreteness x relatedness	-0.14	0.17	-0.86	0.39
Local number x concreteness x relatedness	0.09	0.27	0.35	0.73

Discussion

The broad effects in Experiment 2 were similar to those seen in Experiment 1. When the subject noun phrase had concrete (notionally plural) referents and grammatically plural local nouns, the production of plural (conventionally correct) verb number was easier than when the referents were abstract and less notionally plural. Likewise, when the local noun's grammatical number was consistent with the correct verb number, verb production was easier. These results show again that when notional and local grammatical number were congruent, they facilitated controller number selection. As in Experiment 1, the effect of verb-congruent notional number suggests that the number properties of the subject as a whole changed the production of number agreement, in line with the structural hypothesis.

In contrast to Experiment 1, the grammatical number effects were about the same for abstract and concrete subjects. With notionally plural (concrete) subjects, local singular nouns were not much harder than with notionally singular (abstract) subjects. This can be traced to the typical consequences of grammatical singularity for plural marking: There are rarely any at all. A singular local noun, with no grammatical number specification, does not override the impact of plural marking (Eberhard et al., 2005), which would be the typical value for conjoined subjects. Like the ubiquitous asymmetry between singular and plural local nouns in attraction, plural heads with singular local nouns are no more prone to the production of singular agreement than plural heads with plural local nouns. Though local plurals increase the overall likelihood of plural agreement, their capacity to do this changes only subtly with an enhanced likelihood of notional singularity. This subtle increase may be responsible for the slightly added

difficulty of singular agreement when sentence subjects were abstract. The role of grammatical singular specification will be further explored in Experiment 3.

Whereas notional number showed an effect across the board, lexical-semantic relatedness had almost no impact on agreement performance, replicating Experiment 1. There was nonetheless one suggestive finding that emerged from an exploration of item effects in the raw data. Previous work (Solomon & Pearlmutter, 2004; Gillespie & Pearlmutter, 2011) has been interpreted in terms of a relationship between lexical competition in agreement and scope of production planning. One small subset (six) of our items could have permitted a narrower planning window than the others, and for these items, semantic relatedness seemed to disrupt agreement more than for other items, consistent with lexical competition.

Specifically, for four concrete and two abstract items, there was only one determiner (e.g. *Their destiny and fate*). In the 26 remaining items, there were two determiners (e.g. *The hypothesis and the theory)*. The one-determiner configuration places both nouns in the same phrase, and grammatical structure is a modulator of planning scope. Notably, in the abstract-related-local singular condition, one-determiner responses were 220 ms slower than two-determiner responses (1414 ms vs. 1194 ms) and twice as error-prone (52% to 26%). The one-determiner configuration is more unitary both notionally (Lorimor, 2007) and structurally, and both of these properties should increase the likelihood of contemporaneous preparation of the words for a phrase.

Because contemporaneous preparation is the crux of the scope-of-planning hypothesis about lexical disruptions to agreement, this structural difference is highly suggestive, despite its restrictions. The small one-determiner item sample naturally limits

the value of inferential statistics or modeling, and apart from this single condition there were no striking differences between the one- and two-determiner items. Still, the combined impact on these items of abstractness (a message-level factor), semantic relatedness (a lexical factor), local singularity, and structural simplicity points to a promising avenue for further exploration of lexical and structural effects on producing agreement.

In the following experiments, we examine the contributions of other sources of difficulty in agreement. In Experiment 3, we look at the role of singular number specification and its interactions with linguistic and visual definitions of notional information, examining the interactions between agreement and number cognition.

CHAPTER 5

EXPERIMENT 3: VISUAL AND LINGUISTIC QUANTIFICATION

The previous experiments have demonstrated the role that notional number plays in subject-verb grammatical number agreement and its contribution to structural specifications of number. The present experiment examines the way that notional and grammatical number interact with perceptions of numerosity. Numerosity and enumeration are distinct from notional number but not entirely orthogonal to it: Notional number also reflects an assessment of how many things are in the world.

Specifically, to explore the relationship between numerosity and number agreement, this experiment examined the way that linguistically- and visually- defined grouping affects number agreement. Linguistic group information was conveyed using quantifiers, which are known to affect subject-verb agreement, while visual group information was conveyed using images that vary on properties known to affect object enumeration. The interactions between these two types of notional information and grammatical number were then assessed.

Semanticists and pragmaticists have long been concerned with the way that sets of items are linguistically represented and the way this impacts meaning in utterances. One way this has been investigated is through quantification. Quantifiers (e.g., *each, all, no, at most five*) combine with nouns to specify which of the referents out of all possible referents in the world could be under discussion, meaning that they specify subsets of referents (e.g. Barwise & Cooper, 1981). For example, the phrases *All alligators are hungry* and *No alligators are hungry* denote mutually exclusive groups of referents in the

world. In addition, the sizes of the sets are different. In the first phrase, the set of hungry alligators is numerous, while in the latter, the set of hungry alligators is empty. This draws clear parallels between quantification and notional number.

Similar to the agreement literature reviewed in the previous chapters, there has been research on the collectivity and distributivity of quantified phrases. These properties relate to whether or not a denotation could refer to multiple events. For example, consider the phrase *John and Mary lifted three pianos*. This phrase has two potential meanings, one corresponding to the gloss *John and Mary each lifted three pianos* and one corresponding to the gloss *John and Mary lifted three pianos together*. The former is distributive, while the latter is collective (Gillon, 1987; Schwartzchild, 1994). However, note that this form of distributivity relates to number differently than agreement distributivity does. Distributivity in agreement has been discussed primarily in terms of relations between objects in space, while distributivity in quantification seems to operate across time and space (e.g. Jackendoff, 1991, 1994), relating to the set of possible events that may have taken place—actions being more critical than actors.

Despite the fact that quantification seems to have a relationship with number (it's in the name, even), there has been a paucity of studies examining the relationship between quantifiers and number agreement. The little work that has been done points to a strong link between the two with respect to both grammatical and notional number. Eberhard (1997) demonstrated that quantifiers such as *one* and *each* reduce attraction in the singular head-plural local configuration, and argued that this is due to a singular grammatical specification replacing the typical neutral or default grammatical singular number. She also found effects of quantifier-conveyed notional presuppositions, with *one*

(notionally singular) eliciting less attraction than *each* or *every* (notionally plural). Further support of the notional role of quantifiers comes from quantified phrases in Serbian for which singular and plural verbs are considered equally acceptable. In these items, the implied level of referent individuation increases rates of plural verb usage (Mircovic & MacDonald, 2013).

More evidence suggesting a link between quantification, object enumeration, and agreement comes from the developmental literature. Quantification is linked with object enumeration in development, and both are in turn linked with knowledge of plural morphology. The crucial finding is that English-learning children cannot compare sets of objects outside of their subitization range until they learn to use plural morphology (e.g. Barner et al, 2007). This is likely to not be due to agreement, however, but due rather to the linguistic notion of object grouping that is common to all languages but specified in different ways, through numerals, classifiers, quantifiers, or integers: The learning pattern on set comparison tasks is comparable across languages with varying implementations of number agreement (Li et al, 2009). Children's appropriate interpretation of common quantifiers (including *a, most, some, all)* is also correlated with their ability to count, a correlation that remains robust when age is partialled out (Barner, Chow, & Yang, 2009). However, a caveat is that their ability to use this information may not be due to counting, but could be instead derived from use of approximate number systems or object tracking (see Halberta, Taing, & Lidz, 2008; Hurewitz, Papafragou, Gleitman, & Gelman, 2006).

Sets of objects can also be defined in terms of space. Perceptually, a group seems to be a set of things that "go together", as defined by gestalt properties (e.g. Wertheimer, 1923a,b). These spatial properties include proximity, object motion, and object similarity,

and they can change numerosity perception, making a set of objects seem more like one thing, or more like multiple things.

The visual cognition literature has extensively examined the types of factors that influence object identification, spatial relationships, and numerosity estimation. In particular, there is strong evidence that adding space between items inflates numerosity judgments (e.g., Kruger, 1972, Allik & Tuulmets, 1991, Ginsberg & Goldstein, 1987). This phenomenon is driven by the total area occupied by an array of items, such that the larger the lattice of items, the more inflated the numerosity estimation is (Vos, Oeffelen, Tibosch, & Allik, 1988), and the more empty space around a set of items, the less numerous it is judged to be (Allik, Tuulmets, & Vos, 1991; Sophian & Chu, 2008).

There are visual properties that can make numerosity estimation more accurate as well. If objects are collected into units, by joining with lines (Franconeri et al, 2009) or by clustering them into tight groups (Trick & Enns, 1997), numerosity estimation becomes successful. The common thread is that the more easily objects are viewed as wholes, rather than parts, the more group-like (and the more singular) they become. Much like linguistic sources of number, the collectedness or the distributedness of items can vary, and it affects perception of numerosity.

Experiment Outline

In the present experiment, we tested interactions between these two types of grouping information, linguistic and visual, by crossing the two and examining their contribution to number agreement. Linguistic number information was carried by quantifiers varied on notional number but with constant singular grammatical number, while visual number information was carried through a spatially spread or spatially

clustered array of objects. Both properties are predicted to impact enumeration of items and the question is whether (and how) they affect number agreement.

The quantifiers used in this experiment include those used in Eberhard (1997)—*each* and *every* (presupposed notional plural), and *one* (presupposed notional singular), with the addition of *a* (presupposed notional singular). Despite their notional differences, these are all grammatically singular and can only be used with a singular head noun. They therefore denote a form of *specified* singular number, in contrast to the typical unspecified nature of the singular. To contrast against these, *the* and *no* were used as a baseline. *The* refers to a single, definite item when paired with singular noun, but also occurs with plural head nouns. *No* refers to a null set that is compared to a plural comparison set, but it can occur with both singular and plural head nouns. This means that though both *the* and *no* are unspecified for grammatical number, they vary on their presupposed notional number—*no* is more notionally plural.

The visual manipulation adjusted the space between photos of the referents in an array. This was designed to create visual distributions that looked like one thing or more than one thing without changing the object content: All arrays were composed of the same number of items (six), which is outside of the subitizing range but small enough such that a viewer can easily recognize the objects on the screen.

It is predicted that all sources of number, presupposed notional number, grammatical specification, and visual distribution, will affect agreement. Presupposed notional number and grammatical specification should replicate the notional and grammatical patterns in the two previous experiments and Eberhard (1997): The presupposed plural quantifiers *each, every,* and *no* should increase rates of plural

agreement and response latencies, while the grammatically-specified singular quantifiers *a, one, each,* and *every* should decrease rates of plural agreement and response latencies. Visual number is also predicted to affect number agreement but since it has not been addressed in the literature to date, predictions for this are not as fleshed out. In particular, visual information might only affect agreement in the cases that the speaker has conflicting notional and grammatical number, amplifying notional agreement patterns, or it might dampen the effect of notional number, replacing whatever mental model the speaker may have created of the utterance.

Furthermore, by examining the interactions between these sources of information, we can clarify the interpretation of diffusion parameters. Previous work has shown that singular quantifier specification reduces attraction (e.g Eberhard, 1997). The question posed under the diffusion framework is whether this happens because of a stronger bias (z), reduction of the criterion (a), or an increased evidence accumulation rate (v). In short, this modeling technique can look at why singular specification differs from default singular number, and can look at how it interacts with other sources of information.

Method

Participants. In exchange for course credit or $7.00 compensation, 80 members of the University of Illinois community participated in the experiment. Participants were excluded if they had less than 66% of usable critical trials (N=4) or if they learned another language before learning English (N=2), leaving 72 participants.

Equipment. Stimuli were presented using Matlab R2012b with Psychtoolbox-3 on a dual-core iMac with a 24-inch flat screen. As in Experiment 1, audio was presented

to participants using Koss headphones and their speech was digitally recorded using a Sennheiser directional microphone and TubeMP preamplifier.

Table 10
Example stimuli from Experiment 3.

Notional number	*Grammatical specification*	*Quantifier*	*Root preamble*	*Predicate*
Notionally singular	Singular	A(n)	… alligator	HUNGRY
	Singular	One	with	
	Neutral	The	humongous	
Notionally plural	Singular	Each	claws	
	Singular	Every		
	Neutral	No		

Materials. There were 72 items designed to serve as sentence preambles. These were complex noun phrases containing a singular head and plural local noun inside a prepositional phrase modifier. The first word in each item was a quantifier that varied between *each, every, no, the, one,* and *a,* all of which can be used with grammatical singular number but vary on their presupposed notional number: *Each, every,* and *no* are notionally plural while *the, one,* and *a* are notionally singular. Additionally, *each, every, one,* and *a* can only be used with singular nouns, making them specified for singular grammatical number, while *no* and *the* can be used with singular and plural nouns, making them unspecified for number. Aside from the different first word, the rest of the preamble was identical across each of these six quantifier versions. The preposition

within an item varied between *for, from* and *with,* with an equal number of items using each of the three and all versions of an item using the same preposition. This was to decrease repetitiveness in the experiment. All items were also paired with a unique predicate for the speaker to use in order to encourage productions containing an inflected verb, as in Experiment 1. These predicates were designed to be compatible with the preamble as a whole in each of its six forms. See Table 10 for example stimuli, and see Appendix A for all stimuli.

Norming. To select stimuli, three norming surveys were conducted. In these surveys, 131 items (108 critical items and 111 fillers containing 2 nouns) were evaluated in order to create a stimulus set balanced across plausibility, integration, and notional number. Surveys were presented to a total of 182 raters on Amazon Mechanical Turk. All raters had a US IP-address and had done more than 1000 hits with over 95% approval. No raters did more than one list per task, although some (12 participants) did more than one norming task. Each rater was presented with a list containing a third of the potential items, with each rater in a given task using a different list.

Item lists were created by selecting a third of the critical trials per list (36), with an equal number of items containing each of the six quantifiers. This created 18 base critical item lists. A randomly-selected third of the filler items (37) was added to each of the 18 base critical item lists, for a total of 77 trials per list. Ninety unique lists were generated from the base critical item lists paired with fillers such that each critical item was seen five times in each of its six forms and each filler item was seen thirty times. These ninety lists were then used for all three norming tasks.

The first norming task rated the plausibility of the preamble and predicate using audio clips. These clips comprised of a preamble, a correctly inflected copula verb, and the associated predicate. Data were collected from 91 participants in this task, with one participant excluded for incorrectly answering a pre-survey audio captcha. Items were rated on a 1-7 scale, where 1 represented "not good" and 7 represented "very good"; instructions are displayed in Appendix B. The final set of experimental items was selected based upon the average participant ratings from this task. The final set selection was done to maximize overall plausibility, to equalize plausibilities across each of the six quantifier versions, and to have stimuli divided equally across preposition categories.

The final set of critical stimuli was rated on this measure as fairly plausible (M=5.17, mean rating range by item 2.4 to 7), with the item version containing *no* as the least plausible and *the* as the most plausible on average out of the six quantifiers (*no,* M=4.58; *the*, M=5.58). Importantly, differences were small across prepositions, though items containing *with* (M=5.35) were rated as more plausible than those containing *from* (5.12) or *for* (5.03) Plausibility ratings are displayed in Table 11.

The second norming task was designed to gather information about referential integration in the stimulus set. In this task, 90 participants rated links between underlined nouns in written sentences comprised of the preamble, the correctly inflected copula verb, and the item's predicate. Instructions for these ratings were identical to those in Experiments 1 and 2 (see Appendix B), and raters used a 1-7 scale with 1 representing "not linked" and 7 representing "very linked". Differences within item versions were minimal, with the notionally-plural quantifiers *each, every,* and *no* given lower ratings on average than the notionally-singular quantifiers *a, one,* and *the* (Notionally-plural

M=4.02, notionally singular M=4.12). Differences between prepositions were slightly larger, though still fairly small. The items containing the preposition *with* were rated as the most integrated (M=4.98), followed by those containing *from* (M=4.45) and *for* (4.15). This suggests that this item set is well-controlled for integration, a property that was demonstrated to affect agreement in Experiment 1. See Table 11 for ratings.

The final norming task rated the notional number of the preamble and predicate. In this task, participants rated the notional number of written sentences comprised of a preamble, a correctly inflected copula verb, and the item's predicate. Ratings were collected from 91 participants, with one participant excluded for failing to answer over 33% of trials. The task asked participants to simply rate whether the written phrase represented one thing or more than one thing (a two-point scale, with the left category representing "one thing"). See Appendix B for instructions. These responses were coded with "one thing" as 7 and "more than one thing" as 1 to compute item averages in the same space as the other ratings. For these ratings, there was a substantial difference between the notionally-plural quantifiers *each, every*, and *no* and the notionally-singular quantifiers *a, one,* and *the* (Notionally-plural M=2.42, notionally-singular M=6.41), confirming the validity of the notional number manipulation. Additionally, mirroring the integration ratings, there were also small differences between preposition classes, with items containing *with* rated as the most notionally singular (4.64), followed by those containing *from* (4.45) and *for* (4.15). However, average integration and notional number ratings by item were not significantly correlated, likely due to the low variability in the integration ratings themselves ($r = 0.05$, $p = 0.26$). See Table 11 for ratings.

Table 11
Norming ratings for Experiment 3.

Norming Task	Preposition	Quantifier						
		each	*every*	*no*	*a(n)*	*one*	*the*	*Mean*
Plausibility	*for*	4.94	5.03	4.48	5.32	4.80	5.62	5.03
	from	5.59	5.26	4.57	5.28	4.59	5.44	5.12
	with	5.38	5.23	4.70	5.45	5.66	5.68	5.35
	Mean	5.31	5.17	4.58	5.35	5.02	5.58	5.17
Integration	*for*	3.57	3.52	3.50	3.71	3.80	3.73	3.64
	from	3.61	3.77	3.55	3.60	3.55	3.43	3.59
	with	4.99	4.91	4.73	5.06	5.01	5.20	4.98
	Mean	4.06	4.06	3.93	4.12	4.12	4.12	4.07
Notional number	*for*	1.85	1.95	3.55	6.30	6.50	6.55	4.45
	from	2.01	1.81	2.48	6.00	6.11	6.50	4.15
	with	2.40	2.20	3.50	6.55	6.50	6.70	4.64
	Mean	2.09	1.99	3.18	6.28	6.37	6.58	4.41

Combining fillers and critical trials, there were 216 trials in the experiment, plus 8 overt practice trials. These trials were balanced on grammatical number of head and local nouns, as well as verb number: Half of the experimental trials (108) had a singular head and half had a plural head, eliciting half singular and half plural verbs, and the same number of trials had plural and singular local nouns (82 of each). The eight practice trials were also split equally between singular and plural and were comprised of simple noun phrases like "*The sharks* or *The crispy waffle"*. In addition to these practice trials, each

list began with the same 16 filler trials, split equally between singular and plural heads, but with 8 catch trials (50%) rather than the 16% represented in the rest of the experiment. This was to emphasize that there were two types of trials from the onset of the experiment, and to encourage participants to fully process the preambles for all trials.

Recording. Items were recorded in a quiet room by a female speaker from the Chicago area (the author), using a Sennheiser directional microphone routed through a TubeMP preamplifier. Items were recorded within the carrier phrase "The next sentence is", completed with "was" or "were" and the predicate for each item. Audible pauses were removed from the files to increase the speech rate and maintain fluent-sounding stimuli. The carrier and completion were then edited out for the experimental stimuli, and the carrier was edited out for the plausibility norming stimuli.

Images. For each item, fillers included, three pictures were collected to represent the item. These pictures were of objects and animals on white backgrounds and they derived from a number of sources, including freely available images from published work (Brodeur, Dionne-Dostie, Montreuil, & Lepage, 2010; Brady, Konkle, Alvarez, & Oliva, 2008; Moreno-Martinez & Montoro, 2012), Wikimedia Commons, and creative-commons licensed photos on Flickr. Photo-editing was carried out using Gimp for some items to remove backgrounds or adjust contrast and color. Photos were sized to dimensions no larger than 190 by 190 pixels. A second version of each photo was created by shrinking it by 15% (162x162 pixels maximum), decreasing the saturation by 30%, and rotating the color wheel 7 degrees to the right (shifting red toward orange, orange toward yellow, etc), using an ImageMagick script. This meant there were six different photos representing the head nouns of the phrases, as shown in Figure 10.

These six photos were semi-randomly arrayed in a staggered three-by-three grid of 1000 pixels square, centered on a 1920x1200 screen. This array was made by assigning items to the center port of the grid and five surrounding ports, with no two vertically- or horizontally-adjacent ports filled. These constraints were designed to maximize the space-filling of the pictures while maintaining a degree of randomization. Arrays were randomly chosen from the set of arrangements meeting these criteria, and a random vertical and horizontal jitter ranging from 0 to 125 (defined from the top left quadrant of the port) was added to each picture within each array to vary object positions.

For the 72 critical items and a subset of the fillers (24 additional trials, 12 in each condition), this array was transformed in two ways. Far arrays expanded the photos outward 20% relative to each other, making them more spread apart. Close arrays condensed the photos inward 20% relative to each other, making them closer together. Note that size of each photo did not vary, and that the arrangements of photos relative to each other were identical across all array versions, as seen in Figure 8.

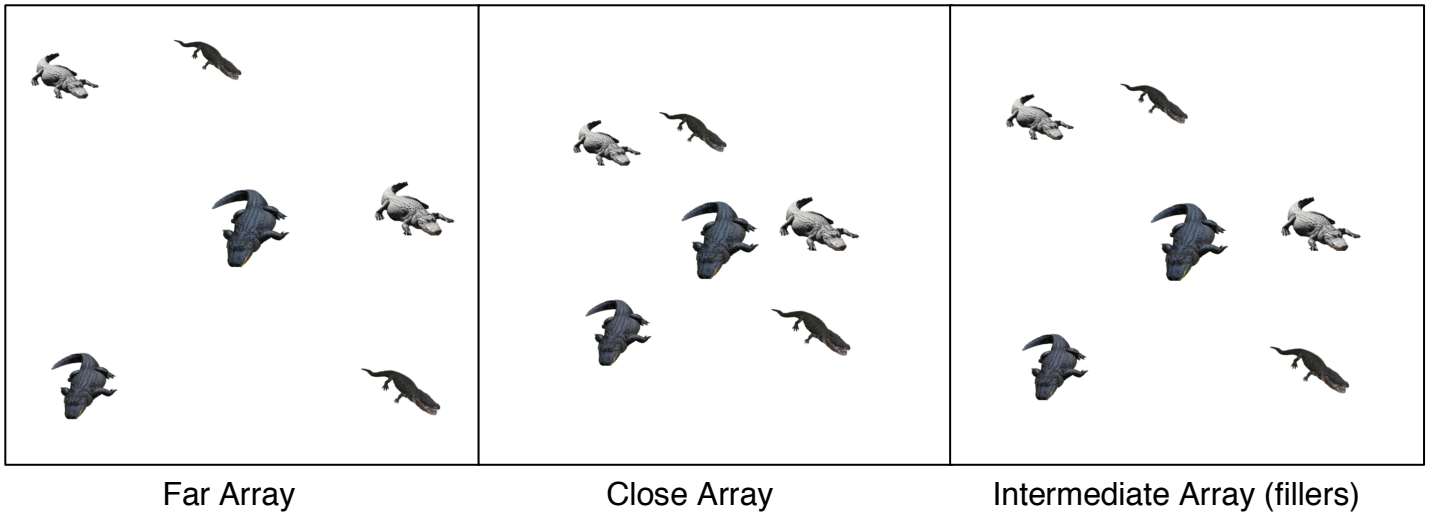


Figure 8. Sample image arrays from Experiment 3.

Procedure. The experimental procedure was similar to Experiment 1. Trials began with a cue to look at the center of the screen, a '+', displayed for 500 ms, followed by a blank screen for 50 ms and then a plausible predicate presented for 200 ms. Then, the picture array was displayed in the center of the screen. The picture array remained on the screen for 500 ms of preview time, and then the preamble was played over the headphones while the array remained on the screen. After the preamble was played, the array was replaced with a response cue, "!" to signal adding a completion only, and "Repeat" to signal repetition of the fragment and addition of a completion. An interval of 2 seconds (regular trials) or 3 seconds (catch trials) elapsed, after which the participant was cued to press a key to move on to the next trial. See Figure 9 for a diagram of the trial sequence.

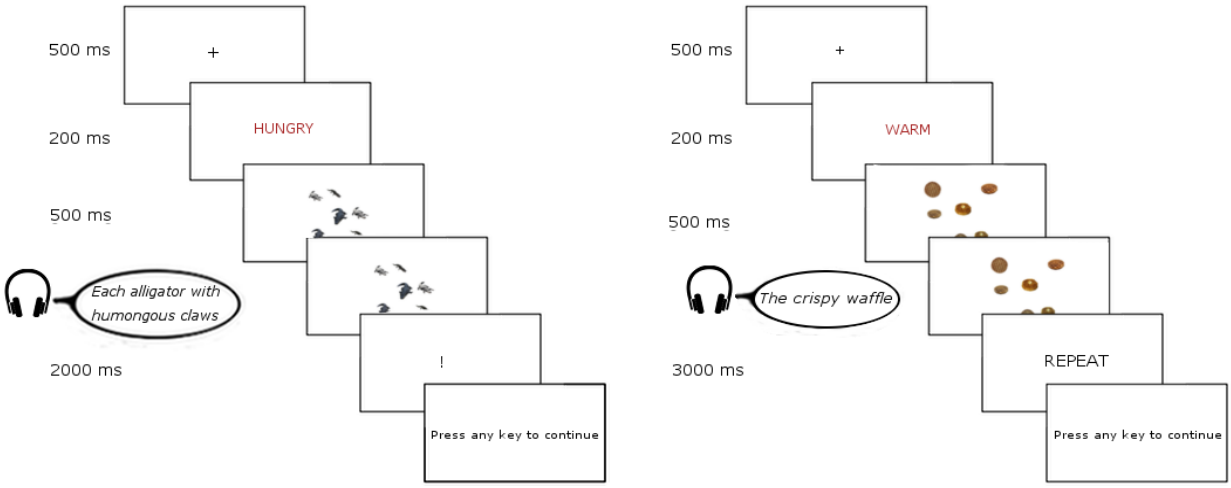


Figure 9. Trial diagram for Experiment 3 standard trials (left) and catch trials (right)

Scoring. As in Experiment 1, responses on the critical trials were scored as valid, miscellaneous, or missing. Valid responses consisted of an inflected form of the copula (*was, is, were, are)* or regular present tense verb (e.g. *seems)* and the correct predicate

adjective for the trial, with no additional modifiers (e.g. *very* or *really),* no speech corrections (e.g. *were- no was nice)* and no disfluencies before the verb. Valid responses were scored as singular or plural according to their verb number, and only valid trials were submitted to analyses. Speech onset latencies were calculated from these valid trials using a Praat script that searched first volume change of at least 50 decibels in the sound files over 3ms binned time windows.

Design. Each participant received one of 24 lists, each list containing one version of all 72 items, 6 in every combination of quantifier and array. These lists were created from 12 base lists balancing the combinations of quantifiers and arrays, and these were counterbalanced in pairs such that each item appeared in the first and the second half of the experiment, for a total of 24 lists. Every list was presented to three participants, so that every item was tested on six participants in each form. The fixed effects in the statistical analyses were quantifier notional number (singular-plural), quantifier grammatical number specification (singular-neutral) and spatial distribution (close-far), fully crossed. In the critical items, each head noun was singular, each local noun was plural, and all conventionally-correct verbs were singular.

Analysis. As in the previous experiments and as outlined in Chapter 2, the main analysis technique involved fitting diffusion models with the fast-dm program. Supersubjects were created by list, with all three participants in a given experimental list composing a supersubject. Five candidate models were run. Out of these, the full model had the highest average K-S *p*-value (0.75), followed by the fixed *Ter* model (0.74), the fixed criterion model (0.69), and the fixed bias model (0.67), with the minimal model trailing far behind (0.15). See Appendix D for distributions of K-S *p*-values by

supersubject. On the predicted versus observed data criterion calculated by effectively running the model backwards, there were minimal differences between predicted and actual median RTs and error rates across the full model and the models eliminating only one parameter (See Appendix C). As in previous experiments, the fixed *Ter* model was selected for further analysis.

Results

Figure 10 displays the error and reaction time data for Experiment 3. Rates of plural verb use were fairly low in the experiment overall, consistent with the reduction of attraction in grammatically-specified contexts. These plural verb rates were influenced by the three experimental factors. Plural verb usage was influenced by quantifier grammatical number: Items with unspecified quantifiers (*the, no*) elicited more errors than those with specified quantifiers (*each, every, one, a),* 16% and 8% respectively. Presupposed notional number had a similar effect, with notionally plural quantifiers (*each, every, no*) eliciting more errors than notionally singular ones (*one, a, the),* 14% and 8% respectively. The visual manipulation had only a minimal effect, with spatially far items eliciting more plurals than close ones when notional number was plural (15% to 13%).

In the singular reaction time measure, grammatical number and presupposed notional number interacted, such that *the* (unmarked grammatical number, presupposed singular) and *each*/*every* (specified singular, presupposed plural) were fast, while *a*/*one* (specified singular, presupposed singular) and *no* (unmarked grammatical number, presupposed plural) were slow. These differences were particularly apparent in the far-spread arrays (See Figure 10).

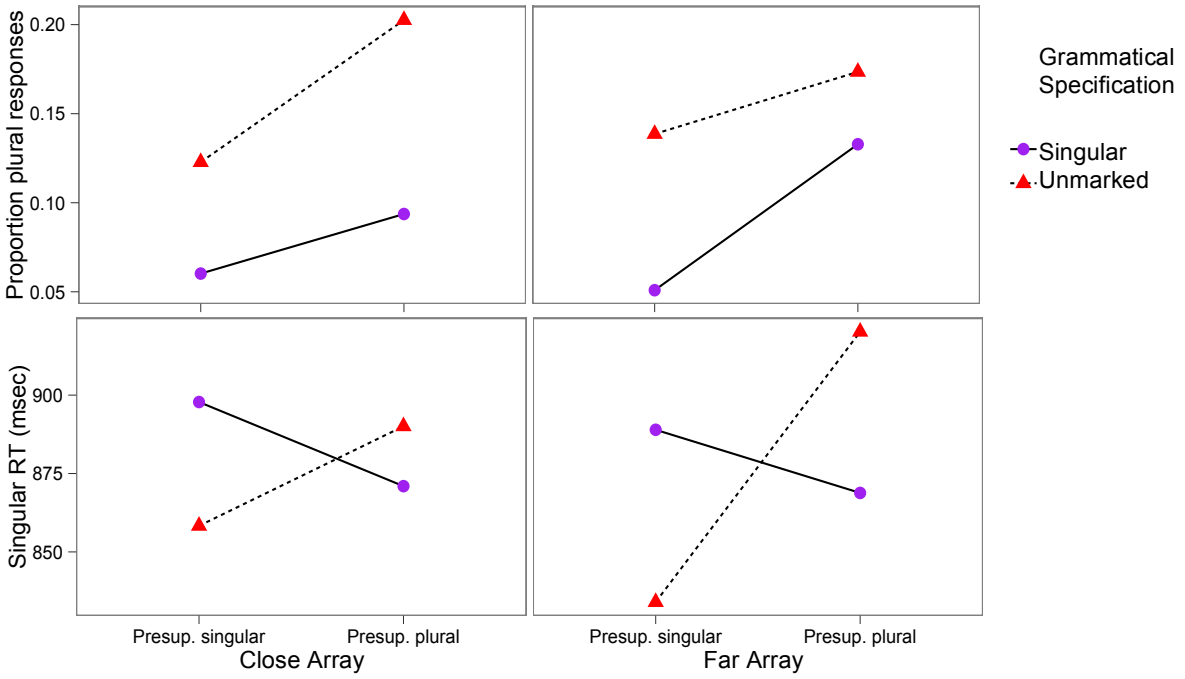


Figure 10. Proportion of plural responses (top) and mean correct response latencies in seconds (bottom) in Experiment 3 varying by notional number, quantifier specification, and spatial distribution of items.

Diffusion results are displayed in Figure 11. The controller number selection parameter (*v*) showed strong effects of both grammatical specification and presupposed notional number, with an interaction between the two such that presupposed notional number had a larger impact when grammatical number was unspecified (unspecified: 0.62 increase of v; specified, 0.25 increase): *A/one* was more difficult than *no,* while *each/every* and *the* were equivalent (see Figure 11). There were also main effects of grammatical and presupposed notional number, such that specified singular quantifiers and presupposed singulars were easier. These patterns were confirmed with mixed effect modeling (see Table 12).

The other two decision-internal diffusion parameters played an important role in these data. These parameters correspond to how conservative the speaker is in

responding (response conservativeness, *a)* and whether the speaker begins with a bias toward one response or the other (response bias, *z,* with the ratio of *z/a* representing relative bias with respect to a conservativeness baseline).

The response conservativeness parameter (*a*) showed an interaction between presupposed notional number and grammatical specification with the most conservative responses elicited when notional and grammatical number were congruent, with notionally presupposed and grammatically specified singular number (*a/one*: 1.76 vs mean of others: 1.57, see Figure 11). This was confirmed with mixed effect modeling (see Table 12).

The relative response bias parameter (*z/a*) showed an interaction between grammatical number and spatial distribution such that close arrays increased the bias toward singular responding in the grammatically-specified conditions (*each/every, one/a)* but decreased it in the grammatically-unspecified conditions (*the, no*) See Figure 11. This was confirmed with mixed effect modeling (see Table 12).

Table 12
Diffusion model parameters from Experiment 3 from fixed Ter diffusion model. P-values are approximated from a standard normal distribution.

	Controller number selection (v)				Response conservativeness (a)				Relative response bias (z/a)			
	Estimate	S.E.	t-value	p(z)	Estimate	S.E.	t-value	p(z)	Estimate	S.E.	t-value	p(z)
Intercept	1.47	0.11	13.80	< 0.001	1.62	0.04	40.43	< 0.001	0.54	0.01	44.56	< 0.001
Grammatical specification	-0.22	0.09	-2.43	0.02	-0.10	0.03	-3.13	< 0.01	-0.01	0.02	-0.94	0.35
Presupposed notional number	-0.43	0.09	-4.68	< 0.001	-0.07	0.03	-1.98	0.05	0.02	0.02	1.01	0.31
Visual distribution	-0.01	0.09	-0.15	0.88	-0.04	0.03	-1.29	0.20	-0.01	0.02	-0.41	0.68
Gramm spec x presup notional num	-0.37	0.16	-2.35	0.02	0.26	0.06	4.51	< 0.001	0.04	0.03	1.28	0.20
Gramm spec x vis dist	-0.21	0.16	-1.38	0.17	0.09	0.06	1.61	0.11	0.07	0.03	2.33	0.02
Presu. notional num x vis dist	0.04	0.16	0.29	0.77	0.04	0.06	0.64	0.52	0.00	0.03	-0.04	0.97
Gram spec x presup notional num x vis dist	-0.16	0.31	-0.50	0.61	0.18	0.11	1.57	0.12	0.02	0.06	0.38	0.70

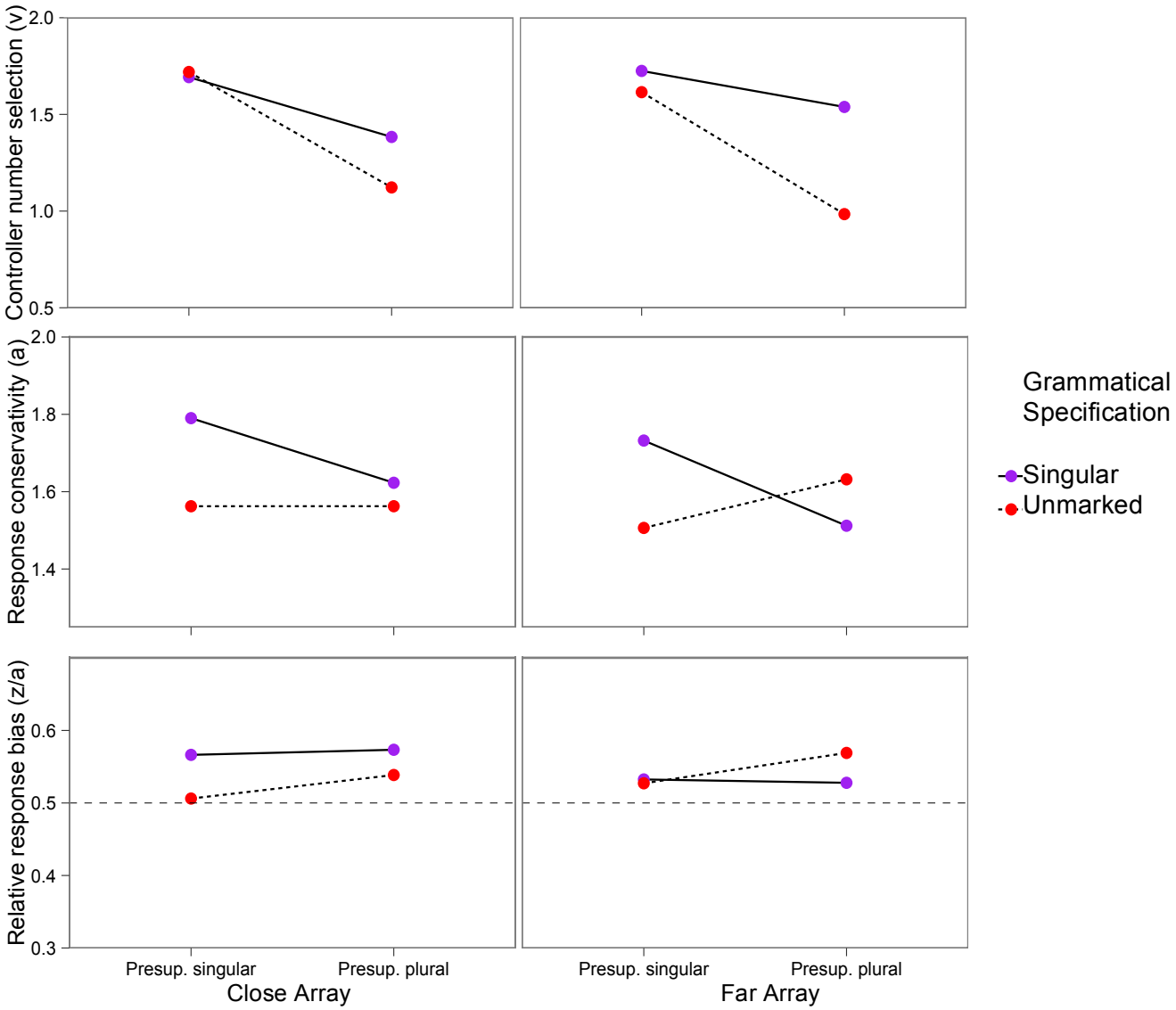


Figure 11. Fast-dm parameters from fixed *Ter* model for Experiment 3. Top: controller number selection (*v*), middle: response conservativeness (*a*), bottom: relative response bias (*z*/*a*). Larger *v* reflects ease of agreement decision, larger *a* reflects a more conservative response decision, and larger *z*/*a* reflects a bias towards singular responses.

Discussion

Replicating Eberhard (1997), in this experiment we demonstrated strong effects of quantifier grammatical specification on attraction: Singular specification makes attraction less likely, with the singular specified quantifiers *a, one, each* and *every* eliciting fewer

errors than the unspecified *no* and *the*. Additionally, as in Eberhard (1997), we demonstrated effects of presupposed notional number: The notionally plural *each, every*, and *no* elicited more plural responses than the notionally singular *a, one,* and *the*.

Response latencies showed a different pattern. For response latencies, the quantifiers *the (*unspecified, presupposed singular) as well as *each/every (*singular specification, presupposed plural) were fast, while *no* (unspecified, presupposed plural), and *one/a* (singular specification, presupposed singular) were slow. These differences were magnified by the visual manipulation, such that *the* was particularly fast and *no* was particularly slow with the far-spread array. This is to say that conflicting presupposed notional number and grammatical specification lead to *faster* completions, making these data hard to explain through notional or grammatical properties. Latency differences are also not easily accounted for with an appeal to item plausibility: *The* and *a* were the most plausible items in the norming study. One appealing possibility is that this difference may be due to pragmatics, such that when reference to a particular visual token is most clear, items are produced more quickly—*the* refers to a definite, single item, while *each* and *every* refer to all of the items in the array. Norming of the acceptability of plural sets of items with each quantifier would clarify this point, and the role of definiteness in number agreement is worthy of future study.

Despite the differences between the latency and error measures, we were able to disentangle the contributions of grammatical number and the two types of notional number using a diffusion analysis. In particular, this analysis showed that singular quantifier specification reduced attraction through the main computation of agreement (controller number selection, *v),* as did quantifier notional number. The two sources of

number interacted, with notional number particularly affecting agreement in the grammatically-unspecified condition. This replicates the results of Experiment 1. Number agreement depends upon a whollistic, structurally-derived specification of notional and grammatical number. In this case, the sources of number information are not carried in the noun, but none the less, they had a clear impact on the agreement decision.

In contrast, visual information affected only the response bias. This suggests that spatial properties can also influence mental representations by biasing the speaker towards a particular representation based upon gestalt principles. Notional number seems to have an origin in the world. In particular, these results suggest that when visually-displayed notional information is congruent with grammatical specification, speakers are particularly biased toward responding in kind. The role of this parameter z will be discussed further in the following chapters, which focus on the roles of comprehension and monitoring in agreement production.

CHAPTER 6

(MIS-)COMPREHENSION IN AGREEMENT PRODUCTION

In order to elicit agreement trouble in production, previous experiments have used a paradigm in which speakers are given a preamble and instructed to use it as a sentence subject. This is what we did in Experiments 1, 2, and 3. The assumption is that errors occur due to agreement production rather than to the comprehension of the preamble in the first place. However, this is an assumption that may not be entirely warranted. To successfully perform the preamble completion task, a speaker may only need to have formed a shallow representation of the utterance, rather than a full mental representation, and may be relying on a different set of cues to make their response than they would in naturalistic production. This means that understanding the role of comprehension in agreement is critical to understanding agreement production. The present experiments aim to clarify the nature of the comprehension processes that go in to preamble completion tasks by examining how agreement and preamble repetition change when participants do not have to speak, and when accuracy is encouraged over speed.

One core difference between production and comprehension, at least in the confines of psycholinguistic experiment, is whether or not the language user starts with a clear message. Put in the simplest terms, production involves making sounds that convey a message, while comprehension involves deriving a message from perceived sounds. In the preamble completion task discussed so far, participants are fed a message, rather than needing to generate one. This is a burden that essentially depends upon comprehension. The process of constructing a message from comprehended material means that in this task a speaker must overcome what is inherently a noisy signal in order to understand

meaning, using their predictive powers to infer what the speaker may say next and to revise their model of the utterance as needed. These predictive powers involve pragmatic constraints—though note that within an experiment, a listener depends entirely on sentence-internal properties for inference, as the context provided in more naturalistic tasks is absent. The lines drawn here are as such merely a sketch of the factors in real-world language comprehension, but a useful approximation none the less.

Successful comprehension of an utterance seems to involve prediction of upcoming elements, and this prediction is tied to production (e.g. Dell & Chang, 2013; Pickering & Garrod, 2012). The fact that prediction in comprehension comes from the utterances produced by others may lead to the systematic patterns that comprehension research has investigated for years (e.g., verb bias, syntactic ambiguities), and possibly to typological properties across languages (McDonald, 2013). Prediction is also at the heart of Bayesian models of comprehension: A prior (aka, a prediction) is set and updated online as information comes in. These Bayesian models also suggest that listeners track word and structural co-occurrence frequencies, and use these to predict upcoming information, adapting their models as an utterance unfolds (e.g. Hale, 2006, Levy, 2008).

Within the agreement literature, there is some debate about the extent to which attraction in production and comprehension derive from the same mechanisms. This is due to the fact that when speaking, utterances can be driven by a structural frame calculated from the intended message, but when listening, words have to come first. Early findings suggested that agreement comprehension and production operate in a similar fashion. In the cases that speakers find difficult, readers slow down. This slowing occurs due to mismatching head and local grammatical number (e.g. Pearlmutter, Garnsey, &

Bock 1999; Nicol, Forester, & Veres, 1997) and due to lexical semantic factors (e.g. Thornton & MacDonald, 2003). This is consistent with a model in which production difficulty directly informs prediction in comprehension.

More recent findings have called this into question, suggesting that the pattern of comprehension attraction is asymmetrical with respect to verb grammaticality. This suggests that agreement attraction in comprehension may be more consistent with reevaluation and memory search for an agreement controller, in line with models of cue based memory retrieval. The major finding that supports this is that in cases of singular heads and local plural nouns, there is a reduction in the magnitude of difficulty experienced at an ungrammatical plural verb, suggesting an "illusion of grammaticality" (Wagers et al, 2009). Importantly, this difficulty is triggered regardless of structure (in both structurally-intervening and non-intervening contexts; Wagers et al, 2009), and the pattern persists with a time-sensitive measure such as ERPs (Tanner, Nicol, & Brehm, 2014). These findings suggest that agreement comprehension may rely more on cued retrieval of controllers than on a structural gestalt that influences number marking. This may mean that agreement production is more reliant on structural cues than comprehension is, suggesting an important difference between the two modalities.

A memory-based model provides strong support for the role of lexical retrieval in comprehension, and hints further at critical differences between agreement comprehension and production. The cued retrieval model ACT-R accounts neatly for both agreement comprehension and pronoun comprehension, and suggests that agreement comprehension relies on both syntactic and morphological cues driving cued retrieval of a controller, while pronoun comprehension relies only on syntactic cues (Dillon et. al,

2013). This highlights a possible difference between comprehension and production: The Dillon et al results are in contrast to Butterfield, Cutler, Cutting, Eberhard, and Humphreys (2006), who found that the morphosyntactic specifications are similar in pronoun and verb agreement production, with the difference between the two being that pronoun agreement is more strongly influenced by notional factors.

In addition to processing differences between production and comprehension, there are issues regarding speakers' understanding of the preambles in production tasks. There is emerging evidence that speakers may not interpret presented materials literally, and that they are particularly prone to interpreting an infrequent or implausible utterance as something less surprising. Listeners seem to maintain uncertainty about syntactic relationships between words (e.g. Levy, Bicknell, Slattery, & Rayner, 2010), even to the extent that they 'correct' a speaker's utterances by adding missing function words, by swapping the order of words to match thematic roles (e.g. Gibson et al, 2013), or by altering morphology on words (e.g. Bergen, Levy, & Gibson, 2012). This may be a reason for preamble errors in production tasks—the structures presented to participants can be infrequent or strange. The number of trials discarded due to preamble error can be large, and can substantially alter the pattern of data elicited on the trials submitted to analyses (compare Solomon & Pearlmutter, 2004 to Brehm & Bock, 2013). This means that even if listeners and speakers use the same mechanisms to perform agreement, it is still necessary to worry about preamble comprehension in production tasks.

To summarize, there is evidence that comprehension relies on production, via prediction. There is evidence that agreement comprehension relies more on lexical cues than agreement production does, and there is evidence that agreement errors themselves

could be due to miscomprehension of preambles. These facts are all intertwined: Not all information is equally predictive, and mis-prediction could lead to mis-comprehension. The approach we will take to address these questions is to simply look at the role of talking in the production tasks. The question is whether removing the speaking element of the task changes the pattern of agreement and preamble errors.

A series of experiments in the literature strikes a middle ground between comprehension and production in a clever way. These experiments use an RSVP preamble completion task (Staub, 2008, 2009, 2010). In this task, speakers receive one word at a time, such as in a self-paced reading or ERP experiment, but they then complete the sentence fragment with a forced-choice selection between a singular and a plural verb. There are elements of production (generate valid next word), as well as elements of comprehension tasks (feedback to participant), and the task still requires preamble comprehension. This makes the paradigm an ideal test for the present questions.

Results using this RSVP paradigm seem to largely track previous agreement production work, demonstrating attraction in the head-singular, local-plural cases, and demonstrating sensitivity to notional number of head nouns (Staub, 2009). However, the paradigm has also been used to examine differences between intervening and non-intervening attraction (e.g, intervening: *The key to the cabinets* versus non-intervening: *The cabinets that the key open)*, and it suggests that the two may not result from the same processes (Staub, 2010). This is consistent with the RSVP task capturing elements of production and comprehension: The products of structurally-driven production matter, as does lexically-derived subject mis-comprehension and re-evaluation.

The current studies aim to continue this work by comparing the ways that

structural and lexical information are used in a paradigm that emphasizes comprehension difficulty. To investigate these questions, the RSVP completion task (Staub, 2009, 2010) was adapted to the materials from Experiment 1 by adding catch trials and a predicate compatibility measure to the task. This task was performed with speeded (Experiment 4) and unspeeded response deadlines (Experiment 5). Preamble errors in catch trials were analyzed in order to determine the comprehension burden in the experiments, and the agreement results were analyzed with a diffusion model. The diffusion analysis allows a standardized comparison of the overall time course of agreement between this task and the more standard preamble-completion one, looking at how the distribution changes across experiments, and how variability is allocated to the decision parameters (e.g. controller number selection, *v*) compared to non-decision time (*Ter)*.

This method also allows the comparison of the parameters *a* and *z* across experiments. In Experiments 1 and 2, only the results of the main parameter of interest, controller number selection (*v*), were discussed. However, allowing the parameters for response conservativeness (*a*) and response bias (*z*) to vary by condition greatly improved the models (see Appendices C and D). Their results are therefore reported here to motivate predictions for Experiments 4 and 5.

In Experiment 1 there was a significant main effect of local number on response conservativeness (*a*): Local-singular preambles promoted less conservative responding than local-plural preambles did. This was confirmed statitically with multi-level modeling (see Table 13). Response bias (*z*) showed a slight bias toward responding with a singular verb, with an average value of .52*a* (Values of *z* greater than .5*a* indicate a bias toward the correct response). There was an additional interaction between local number

and integration such that in the integrated case, local singulars elicited a stronger bias towards singular responding, and in the unintegrated case, local plurals elicited a stronger bias towards singular responding. This was confirmed statistically with multi-level modeling (see Table 13).

In Experiment 2, the response conservativeness parameter (*a*) showed effects of local number and concreteness. Here, local plural nouns elicited more conservative responding, as did abstract pairs. These effects were confirmed by multi-level modeling (see Table 13). Response bias (*z)* was close to the equibiased point, with an average value of .49*a* [95% CI .47 to .51], with no other effects observed. (Note that in this experiment, the upper response boundary was plural, representing the conventionally correct response.) These effects were confirmed by multi-level modeling (see Table 13).

These results from Experiments 1 and 2 suggest roles for response conservativeness and response bias in agreement. In both experiments, congruent local number increased *a*, replicating previous work (Staub, 2008). This suggests that *a* may reflect an awareness of monitoring or grammatical competition: Speakers lower their criterion *a* in response to conflicting information, allowing for the possibility of notional agreement but also creating a higher rate of attraction errors. A prediction for the present experiments is that *a* will increase as the speed-accuracy balance is shifted towards being accurate rather than fast. Furthermore, if the parameter *a* tracks number conflict in particular, it is predicted to be influenced by clashes in notional and grammatical number and not lexical-semantic factors, to the extent that there are observed lexical effects.

A final advantage of a diffusion model as an analysis technique in these experiments is that by equating the scale of the processes of interest across varying time

courses and varying tasks, we can examine the relative changes on controller number selection (*v*) compared to the other two parameters. Controller number selection (*v)* is predicted to vary less across tasks than the ancillary parameters *a* and *z*, and the extent to which it does should be a hallmark of differences in agreement computation across comprehension and production.

Table 13
Response conservativeness and response bias parameters from fixed Ter diffusion model for Experiments 1 and 2. P-values are approximated from a standard normal distribution

Experiment 1: Spoken complex noun phrases								
	Response conservativeness (a)				Relative response bias (z/a)			
	Estimate	S.E.	t-value	p(z)	Estimate	S.E.	t-value	p(z)
Intercept	1.35	0.03	51.70	< 0.001	0.52	0.01	52.46	< 0.001
Local number	0.09	0.02	4.24	< 0.001	-0.01	0.02	-0.39	0.70
Integration	-0.03	0.03	-1.03	0.30	0.02	0.02	1.20	0.23
Predicate compatibility	0.04	0.02	1.77	0.08	-0.01	0.02	-0.50	0.62
Local number x integration	-0.05	0.06	-0.92	0.36	0.12	0.04	3.18	< 0.01
Local number x predicate compatibility	-0.01	0.05	-0.19	0.85	0.02	0.04	0.35	0.73
Integration x predicate compatibility	-0.04	0.06	-0.64	0.52	-0.01	0.04	-0.20	0.84
Local number x integration x predicate compatibility	-0.02	0.08	-0.29	0.77	-0.01	0.06	-0.14	0.89
Experiment 2: Spoken conjoined noun phrases								
Intercept	1.70	0.02	68.42	< 0.001	0.49	0.01	43.15	< 0.001
Local number	0.14	0.03	5.16	< 0.001	0.00	0.02	-0.20	0.84
Concreteness	0.10	0.03	3.65	< 0.001	-0.01	0.02	-0.54	0.59
Relatedness	-0.01	0.04	-0.27	0.79	0.00	0.01	0.15	0.88
Local number x concreteness	-0.08	0.06	-1.47	0.14	-0.02	0.04	-0.56	0.58
Local number x relatedness	-0.07	0.07	-0.98	0.33	0.02	0.03	0.76	0.45
Concreteness x relatedness	0.02	0.05	0.37	0.71	-0.01	0.04	-0.13	0.90
Local number x concreteness x relatedness	0.06	0.13	0.47	0.64	0.09	0.06	1.54	0.12

The current experiments

In Experiments 4 and 5, grammatical number, notional number and lexical-semantic relationships were manipulated in order to assess variations in agreement across comprehension and production. Data were collected in two similar paradigms in order to manipulate the balance of comprehension and production resources used in agreement. The first of these was a speeded button-press task in which participants are asked to quickly judge the next word in a sentence (Experiment 4) and which has previously demonstrated results largely similar to production (e.g. Staub, 2009; Veenstra et al, 2014). The second task was a no-pressure button-press task, which by eliminating incentives towards maintaining running speech is theorized to pick up primarily on comprehension difficulty (Experiment 5).

The experimental design was identical to Experiment 1, with structural (integration) and lexical (lexical-semantic compatibility) variables manipulated alongside local grammatical number. The three factors were predicted to have similar effects to Experiment 1. Referential integration, relating to how many things are in the message, is predicted to cause difficulty due to notional number. Predicate compatibility, a manipulation of connections between words, is predicted to cause difficulty due to lexical-semantic factors. Both factors are predicted to interact with grammatical number, indicating in turn whether local noun number has a structural effect (interacting with referential integration), or a lexical effect (interacting with predicate compatibility). The way that these effects change among experiments indicates the degree to which information is used in comprehension compared to production. In particular, if lexical effects increase and notional effects diminish, this suggests a primarily-lexical effect of

agreement comprehension.

Additionally, by assessing the rates of preamble repetition errors in the catch trials and by examining where they occur, we can address the question of preamble comprehension. In particular, if participants misremember inflections more often in the RSVP experiments compared to the speaking ones, that suggests that comprehension difficulty is greater, and the agreement difficulty observed in this experiment may be due to uncertainty about number information in the preambles.

As the critical difference between Experiments 4 and 5 is in the methodology, and as the procedure was similar across both, the methods are outlined in ensemble below in order to highlight the similarities and differences in the two.

Method

Equipment. Experiments were run on a Mac Mini or Dell desktop computer with a 17-inch monitor, using Matlab R2009b and PsychToolbox-3. Audio was digitally recorded using a Sennheiser directional microphone run through a Tube MP preamplifier.

Materials. Experimental materials were identical to Experiment 1. There were 24 experimental items, varying in integration, local number, and predicate compatibility, all fully crossed. See Table 1 for example stimuli, and Appendix A for the full list. As in Staub (2009, 2010), there were an additional 12 practice trials added to the beginning of the experiment in order to give the participants a chance to familiarize themselves with what is an admittedly strange and novel task.

Procedure. The procedure was adapted from Staub (2009). Stimuli were presented one word at a time in an RSVP paradigm ending in a two-choice button-press

decision of the fragment's verb continuation. All text was centrally displayed in size 14 black Times New Roman font unless otherwise noted. See Appendix B for instructions.

All trials began with a fixation cross presented for 1000 ms, followed by a predicate in all-caps red font presented for 250 ms, followed by a blank screen for 150 ms. The preamble followed, presented one word at a time in the center of the screen for 250 ms each with a 150 ms ISI. Critical preambles all had six words; fillers varied in length from two to nine words.

For the standard trials, after the preamble was presented, the words "WAS" and "WERE" appeared on the screen, with the word "WAS" on the left. Participants selected the appropriate continuation to the preamble by pressing the F or J key, corresponding to "was" and "were", respectively. They were then given feedback on whether their decision was correct with the text "Correct" or "Incorrect". For Experiment 4, there was a 1.2 second time deadline allowed for responding, after which a buzzing tone was presented and "TOO SLOW!" was displayed on the screen. For Experiment 5, participants were allowed to take as much time as they wished. See Figure 12 for a trial diagram.

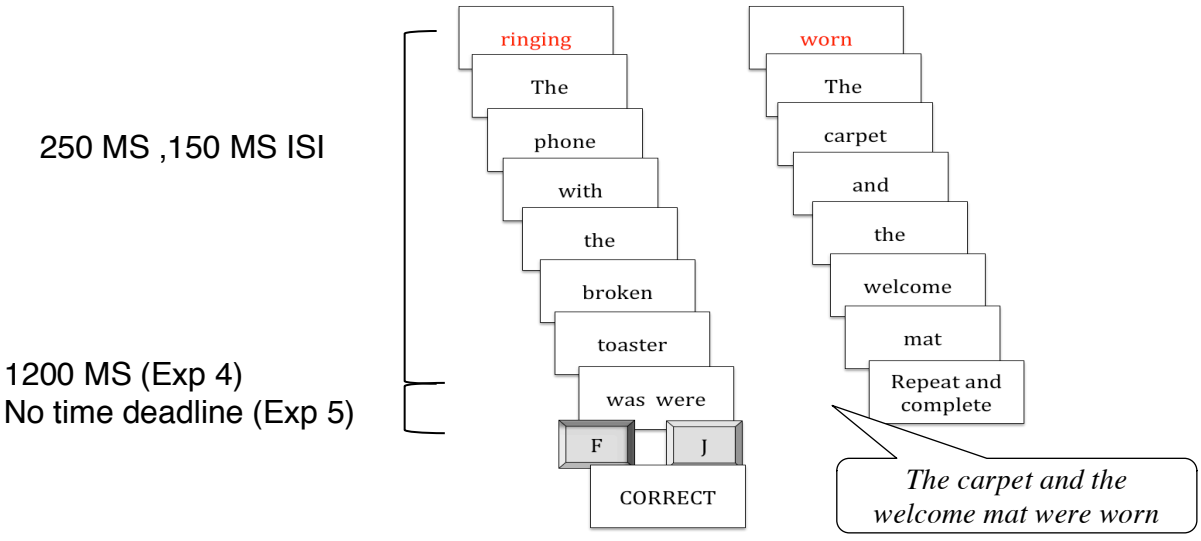


Figure 12. Trial diagram, for standard (left) and catch trials (right) in RSVP experiments.

These standard trials were interleaved with catch trials that involved speaking aloud (Experiment 4) or typing full sentence completions (Experiment 5) in lieu of the forced-choice decision. These trials were cued by 'Repeat and complete' (Experiment 4), or 'Type response' (Experiment 5). In the typed completions, participants were allowed to use the delete key to edit their responses, as this was deemed more natural. The addition of catch trials marks a change from Staub (2008, 2009, 2010), but maintains the parallels between Experiments 1 and 2 and the current projects and allows the examination of preamble repetition errors. See Figure 12 for a trial diagram.

Scoring. For standard trials, responses of "F" were coded as singular, and responses of "J" were coded as plural. For catch trials, verbs were scored as singular or plural if they bore number marking (e.g., copular verbs such as "is" or "were", regular present tense verbs such as "seems"), scored as uninflected if they contained verbs with non-transparent morphology (e.g. past tense regular verbs, such as "seemed"), and scored as miscellaneous otherwise. Predicates were scored as correct or incorrect, with spelling errors and one-phoneme sound changes counted as correct. Preambles were scored as correct or incorrect, and as with the predicates, small changes to preambles that preserved meaning, structure, and morphology were also accepted as correct. This included one-phoneme sound substitutions that did not alter morphology or transform the word into another word (e.g. "tasi" for "taxi"), omission/alteration of determiners (e.g. omitting "the" or changing it to "a"), and spelling errors (e.g. "glitery" for "glittery"). This was intended to give as conservative a measure of preamble errors as possible and to make sure that speaking and typing themselves did not alter the data pattern, given their different articulatory constraints (mouths versus fingers) and given that the typed task

allowed editing. Incorrect preambles were subdivided into errors on noun lemmas (omission, exchange or substitution of noun forms), errors on noun inflectional morphology (omission/addition of -s), and miscellaneous errors (everything else).

Design. Identical to Experiment 1. Each participant received one of 16 lists, each list containing one version of all 24 items, three in every condition, and with pairs of lists counterbalancing item presentation order. Every list was presented to 8 participants, so that every item was tested on 16 participants. The fixed effects in the statistical analyses were integration (integrated-unintegrated), lexical-semantic compatibility (head-local), and local-noun number (singular-plural), all fully crossed.

Analysis. Critical trial analysis was identical to Experiment 1, using diffusion modeling as outlined in Chapter 2. The catch trial analysis was performed using multiple linear regression, with critical trial error rates as dependent measures and experiment and catch trial error rates as predictors.

Experiment 4: Speeded button pressing

Introduction. This experiment was designed to compare the production task used in Experiment 1 with a task that blends together components of comprehension and production (e.g. Staub, 2009). This task involves a two-alternative forced-choice decision, making it similar to certain comprehension tasks (e.g. Nicol et al, 1997) and to the types of judgments typically submitted to diffusion analyses (e.g. Ratcliff, 1978; Ratcliff et al, 2004; Nozari & Dell, 2009). The goal is to further disentangle structural from lexical contributions to number agreement, compare the inner workings of comprehension and production, and to begin to coherently interpret the response bias and response conservativeness parameters.

Method

Participants. Data were collected from 141 members of the University of Illinois community. Participants were compensated with course credit or $7 compensation. Data were excluded from 9 participants who failed to provide usable responses for over 33% of critical trials, from one participant who was not a native English speaker, and from three participants due to technical difficulties. This left 128 participants. All participants were right-handed.

Procedure. As outlined above, with a 1.2 second response deadline for verb selection for standard trials, and speak-aloud catch trials in which participants repeated back the preamble with its predicate. Participants were run one at a time in a quiet room with the experimenter sitting beside them.

Analysis. Diffusion models were fitted using the fast-dm program, as outlined above. Five candidate models were run, and the two model fitting procedures outlined in Chapter 2 were performed. The full model had the highest average K-S *p*-value (0.80), followed by the fixed criterion model (0.75), the fixed bias model (0.73), and the fixed *Ter* model (0.70), with the minimal model trailing far behind (0.20). Despite the differences in K-S *p*-value, there were minimal differences between predicted and actual median RTs and error rates across the full model and the models eliminating only one parameter (See Appendix C). The fixed *Ter* model was therefore selected to match with the previous experiments, though compared to previous experiments, this model fitted the data less well (See Appendix D).

Results and Discussion. Correct reaction times and the proportion of error responses roughly mirrored Experiment 1 (See Figure 13). Across conditions, average

response speed and error rates were close to those in Experiment 1 (E1: Mean correct RT= 807 ms, 13% plural responses; E4: Mean correct RT= 722 ms; 14% plural responses). Additionally, as in Experiment 1, local plural nouns elicited slower correct responses and more errors than local singular nouns (Local plural: mean correct RT = 747 ms, 22% plural responses; Local singular: mean correct RT = 697 ms, 5% plural responses), as did unintegrated fragments compared to integrated ones (Unintegrated: mean correct RT = 731 ms, 19% plural responses; Integrated: mean correct RT = 713 ms, 8% plural responses). There was only a small effect of the predicate compatibility measure in latency and error measures (Local-compatible: mean correct RT = 717 ms, 15% plural responses; Head-compatible: mean correct RT = 727 ms, 13% plural responses). Mean response latencies by condition are displayed in Table 14.

Table 14.
Response latencies (in milliseconds) from Experiments 4 and 5.

Experiment	*Integration*	*Predicate compatibility*	*Singular Response*		*Plural Response*	
			Local singular	*Local plural*	*Local singular*	*Local plural*
Experiment 4: Speeded button press	*Integrated*	*Head-Compatible*	692	727	849	702
		Local-Compatible	680	752	806	751
	Unintegrated	*Head-Compatible*	715	774	803	767
		Local-Compatible	700	736	797	808
Experiment 5: No-pressure button press	*Integrated*	*Head-Compatible*	1140	1177	1264	1243
		Local-Compatible	1112	1253	965	1167
	Unintegrated	*Head-Compatible*	1260	1419	1699	1457
		Local-Compatible	1185	1418	1303	1270

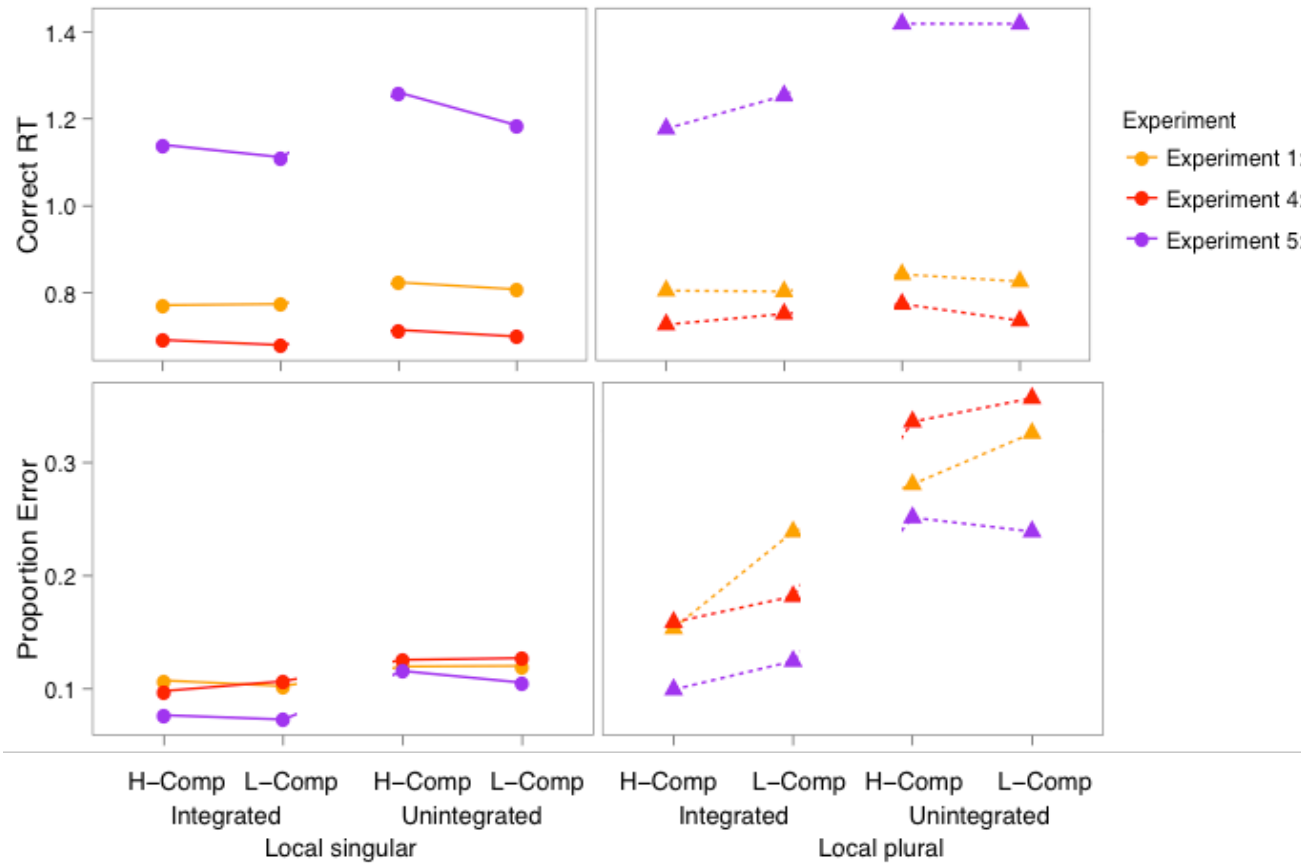


Figure 13. Mean singular response latencies in seconds (top) and proportion of plural responses (bottom) in Experiments 1 (speaking), 4 (speeded button pressing), and 5 (unspeeded button pressing),varying by local number, integration, and predicate compatibility.

The diffusion analysis of the results is presented in Table 15. On the controller number selection parameter (*v*), there was an interaction between local number and integration such that unintegrated preambles were particularly difficult in the presence of a local plural noun, with a reduced *v*. There were also main effects of local number, integration, and predicate compatibility, such that local plural nouns were more difficult than local singulars, unintegrated preambles were more difficult than integrated preambles, and local-compatible predicates were more difficult than head-compatible ones (See Figure 14). For the response conservativeness parameter (*a*), there were no

significant effects (See Figure 14), but the value was close to that of Experiment 1 (E1: a = 1.35, E4: a = 1.16). The relative response bias (z/a) was relatively neutral (z = .51a), with a three-way interaction such that the unintegrated-head compatible-local plural case was biased toward plural responses, as well as a main effect of local number such that local singulars were more biased toward singular responses than local plurals were (see Figure 14).

In general, the parallels in results between this experiment and Experiment 1 are strong, suggesting the robustness of the RSVP button-pressing paradigm as a measure of difficulty in subject-verb agreement, as well as the strength of the diffusion analysis technique for subject-verb agreement. Despite the fact that this task involves a more covert production element than the task in Experiment 1, the predictive processes inherent to the task show similar results: Again, local plural nouns and unintegrated fragments elicited slower and less accurate responses, translating to reduced rates on the controller number selection parameter (v). Again, the structural variable of integration demonstrates larger effects than the lexical variable of predicate compatibility. However predicate compatibility had a somewhat larger effect here than in Experiment 1, suggesting the importance of lexical information in prediction as in Wagers et al (2009). The combination of these findings suggests that contrary to some previous findings (e.g. Wagers et al, 2009), prediction can rely on both lexical and structural sources of number agreement, and suggests that the type of prediction tapped in this task may be covert production, as argued in Staub (2009).

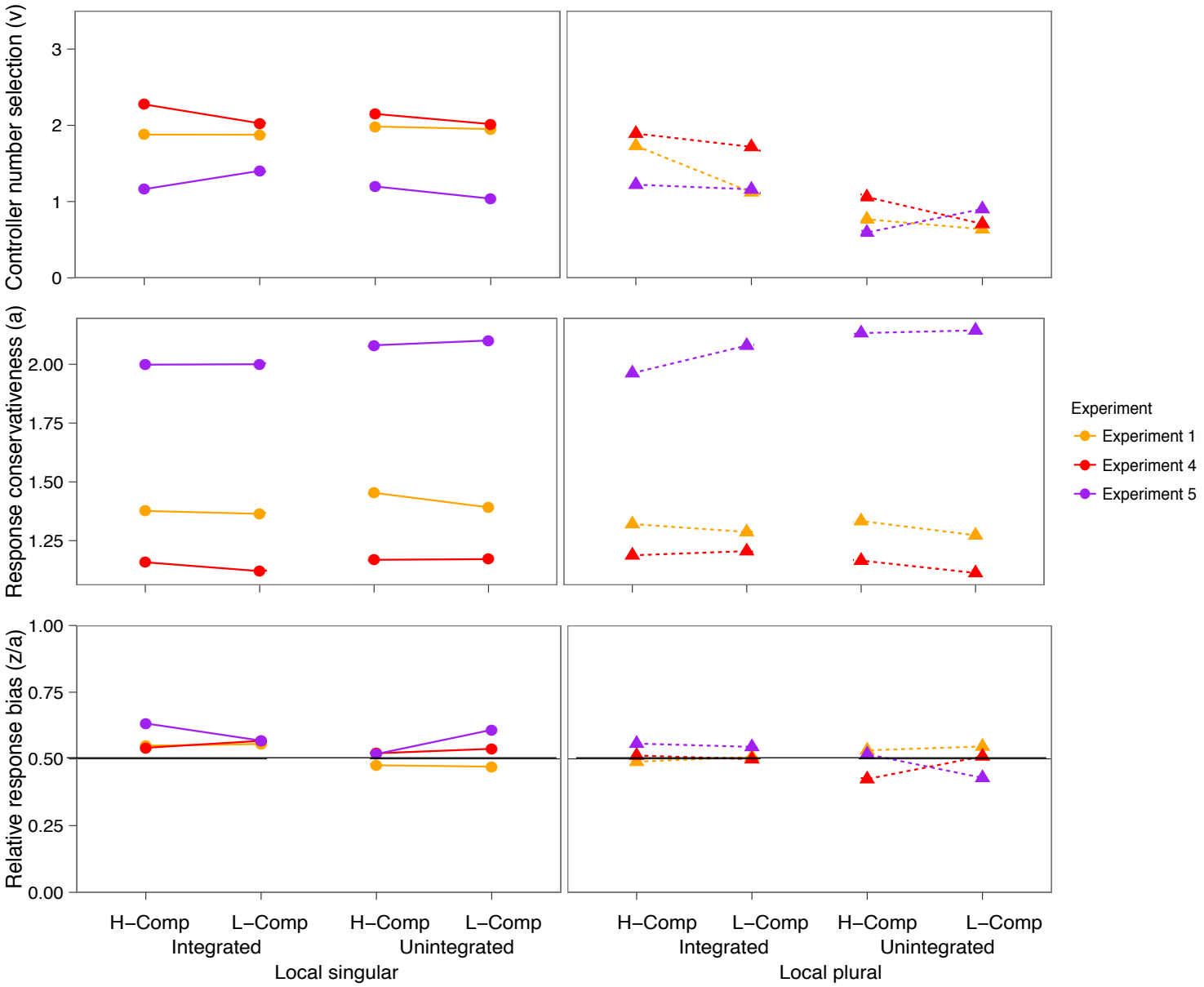


Figure 14. Fast-dm parameters from fixed *Ter* model for Experiments 1 (speaking), 4 (speeded button pressing), and 5 (unspeeded button pressing). Top: controller number selection (*v*), middle: response conservativeness (*a*), bottom: relative response bias (*z*/*a*). Larger *v* reflects ease of agreement decision, larger *a* reflects a more conservative response decision, and larger *z*/*a* reflects a bias towards singular responses.

Table 15.
Diffusion model parameters from Experiments 4 and 5 from fixed Ter diffusion model. P-values are approximated from a standard normal distribution.

	Controller number selection (v)				Response conservativeness (a)				Relative response bias (z/a)			
	Estimate	S.E.	t-value	p(z)	Estimate	S.E.	t-value	p(z)	Estimate	S.E.	t-value	p(z)
Experiment 4: Speeded button press												
Intercept	1.73	0.06	29.11	< 0.001	1.16	0.02	64.8	< 0.001	0.51	0.01	45.35	< 0.001
Local number	0.77	0.13	5.77	< 0.001	-0.01	0.02	-0.55	0.58	0.06	0.02	2.71	< 0.01
Integration	0.49	0.11	4.33	< 0.001	0.01	0.02	0.57	0.57	0.03	0.02	1.77	0.08
Predicate compatibility	0.23	0.1	2.27	0.02	0.02	0.02	0.88	0.38	-0.03	0.02	-1.62	0.11
Local number x integration	-0.85	0.18	-4.81	< 0.001	-0.09	0.05	-1.91	0.06	-0.01	0.03	-0.46	0.65
Local number x predicate compatibility	-0.07	0.2	-0.36	0.72	0	0.05	0	1	0.01	0.03	0.55	0.58
Integration x predicate compatibility	-0.03	0.25	-0.13	0.9	-0.02	0.04	-0.36	0.72	0.04	0.03	1.36	0.17
Local number x integration x predicate compatibility	0.29	0.49	0.6	0.55	0.11	0.1	1.15	0.25	-0.11	0.05	-2.03	0.04
Experiment 5: Unspeeded button press												
Intercept	1.09	0.05	20.47	< 0.001	2.06	0.07	30.32	< 0.001	0.55	0.01	53.93	< 0.001
Local number	0.23	0.07	3.54	< 0.001	-0.03	0.05	-0.66	0.51	0.07	0.02	3.53	< 0.001
Integration	0.3	0.07	4.36	< 0.001	-0.1	0.04	-2.38	0.02	0.06	0.02	3.6	< 0.001
Predicate compatibility	-0.08	0.07	-1.16	0.25	-0.04	0.03	-1.08	0.28	0.02	0.02	1.15	0.25
Local number x integration	-0.28	0.16	-1.74	0.08	0.03	0.08	0.34	0.73	-0.04	0.04	-1.11	0.27
Local number x predicate compatibility	0.08	0.15	0.54	0.59	0.05	0.09	0.57	0.57	-0.06	0.03	-1.96	0.05
Integration x predicate compatibility	-0.01	0.12	-0.12	0.9	-0.04	0.08	-0.54	0.59	0.04	0.03	1.34	0.18
Local number x integration x predicate compatibility	-0.77	0.3	-2.55	< 0.01	0.12	0.18	0.69	0.49	0.23	0.07	3.38	< 0.001

Experiment 5 Unspeeded button pressing

Introduction. This experiment was designed to further examine the interplay between sources of number information in production and comprehension by removing the speed deadline in the button press paradigm (e.g. Staub, 2010). The hypothesis is that due to the removed time pressure, preamble comprehension becomes the major locus of difficulty in the task. This was further emphasized with typed, not spoken, catch trials.

Method.

Participants. Data were collected from 136 members of the University of Illinois community. Participants were compensated with course credit or $7 compensation. Data were excluded from five participants who had a critical trial accuracy level under 60%, from one participant who did not keep her hands on the keyboard, and from two participants due to technical difficulties. This left 128 participants, all right-handed.

Procedure. As outlined above, with no time deadline for standard trials, and with catch trials in which participants repeated back the preamble and predicate. Participants were run alone or in groups of up to 3 with the experimenter sitting in the same room.

Analysis. A single trial was excluded as it had a reaction time of 17 seconds and was deemed to be a contaminant. Diffusion models were fitted on the rest of the data using the fast-dm program, as outlined above. Five candidate models were run, and the two model fitting procedures outlined in Chapter 2 were performed. The full model had the highest average K-S *p*-value (0.79), followed by the fixed criterion model (0.78), the fixed *Ter* model (0.74), and the fixed bias model (0.65), with the minimal model trailing far behind (0.26). Again, there were minimal differences between predicted and actual median RTs and error rates across the full model and the models eliminating only one parameter, though due to the low error rates in the experiment, there was a great deal of variability in the predicted error reaction time (See Appendix C). The fixed *Ter* model was again selected for analyses, as it provided a similar fit to the data as the maximal model (see Appendix D), and as it was the model used in previous experiments.

Results and Discussion. As in the previous experiments, local plural nouns elicited slower correct responses and more errors than local singular nouns (Local plural:

mean correct RT = 1317 ms, 15% plural responses; Local singular: mean correct RT = 1174 ms, 5% plural responses), and unintegrated fragments elicited slower correct responses and more errors than integrated ones (Unintegrated: mean correct RT = 1321 ms, 15% plural responses; Integrated: mean correct RT = 1171 ms, 5% plural responses; See Figure 13). There was minimal effect of the predicate compatibility measure in either of the latency or error measures (Local-compatible: mean correct RT = 1242 ms, 10% plural responses; Head-compatible: mean correct RT = 1249 ms, 10% plural responses). In contrast to Experiment 1 and Experiment 4, responses were slower and more accurate, particularly in the local plural condition (See Figure 13). Additionally, in contrast to the previous experiments, errors were not always slower than correct responses: In fact, errors tended to be faster than correct responses in the local-compatible conditions. Mean response latencies by condition are displayed in Table 14.

Results of the diffusion analysis are reported in Table 15. For the controller number selection parameter (v), there was a three-way interaction between local number, integration, and predicate compatibility such that the local plural-unintegrated-local compatible condition was easier than otherwise predicted. There were also main effects of local number and integration such that local plurals and unintegrated phrases were especially difficult (see Figure 14). For the response conservativeness parameter (a), there was a main effect of integration such that unintegrated fragments elicited more conservative responses than integrated ones. Responses were also far more conservative across the board than in Experiments 1 or 4 (2.06 to E1=1.35, E4=1.16; See Figure 14). Relative response bias (z/a) was again biased slightly toward singular responses (z=.55a), and there were main effects of local number and integration such that local singular

nouns and integrated preambles were particularly biased towards a singular response. There was also a three-way interaction between local number, integration, and predicate compatibility, such that the local plural-unintegrated-local compatible condition was biased toward plural responses (see Figure 14).

In this experiment, participants were fairly accurate and quite slow in their responding. This translated to reduced controller number selection (v) and increased response conservativeness (a) compared to the previous two experiments, but a similar relative response bias (z/a). As in the previous experiments, local plural noun number elicited slower responses and more errors, and decreased the controller number selection parameter (v). Similarly, unintegrated fragments were more difficult than integrated ones, eliciting slower responses and more errors, as well as decreasing the controller number selection parameter (v) and increasing the response conservativeness parameter (a). As in Experiment 1, there were limited effects of the predicate compatibility measure. In this experiment, predicate compatibility only elicited effects via complementary three-way interactions on the controller number selection parameter (v) and the relative response bias parameter (z/a) in an unintegrated fragment with a local plural noun. In this case, a local-compatible predicate was less likely to lead to a plural response than would be otherwise predicted for the controller number selection parameter (v), but more likely to lead to a plural response due to response bias (z/a). However, though significant, these results may be as a result of the task's low error rates and the corresponding reduced reliability of the model fitting procedure.

Catch trial analysis

Given the differences among experiments on critical trial performance, and given

the previous literature on noisy channel comprehension (e.g. Gibson et al, 2013; Bergen et al, 2012), an additional analysis was carried out in order to examine the types of errors on catch trials. Catch trial preamble repetitions were coded for errors, and errors were classified as specified in *Scoring*. In particular, what was of interest were qualitative changes to the types of errors across experiments, as well as the relationships between the by-subject rate of inflectional morphology errors, the by-subject rate of noun and adjective errors, and performance in the critical trials.

In general, Experiment 1 elicited fewer catch trial errors than either Experiments 4 or 5. In Experiment 1, 4% of trials contained a preamble error, compared to 16% (Experiment 4) and 15% of trials (Experiment 5). A similar pattern occurred for predicate adjective errors, which occurred on less than 1% of catch trials in Experiment 1, but occurred on 9% of catch trials in Experiment 4, and 11% of catch trials in Experiment 5. However, the ratios of error types remained relatively comparable across experiments, with the most errors on inflectional morphology (56% of Experiment 1 errors, 35% of Experiment 4 errors, and 42% of Experiment 5 errors), followed by errors on noun lemmas (E1: 33% of errors, E4: 33% of errors, E5: 34% of errors), followed by other error types (E1: 11%, E4: 32%, E5: 24%).

One potential driver of the attraction effect in the present experiments is a failure to accurately encode the morphology on nouns. Improperly encoding a singular head noun as plural would lead to a response pattern that would look like attraction, while failure to notice the –s ending on attractors would lead to participants being immune to attraction. To address this question, by-participant rates of attraction (errors in local plural trials minus errors in local singular trials) were predicted from rates of

morphological or non-morphological errors across experiments using multiple linear regression. Participants who produced many morphological errors in catch trials produced more attraction errors in critical trials (β=0.55, t=3.17, p<.01), and while participants in Experiment 1 (speaking) and Experiment 4 (speeded) produced more attraction errors than Experiment 5 (unspeeded), the relationship between morphological errors and attraction errors did not differ by experiment ($F(1,378) = 0.18$, $p = 0.83$). This relationship was not due to generally-error prone responding, as attraction was not predicted by non-morphological preamble errors (β=-0.05, t=-0.30, $p = 0.77$), and did the relationship between non-morphological errors and attraction errors did not differ by experiment ($F(1,378) = 1.57$, $p = 0.21$).

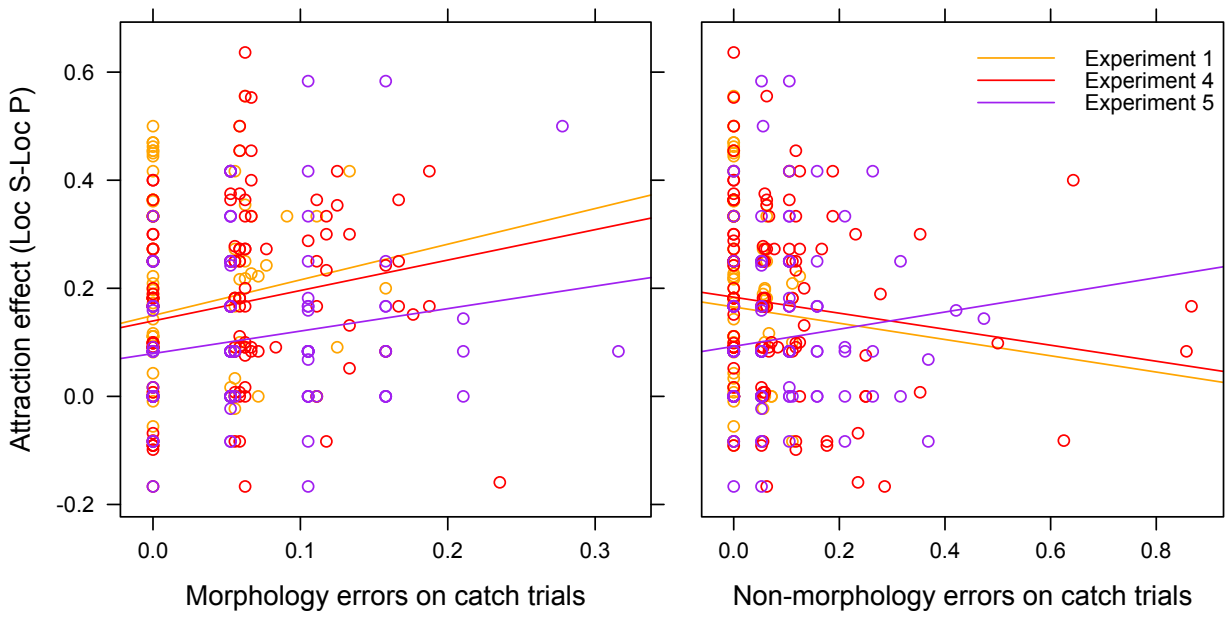


Figure 15. By-participant relationship between magnitude of attraction effect and catch trial errors for Experiments 1 (speaking), 4 (speeded button pressing), and 5 (unspeeded button pressing). Left panel represents preamble repetition errors involving inflectional morphology, right panel represents all other catch trial preamble repetition errors.

The second question of interest relates to the interaction in Experiment 5 between local number, integration, and predicate compatibility, in which local plural-unintegrated-local compatible trials are less difficult than would otherwise be predicted. This pattern of data could be due to participants ignoring the predicates. To address this question, the rates of incorrect predicate adjective use were compared across experiments and subjects. Variations in performance did not predict predicate compatibility. There was no systematic relationship between predicate adjective errors and the predicate compatibility effect (as measured by a difference score between head-compatible and local-compatible adjectives for agreement errors), β=-0.34, t=-0.63, p=0.53.

Discussion

Experiments 4 and 5 compared the spoken preamble completion paradigm to a button-pressing paradigm in order to examine the role of comprehension in production. Experiment 4 used a timed-button pressing task similar to previous work (e.g. Staub, 2009), while Experiment 5 removed the time deadline (similar to Staub, 2010). Experiment 4 had the same catch task as in Experiment 1, while Experiment 5 used a typed catch task. This allowed us to look at the role of speaking in agreement production and in preamble repetition in order to assess the contributions of prediction and comprehension errors in agreement production.

Comparing Experiment 1 and Experiment 4 allows us to look at the contrast between speaking and the speeded button pressing task. These experiments had a similar overall speed range. However, the speeded button task elicited a stronger effect of predicate bias than speaking did, in overt errors and in the controller number selection parameter (v). This fits with the notion that agreement comprehension relies more on

lexical cues than agreement production does. Experiments 1 and 4 were similar on the response conservativeness (*a*) and response bias (*z*) parameters. Speeded button pressing was slightly less conservative than speaking, with a slightly reduced response conservativeness parameter (*a*). Both experiments had a similar overall relative response bias (*z*/*a*), and relative response bias (*z*/*a*) was affected by local noun number in both experiments, though in speeded button pressing this was in the form of a main effect and in speaking this presented in interaction with integration. In both experiments, local plural nouns elicited a bias slightly toward error-prone responding, though in Experiment 1 this bias was modulated by contrasting grammatical and notional number.

Comparing Experiment 4 and Experiment 5 allows us to look at the contrast between the speeded and no-pressure button pressing tasks. The untimed responses tended to be substantially slower, with a wider response range (See Table 14). Additionally, errors in the no-pressure button task were faster than correct responses in the local-compatible conditions. The biggest difference between the two tasks was that the no-pressure button pressing task was extremely conservative, with a large response conservativeness parameter (*a*) that increased with notional plurality. There were also reduced differences in the no-pressure button task between conditions in the amount of errors elicited and in the magnitude of the controller number selection parameter (v), though the relative difficulty of conditions was similar.

Comparing performance on the catch trials across experiments demonstrates a relationship to attraction in critical trials and morphological errors in preamble repetition on catch trials. This relationship was consistent across experiments, and suggests that processing difficulty lead to some of the agreement errors observed. However, this

relationship did not hold for non-morphological errors, suggesting that it is not general error-prone comprehension that elicits attraction, but something particular to morphological cues to number information. Rather than simple "noise" in the system, there is something systematic about ability to process inflectional information, and this information is used across modalities.

Though the experiments differed in the magnitude of the predicate compatibility effect in critical trials, and also differed in rates of correct predicate adjective use in catch trials, there was no relationship between the two by experiment or by speaker. Experiment 4 showed the strongest predicate compatibility effect, but had more predicate adjective errors than Experiment 1. This suggests that the differences in predicate compatibility effect across experiments do not boil down to simple attention (or inattention) to predicates, but are connected with something more nuanced. Again, this suggests that there are not clear lines to be drawn between the more production-oriented and more comprehension-oriented tasks.

These facts are underscored by the similarities across experiments in the primary diffusion parameter, controller number selection (v), which show us that similar information is used across experiments. This suggests that information use is similar in agreement production and in prediction for comprehension, consistent with Dell and Chang (2013) and MacDonald (2013). What is difficult in the typical preamble-completion task relies to some extent upon preamble comprehension—preamble repetition was not perfect, and in fact, varied in systematic ways.

However, there was one key difference between the three experiments, appearing on the controller number selection parameter (v). In the speeded button pressing task,

there was a stronger role of lexical information than in speaking, suggesting that agreement comprehension relies more heavily on lexical cues than agreement production. This is consistent with the cue based memory retrieval model of agreement comprehension (e.g. Dillon et al, 2013), though note that structural information also played an important role in both button pressing tasks, suggesting that prediction can also rely on structural information. In addition to a reliance of production upon comprehension, structural comprehension relies (to a degree) upon structural production.

The large difference in response speed and accuracy in these experiments was primarily captured by the response conservativeness parameter (*a*), replicating work in other domains that shows *a* to reflect pressure to respond quickly versus accurately (e.g. Ratcliff & McKoon, 2008). There are also hints towards a role of *a* in accounting for notional agreement—if *a* is lowered, notional agreement follows. This is further explored in the following chapter.

Response bias (*z)* changed minimally across conditions and experiments, though note that allowing it to vary by participant and condition improved model fits. The parameter *z* may pick up differences in how participants approach the task, and should be explored in future work by examining how response bias changes across participants.

The following chapter takes a first step toward examining how participant strategy differences affect these parameters *a* and *z*. In the following chapter, the contributions of the parameters *a* and *z* will be further explored in agreement by altering incentives to be accurate, in order to determine their connection to response monitoring and error production.

CHAPTER 7

EXPERIMENT 6: MONITORING AGREEMENT

It seems to be a common naive belief that speech errors arise because speakers are inattentive. This is the view exemplified by prescriptive style guides, such as Strunk and White's *The Elements of Style*, and Safire's New York Times column *On Language*. The truth in this view is that speakers tend to be fairly good at avoiding errors, and when they do make errors, they often notice them. However, a monitor may not be necessary to explain the error patterns that are produced, making the role of the monitor in language production somewhat controversial.

Experiment 6 was designed to examine the role of monitoring in agreement by examining how patterns of responding changed when speakers are made aware of the grammatical rules behind agreement. The idea is that this can dissociate notional number agreement (not an error) from attraction (an error), demonstrating if and when speakers notice mistakes in their own speech.

This manipulation also has the advantage of creating a between-subjects group of the sort that has previously been shown to affect the non-drift rate parameters of non-decision time, response conservativeness, and response bias. These often vary by individual and task-differences in more canonical decision tasks (e.g. Ratcliff et al, 2012; Ratcliff, Thapar & McKoon, 2006; Spaniol, Madden & Voss, 2006), and manipulating such a property will allow the further examination of the decision processes behind agreement and their mapping to the model.

Monitoring has been invoked as a driver of a wide range of biases in speech

errors, including the lexical bias, in which slips are more likely to occur if they produce words (e.g. *barn door* slips to *darn bore* more than *dart board* slips to *bart doard*; Baars, Motley, & McKay, 1975), though note that this pattern could also occur due to interactivity between semantic and phonological information (e.g Dell, 1986). Monitoring has also been invoked to explain the fact that errors tend not to create taboo words: *Tool carts* is more likely to slip that *Tool kits* (e.g. Motley, Camden, & Baars, 1981, 1982). More qualitatively, speakers regularly notice that they make mistakes and correct their own speech (e.g Levelt, 1983). These all point to a potential role for some type of speech monitor, either internal (pre-speech) or external. However, the extent to which this monitor impacts the patterns of errors typically produced is under debate (compare Hartsuiker, Corley, & Martensen, 2006 with Nozari & Dell, 2009; and compare Levelt Roelofs, & Meyer, 1999 with Dell, 1986).

In agreement, there is some evidence for monitoring and repairing speech. Speakers correct themselves occasionally when they make an agreement error. For example, in Experiment 1, there were 46 cases in which participants produced a verb correction in a critical trial (out of 3072 utterances). Forty of these changes involved changing an incorrect plural verb to a correct singular one, and 31 of the changes to a correct verb were in the local-plural condition. This suggests that attraction errors in particular are noticed at least some of the time.

The role of monitoring other types of information in agreement is less clear--Hartsuiker (2006) points out that monitoring of semantic information in agreement is hard-pressed to explain notional agreement, drawing a distinction between grammatical number and notional number in monitoring for errors. The evidence he uses to explain

this shows a relationship between working memory load and attraction errors, but not between working memory load and notional agreement (Hartsuiker & Barkhuysen, 2006). If we assume that the working memory burden involves some cognitive processes also used in monitoring (as Hartsuiker, 2006, does), this would suggest a dissociation between notional agreement and attraction. Error monitoring picks up only attraction, not notional agreement. This fits with the idea that the former is an error, and the latter is not.

In the present experiment, participants were given a grammar test either before or after they performed the unspeeded RSVP completion task described in the previous chapter (Experiment 5). The grammar pretest is theorized to encourage monitoring for errors, and the prediction is that those who receive the grammar pretest will make few mistakes and will respond slowly, while those who receive the grammar posttest will replicate the Experiment 5 response pattern. Furthermore, this paradigm will allow the separation of notional agreement and attraction errors, as outlined above: If notional agreement is not an error, and if notional information is not monitored, then it is predicted that the attraction rate will diminish more than the notional agreement rate.

This experiment will also serve as a further exploration of the diffusion parameters in agreement. Previous work has provided tentative support for monitoring external to the primary decision process: Nozari and Dell (2009) explored the role of self-monitoring in a lexical decision task using a variant of the diffusion model (EZ-diffusion), and found that the parameter *Ter (*non-decision time) was affected by editing in speech, but *MDT* (the remainder of reaction time, and a combination of *v*, *a*, and *z*) was not. The different modeling procedure, which did not examine differences in *v, a,* or *z,* means that these results may not directly inform the present study. However, it generates

a prediction that though reaction times may be slower for the grammar pretest group, slowing would be reflected on *Ter* and not on any other parameter (*v, a,* or *z).*

Counter to this hypothesis, it is also possible that monitoring-related slowing is due to decision-internal processes, suggesting an internal monitor. In this case, monitoring would affect response conservativeness (*a),* or response bias (*z).* This is to say that responses may become slower because participants are performing agreement itself more cautiously, or have altered their bias away from singular responding. The prediction from this hypothesis is that the grammar pretest group would have a larger *a* or smaller *z* than the grammar posttest group.

Method

Participants. Data were collected from 480 workers on Amazon Mechanical Turk with IP addresses in the USA. All had done more than 1000 hits with over 95% approval, and were paid $1.50. Out of these participants, 52 were excluded for demographic reasons (46 for self-reporting being left-handed, and 6 for self-reporting learning a language before English). An additional 35 were excluded due to performance (10 for quitting the task before the end, 6 for performance below chance on the grammar test, 16 for performance below chance on the RSVP task, and 3 for having a difference between their mean and median RT of over 30 seconds, suggesting the presence of many slow trials in which they were not paying attention). Data were also excluded from participants who had already completed the task, eliminating 9 subject-runs. These data came from one participant who completed the task four times and from five participants who completed it twice. This left 384 participants, 192 in each group.

Equipment. The experiment was run through Amazon Mechanical Turk on the

participants' home computers, using a script written in the IbexFarm platform (http://spellout.net/ibexfarm/). This platform is open source and uses JavaScript to time script events in a user's browser. IbexFarm functions for RSVP and timed acceptability tasks were adapted for the present experiment.

Procedure. All participants performed the no-pressure RSVP experimental task outlined in the previous chapter, with written catch trials. Half of the participants (N=192) were given a pretest in which they were asked to identify the grammatically correct option out of pairs of sentences, and the other half (N=192) were given the same as a posttest.

Materials. Experimental materials for the RSVP task were identical to Experiment 1. There were 24 experimental items, varying in integration, local number, and predicate compatibility, all fully crossed. See Table 1 for example stimuli.

The grammar test contained subject-verb agreement violations (See Appendix E). These stimuli were made up of simple noun phrases, conjoined noun phrases, and noun phrases with prepositional phrase modifiers, varied on notional number and balanced across the four combinations of head and local noun number (singular-singular, singular-plural, plural-singular, plural-plural). Verbs in this test were varied between auxiliary verbs, semantically-light lexical verbs, and semantically-heavy lexical verbs, one third of each. Half of the items had a correct singular response and half had a correct plural response, with answers counterbalanced such that the correct answer was listed first half the time and each verb response was listed first half of the time. The same list was used for all participants.

Procedure. The procedure was similar to Experiment 5. Stimuli were presented

one word at a time in an RSVP paradigm ending in a two-choice button-press decision to select the fragment's appropriate verb continuation, by pressing "F" for 'was' and "J" for 'were'. There were catch trials interleaved with these standard trials that involve typing full sentence completions in lieu of the forced-choice decision. See Chapter 6 for more details on the paradigm. For the grammar test, the instructions were displayed on one screene, followed by a single screen per question. Participants were allowed to take as much time as they wished and moved between screens by pressing a "submit" button, allowing them to change their answer before moving on if they so desired.

Scoring. As in Experiment 5. For standard trials, responses of "F" were coded as singular, and responses of "J" were coded as plural. For catch trials, verbs were scored as singular or plural if they bear number marking (e.g., copular verbs such as "is" or "were", regular present tense verbs such as "seems"). Predicate adjectives and preambles were scored as correct or incorrect, with leniency for changes preserving meaning, structure, and morphology.

Unlike in the previous experiments, participants were not supervised by an experimenter as they ran through the experimental task. This meant that there was an increased potential for trials with invalid reaction times due to the participant failing to pay attention to the task. To address this, all data points with reaction times above 20 seconds were excluded from the data set (N=23, 0.2% of trials), as were all participants who had mean scores over 30 seconds above their median score, suggesting the presence of contaminant reaction times (N=3 participants). Once these trials and participants were excluded, two other trial exclusion cutoffs were calculated, at three standard deviations (eliminating 160 trials, or 1.7% of the data) and four standard deviations (eliminating 47

trials, or 0.5% of the data) above participants' means. Diffusion models were run on both data sets, and data with the three standard deviation exclusion criterion are reported for reasons outlined in the Analysis section.

Design. As in Experiment 1, each participant received one of 16 experimental lists for the RSVP task, each list containing one version of all 24 items, three in every condition, and with pairs of lists counterbalancing item presentation order. Every list was presented to 24 participants, so that every item was tested on 48 participants. These within-subject variables were fully crossed with the grammar test manipulation. The fixed effects in the statistical analyses were integration (integrated-unintegrated), lexical-semantic compatibility (head-local), local-noun number (singular-plural), and grammar monitoring (pretest-posttest), all fully crossed.

Analysis. Diffusion models were fitted using the fast-dm program, as outlined above. However, as mentioned in the previous section, there were issues with contaminant reaction times in this experiment. As a result of this, two sets of candidate models were run, based upon a data set with a three standard deviation over subject mean exclusion criterion, and a data set with a four standard deviation over subject mean exclusion criterion. From these data sets, data were aggregated in to the same randomly chosen super-subjects (three per experimental list). Five candidate models were run in each of these two data sets, (a full model, a minimal model, and the three models eliminating response conservativeness (*a*), response bias (*z*), or *Ter*). Using the four standard deviation exclusion criterion data set, the model converged on invalid parameter estimates for a number of subjects (e.g. relative response bias *z/a* outside of the 0 to 1 range, and/or response bias *z* or non-decision time *Ter* outside of the valid range for a

given supersubject as specified by the internal variability parameters *sz* or *st0*). These invalid estimates are likely to be due to slow reaction times that mathematically, do not belong to the same reaction time distribution as the bulk of the data. For this reason, the more conservative three standard deviation exclusion criterion was selected, which ended up trimming out 1.7% of the data.

For the three standard deviation exclusion criterion data set, the full model had the highest average K-S *p*-value (0.82), followed by the fixed criterion model (0.77), the fixed *Ter* model (0.74), and the fixed bias model (0.72), with the minimal model trailing far behind (0.21). All models tended to under-predict reaction times, but the fixed *Ter* model was the least innacurate (See Appendix C). As in previous experiments, the fixed *Ter* model was selected for analyses, as it provided the best-recovered data and a similar fit by subject as the maximal model (see Appendix D).

Results

The pretest and posttest groups had comparable accuracy levels (15% plural responses). However, responses from the pretest group were slightly faster than those from the posttest group, for correct responses (pretest 1430 to posttest 1480) and for errors (pretest 1739 to posttest 1757). This speeding of the pretest group was particularly pronounced in unintegrated fragments (See Figure 16). Compared to the in-lab unspeeded-response RSVP experiment, the posttest group was slower and slightly more accurate. (Posttest mean correct RT =1480 ms, 15% plural response; Experiment 5 mean correct RT = 1234 ms, 10% plural responses).

Accuracy and latency results for the within-subject variables replicated previous experiments. As in the previous experiments, local plural nouns elicited slower correct

responses and more errors than local singular nouns (Local plural: mean correct RT = 1439 ms, 9% plural responses; Local singular: mean correct RT = 1475 ms, 21% plural responses), and unintegrated fragments elicited slower correct responses and more errors than integrated ones (Unintegrated: mean correct RT = 1381 ms, 9% plural responses; Integrated: mean correct RT = 1542 ms, 21% plural responses; See Figure 16). Predicate compatibility affected latencies but not errors (Local-compatible: mean correct RT = 1470 ms, 15% plural responses; Head-compatible: mean correct RT = 1441 ms, 15% plural responses). See Table 16 for mean latencies by condition; see Figure 16 for mean correct latencies and proportion error by condition.

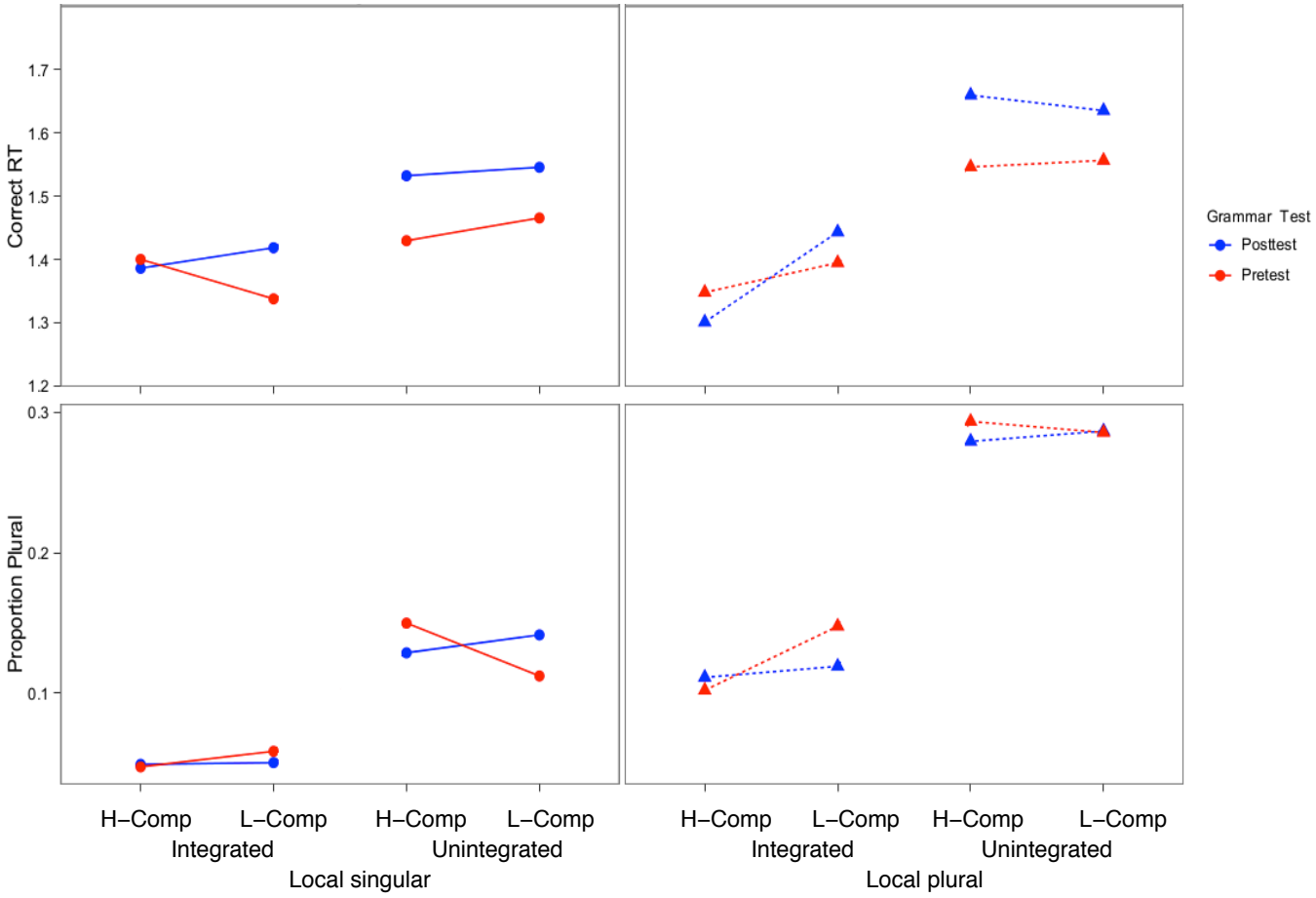


Figure 16. Mean singular (correct) response latencies in seconds (top) and proportion of plural (error) responses (bottom) in Experiment 6 varying by local number, integration, predicate compatibility, and grammar test group.

Table 16.
Response latencies (in milliseconds) from Experiment 6.

Grammar test group	*Integration*	*Predicate compatibility*	*Singular Response*		*Plural Response*	
			Local Singular	*Local Plural*	*Local Singular*	*Local Plural*
Posttest	*Integrated*	*Head-Compatible*	1386	1302	1621	1952
		Local-Compatible	1418	1443	1731	1730
	Unintegrated	*Head-Compatible*	1533	1660	1695	1674
		Local-Compatible	1546	1635	1627	1894
Pretest	*Integrated*	*Head-Compatible*	1400	1348	1802	1566
		Local-Compatible	1350	1402	1898	1927
	Unintegrated	*Head-Compatible*	1430	1546	1809	1739
		Local-Compatible	1466	1557	1692	1641

Diffusion results largely mirrored previous experiments, with minimal differences between grammar test groups. On the controller number selection parameter (v), there was an interaction between local number and integration, such that local plurals with unintegrated phrases were more difficult than those with integrated phrases, as well as main effects of local number and integration (see Figure 17, Table 17). This replicated the results of Experiment 1 and Experiment 4. Additionally, as in Experiment 5, there was also an interaction between local number, integration, and predicate compatibility such that the local plural-unintegrated-local compatible condition was easier than otherwise predicted (See Figure 17, Table 17). There was also a marginal interaction between local noun number and grammar test, such that the pretest group was slower than the posttest group when local nouns were singular (see Figure 17, Table 17).

For the response conservativeness parameter (a), there was an interaction between integration and local number, such that the local singular-integrated and local-plural unintegrated conditions elicited the most conservative responses. This went along with a

main effect of local number, such that responses were more conservative for local singulars. There was also a marginal interaction between integration, predicate compatibility, and grammar test group such that the head-local compatibility difference was present for the pretest group across levels of integration and for the posttest group when preambles were integrated, but was absent for the posttest group in unintegrated preambles. See Figure 17 for estimates by condition and Table 17 for analysis results.

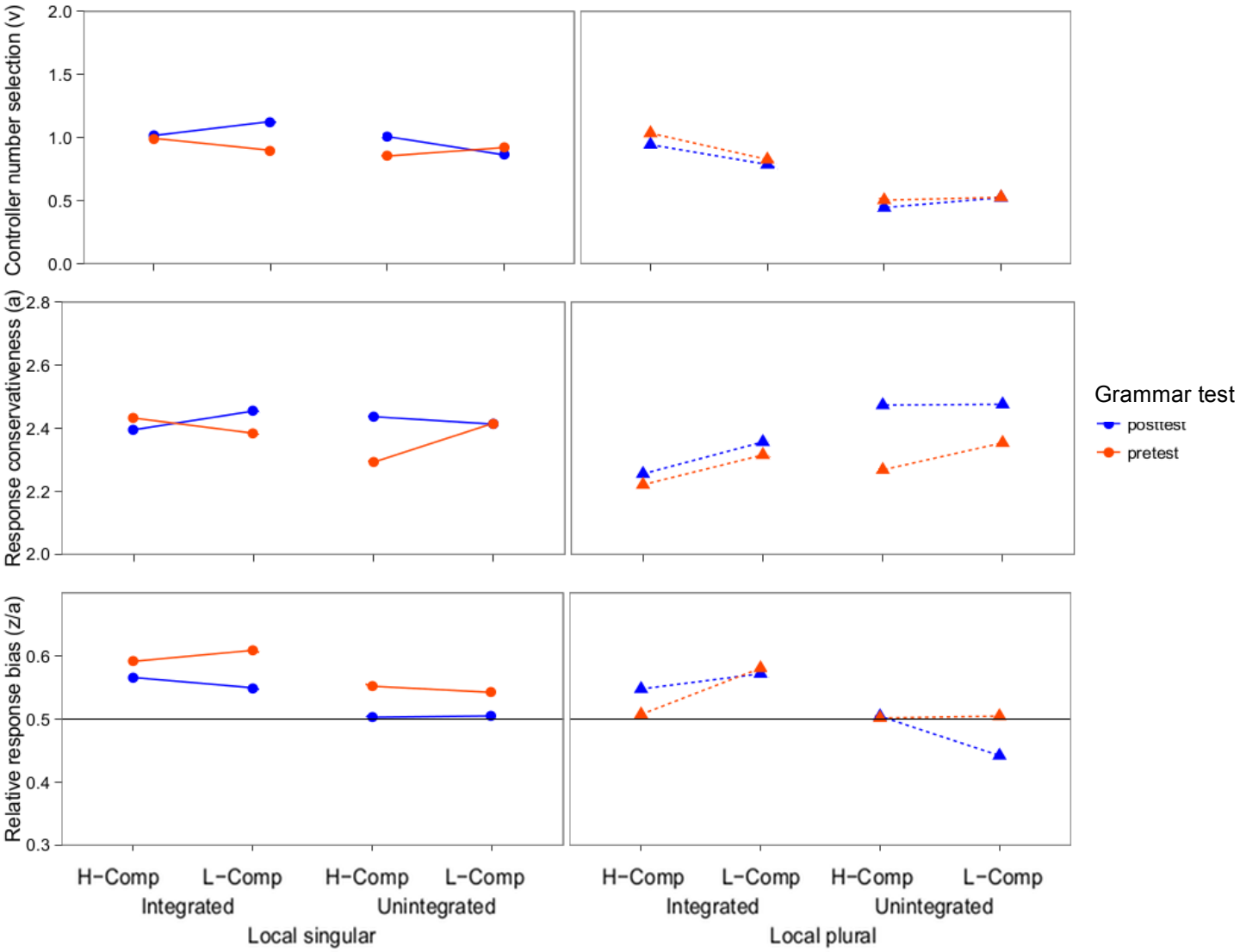


Figure 17. Fast-dm parameters from fixed *Ter* model for Experiment 6. Top: controller number selection (*v*), middle: response conservativeness (*a*), bottom: relative response bias (*z*/*a*). Larger *v* reflects ease of agreement decision, larger *a* reflects a more conservative response decision, and larger *z*/*a* reflects a bias towards singular responses.

As in previous studies, the relative response bias parameter (z/a) disclosed a slight bias toward singular responses (z=.54a) and an interaction between integration and predicate compatibility such that the difference between head and local compatible predicates was largest in the unintegrated condition. There was also a marginal main effect such that pretest group had an increased bias toward singular responses.

The non-decision time parameter *Ter* was allowed to vary by supersubjects only in the model presented here (pretest: 567 ms, posttest: 533 ms). It did not vary significantly by grammar test group, $t(94) = 0.84$, $p = 0.40$.

Table 17.
Diffusion model parameters from Experiment 6 from fixed Ter diffusion model. P-values are approximated from a standard normal distribution.

	Controller number selection (v)				Response conservativeness (a)				Relative response bias (z/a)			
	Estimate	S.E.	t-value	p(z)	Estimate	S.E.	t-value	p(z)	Estimate	S.E.	t-value	p(z)
Intercept	0.83	0.03	25.65	< 0.001	2.37	0.05	48.43	< 0.001	0.54	0.01	77.80	< 0.001
Local number	-0.26	0.04	-6.58	< 0.001	-0.06	0.03	-2.07	0.04	-0.03	0.01	-2.77	0.01
Integration	0.25	0.04	6.09	< 0.001	-0.04	0.03	-1.20	0.23	0.06	0.01	5.07	< 0.001
Predicate compatibility	0.04	0.04	1.04	0.30	-0.05	0.03	-1.83	0.07	0.00	0.01	-0.35	0.73
Grammar test	-0.02	0.06	-0.29	0.77	-0.07	0.10	-0.74	0.46	0.02	0.01	1.81	0.07
Local number x integration	0.30	0.07	4.21	< 0.001	-0.13	0.05	-2.62	0.01	0.01	0.02	0.51	0.61
Local number x pred comp	0.05	0.08	0.60	0.55	-0.04	0.05	-0.85	0.40	-0.01	0.02	-0.56	0.58
Local number x gram test	0.13	0.08	1.69	0.09	-0.06	0.06	-0.93	0.35	-0.04	0.02	-1.54	0.12
Integration x pred comp	0.09	0.08	1.24	0.22	0.00	0.05	-0.09	0.93	-0.04	0.02	-2.08	0.04
Integration x gram test	-0.02	0.08	-0.26	0.80	0.09	0.07	1.38	0.17	-0.02	0.02	-1.01	0.31
Pred comp x gram test	0.03	0.08	0.33	0.74	-0.03	0.05	-0.54	0.59	-0.03	0.02	-1.50	0.13
Loc number x integration x pred comp	0.28	0.13	2.08	0.04	-0.10	0.10	-0.97	0.33	-0.07	0.04	-1.86	0.06
Loc number x integration x gram test	0.11	0.14	0.78	0.44	0.07	0.10	0.70	0.48	-0.04	0.04	-1.13	0.26
Loc number x pred comp x gram test	0.06	0.17	0.34	0.74	-0.02	0.10	-0.18	0.86	-0.05	0.04	-1.17	0.24
Integration x pred comp x gram test	0.21	0.15	1.37	0.17	0.17	0.10	1.70	0.09	-0.01	0.04	-0.37	0.71
Loc number x integration x pred comp x gram test	-0.42	0.27	-1.57	0.12	-0.17	0.20	-0.81	0.42	0.06	0.08	0.78	0.44

Discussion

This experiment was designed to examine the role of attention and monitoring in agreement production, encouraging half of the participants to attend to agreement errors by giving them a grammar test before the experimental task began. The idea was that this pretest would create an individual difference variable of the sort that have been previously shown to affect the non-drift rate diffusion parameters (non-decision time, response conservativeness, and response bias; e.g. Ratcliff et al, 2012; Ratcliff et al, 2006; Spaniol et al 2006).

However, there were minimal differences between grammar test groups, and those that were present were in the opposite direction than predicted. Both groups were equally accurate, but the pretest group was slightly faster than the posttest group, particularly in the unintegrated conditions. This suggests that attending to the agreement decision may make agreement easier to carry out, rather than making speakers more cautious. This may relate to notional variations in particular, not to attraction, as the between-group difference was strongest in the notionally-plural unintegrated condition. In particular, this difference may reflect practice or priming of some sort.

The diffusion results also did not vary much by grammar test group. There were marginal effects of grammar test group on all decision-internal parameters, suggesting that if monitoring is occurring, it is internal to the agreement process. In particular, the small speedup of the grammar pretest group may be due to an increased bias toward singular responding. Major caveats apply, given the large number of subjects and the merely marginal results, as well as the issues with fitting diffusion models to data sets with contaminant reaction times, but a replication with a stronger incentive to monitor for

errors would be worth pursuing.

Both participant groups replicated the Experiment 5 results. All parameters were in a range similar to Experiment 5, suggesting the replicability of the general procedure, even in a non-lab setting. Again, this experiment demonstrates that agreement can be modeled as a decision process, and that the main parameter of this process, controller number selection, occurs by combining grammatical number with lexical and notional sources of information.

CHAPTER 8: GENERAL DISCUSSION

In these six studies, we examined the ways that linguistic information (including grammatical number, notional number, and lexical-semantic relatedness) is used in subject-verb agreement and the way that it interacts with difficulty in other domains, including number cognition, memory, and monitoring. Using a combined measure of accuracy and speed, we looked at how number-inflected verbs were produced after complex subjects. Our aim was to evaluate the predictions of structural and lexical accounts of number agreement and to assess the interactions between agreement and other cognitive domains.

In Experiment 1, sentence subjects had singular head nouns with local nouns that varied between singular and plural (e.g. *The phone with the broken toaster; The phone with the broken toasters)*. This local number variation created the standard attraction effect, with local plural nouns eliciting increased rates of erroneous plural agreement and a slight slowing of responses. Notional number, in the form of referential integration, also varied: Preambles were highly integrated (e.g. *The phone with the missing button)* or unintegrated (e.g. *The phone with the broken toaster)*. Consistent with their expected plural notional valuation, unintegrated preambles increased erroneous plural agreement and severely slowed response latencies. Lexical-semantic compatibility was manipulated by using predicates that were semantically well suited to properties of the head noun (e.g. *ringing* goes with *phone* but not *button* or *toaster)* or the local noun (e.g. *plastic* goes with *button, shiny* goes with *toaster*). This manipulation was relatively ineffective, with local-matching predicates barely promoting erroneous agreement without increasing latencies.

To reconcile differences in the latency and error outcomes, we used diffusion modeling (Ratcliff, 1978; Voss & Voss, 2007) to derive a measure of general difficulty. This measure suggested that preambles containing local singular nouns (e.g. *The phone with the broken toaster; The phone with the missing button)* were uniformly easy, regardless of notional number or lexical-semantic compatibility. However, when a plural local noun was present, notionally plural preambles (e.g. *The phone with the broken toasters)* were more difficult than notionally singular preambles (e.g. *The phone with the missing buttons)*. Incompatible predicates, however, were only negligibly harder than those that were compatible.

In Experiment 2, sentence subjects were made up of conjoined noun-phrases. The phrases contained singular first nouns combined with second (local) nouns that varied between singular and plural (e.g. *The dish and the plate; The dish and the plates)*. Local number affected error rates, with local plurals (consistent with the expected verb number) associated with increased rates of correct plural agreement, but again, with minimal impact on latencies. Concreteness, a notional variable that affects the ease of individuating (counting) referents, also influenced agreement: Concrete, notionally plural preambles (e.g. *The dish and the plate)* elicited more correct plural agreement than abstract, notionally singular ones (e.g. *The hypothesis and the theory)*, but with no effect on latencies. Finally, lexical-semantic relatedness of the two conjoined nouns (e.g. related, *The dish and plate*; unrelated, *The dish and cat)* had no effect on either error rates or response latencies. The generalized difficulty measure mirrored these patterns: Local plural nouns were easier than local singulars, concrete referents were easier than abstract ones, and lexical-semantic relatedness had little effect.

In Experiment 3, sentence subjects varied on the grammatical and notional specification of a quantifier, as well as the spatial distribution of item images. Preambles containing a grammatically-singular specified quantifier (e.g. *One alligator with humungous claws)* elicited fewer errors than those containing a number-unspecified quantifier (e.g. *The alligator with humungous claws)*. Similarly, preambles with notionally-singular quantifiers (e.g. *One alligator with humungous claws)* elicited fewer plural responses than preambles with notionally-plural quantifiers (e.g *Every alligator with humungous claws)*. These did not straightforwardly impact response speed. A diffusion analysis mirrored the error results, with grammatical-specification and notional number interacting on controller number selection (*v*). Spatial information affected response bias (*z)* when it matched quantifier grammatical specification, such that specified singular quantifiers elicited a larger bias toward singular responses in close-spread arrays, and a smaller bias toward singular responses in far-spread arrays.

In Experiments 4 and 5, the stimuli were identical to Experiment 1, with variations in notional, grammatical, and lexical information. Experiments 4 and 5 both used variants on a button-pressing task that is theorized to blend elements of comprehension and production. In Experiment 4, participants were made to respond quickly with a time-cutoff, while in Experiment 5, participants were not pressured to respond fast. In Experiment 4, the results of Experiment 1 were largely replicated, though Experiment 4 elicited a stronger effect of predicate bias, in overt errors and in the controller number selection parameter (*v*). This supports a larger role of lexical information in comprehension than in production. Experiment 5 was extremely slow and extremely accurate, with reduced differences in the controller number selection parameter

(*v*) and more conservative responding that increased in conjunction with notional plurality (*a)*.

Experiment 6 used the same stimulus set as in Experiment 1 and the same task as in Experiment 5 in conjunction with a grammar test. This was designed to encourage monitoring for errors and to encourage dissociation of notional agreement and grammatical errors. There were only marginal differences between grammar test groups, with the data largely replicating Experiment 5. The small differences between groups were in line with process-internal monitoring that is differentially sensitive to notional number, but no firm conclusions can be drawn from these data.

The overarching goal in the experiments was to determine the contributions of structural (wholistic) and lexical (piecemeal) information in agreement, as well as cognitive factors (visual distribution, memory, and attention). First we turn to the comparison of structural and lexical information use. Agreement involves linking features of controllers and targets. In processing terms, this could be accomplished on the basis of individual words or through the syntactic representations constructed from global message properties. The issue that separates lexical and structural accounts is which of these is the dominant force in agreement.

Implications for structural accounts of agreement

Broadly speaking, the findings favor structural over lexical control as the more critical force in the production of agreement. In a structural account, agreement is one product of a process in which evaluation of notional number in a mental model of the subject's referent accompanies the unpacking of a non-verbal message into linguistically viable pieces. Sentences are composed with an emerging syntactic frame that supports

yet-to-be specified morphological and phonological properties (e.g., Dell, 1986). As this mapping and unpacking occur, agreement is created through a combination of referential indices, structural properties, and words. Agreement difficulty arises when the notional number of the subject's referent conflicts with the grammatical number associated with retrieved pieces of words.

Our test of this account relied on three manipulations of notional number, referential integration in Experiment 1, 4, 5 and 6; concreteness in Experiment 2; and quantifier notional number in Experiment 3. These notional factors were designed to alter message-level properties in similar ways: Low levels of integration and high levels of concreteness tend to individuate referents and make them notionally more plural. Quantifiers determine a subset of items, affecting the notional number of the message. In company with the notional factors, we also manipulated semantic properties associated with individual lexical items. Critically, we also manipulated the grammatical number of the local noun, which reliably affects the outcome of agreement—verb number. The evidence for the structural account comes from the overriding impact of notional number on this outcome. That is, a property of the sentence subject as a whole swayed the linguistic signals of agreement.

In Experiment 3, we also manipulated a visual representation of referent number, through spatial properties of a referent display. This interacted with grammatical specification on response bias. This suggests that not only does notional information conveyed in language affect number agreement, so may actual display properties reflecting the message at hand. This is consistent with a structural account of number agreement in which notional number is calculated from the world and the speaker's

mental model of it.

From a structural perspective, the basic mechanisms of grammatical number agreement are rooted in a speaker's construals of numerosity within an intended message. On this account, the meanings of individual words should play a subsidiary role in the production of agreement. The outcomes aligned well with the structural view. The relative ease of agreement, in terms of speed and accuracy combined, was more heavily determined by how notional number meshed with grammatical plurality than by the lexical-semantic properties of component words. This interplay between notional and grammatical number is at the heart of structural arguments about language production (e.g. Bock & Ferreira, 2014).

Implications for lexical accounts of agreement

In contrast to the structural view, a lexicalist account predicts that lexical-semantic associations among words in the sentence are central to the production of agreement. The findings of the experiments suggest limits on the workings of lexical factors. A lexical account of sentence production and agreement is one that hinges on assembling components. Words must first be retrieved, then pieced together into a syntactically acceptable sequence, relying on the words' co-occurrence restrictions (subcategorizations). Finally, the features of words in certain positions must be given agreeing values. On this view, the locus of agreement difficulty is in assembling the pieces of a phrase and identifying an agreement controller.

The lexical factors that we manipulated have been shown to impact language production in previous work. The lexical-semantic compatibility variable in Experiment 1 was inspired by Thornton and MacDonald's results (2003). Their basic finding was that

predicates that could plausibly modify a local noun promoted attraction. When paired with *The album by the classical composers*, the predicate *praised* was harder than *played,* as shown by increased agreement errors and slower reading times. Thornton and MacDonald interpreted this result in terms of the lexical linkages that follow from plausibility. In Experiment 1, we aimed for an operationalization of these factors that rested on lexical semantic compatibility. A weak trend appeared in the same direction seen by Thornton and MacDonald, but the effect was weak and clearly secondary to the influence of notional number.

In Experiment 2, we manipulated lexical-semantic relatedness between the head and local nouns. In previous work, it has been shown that semantic relationships between successive nouns increase interference in production of later nouns (e.g., Damian, Vigliocco & Levelt, 2001; Howard, Nickels, Coltheart, & Cole-Virtue, 2006; Schriefers, Meyer, & Levelt, 1990; Rahman & Melinger, 2007; Wheeldon & Monsell, 1994). In agreement, lexical competition has been hypothesized to raise processing difficulty as the speaker prepares to produce the second noun (Solomon & Pearlmutter, 2004). This in turn disrupts selection of a lexical controller, especially in the presence of a mismatching grammatical number feature (e.g., as outlined by Dillon et. al., 2013). Counter to this hypothesis, despite the presence in Experiment 2 of the kinds of relationships that promote lexical interference, notional number was the only dominant factor. Veenstra (2014) obtained this same agreement result in the presence of a direct measure of accompanying interference.

There was one provocative hint in Experiment 2 about a potential lexical disruption to agreement. In a small subset of items, the conjunction of nouns that formed

the sentence subject had only one determiner (e.g. *Their destiny and fate*). In these items, semantic relatedness between the nouns impeded the production of an agreeing verb, implying that lexical competition within a tighter or shallower constituent may be more likely to interfere with the implementation of agreement (cf. Gillespie & Pearlmutter, 2011). This possibility deserves further attention.

Agreement by content-addressable memory retrieval

Although the outcomes of the present experiments run counter to certain claims about lexical bases of agreement, there are hypotheses about lexical mechanisms that our research was designed to address only indirectly. One of them centers on the retrieval from memory of the sentence elements that play a role in agreement (e.g., Dillon et al, 2013; Lewis & Vasishth, 2005). On this account, retrieval cues (such as noun and verb inflections) serve as pointers to other words that can then be directly retrieved from a content-addressable memory store. For implementing number agreement, a target verb might cue retrieval of a previously occurring noun controller in order to recover the noun's number. The noun number can then guide the selection of the verb's inflection. If a cue leads to retrieval of a wrong noun, agreement problems like attraction may result.

A retrieval process of this kind has been proposed for explaining disruptions in language comprehension that resemble agreement attraction in production. In comprehension, a mismatch between a local- or non-head-noun plural and an ungrammatical plural verb can slow reading, for instance when the verb *were* is encountered in a sentence like *The key to the cells unsurprisingly were rusty from many years of disuse*. Existing findings about the patterns of these disruptions is more consistent with difficulty in a lexical-retrieval process than with retrieval from a syntactic

structure or message-based representation (Wagers et. al., 2009; Tanner et al, 2014; see Shen, Staub, & Sanders, 2013 for analogous results in listening comprehension). Lexical retrieval similarly offers an account of comprehension problems in syntactically difficult structures (e.g. Lewis & Vasishth, 2005) and of differences between subject-verb agreement comprehension and reflexive dependency resolution (e.g. Dillon et al., 2013).

Notably, the verb and reflexive differences in attraction found in Dillon et al. are absent in the production results of Bock et al (2006). In the latter, a structural relationship offers a better account of both verb and reflexive attraction. This and other results cast doubt on whether memory retrieval plays a critical role in producing agreement. In production, the best evidence for the use of a retrieval mechanism comes from work on grammatical gender agreement by Badecker and Kuminiak (2007) in Slovak. The problem with generalizing the Badecker and Kuminiak findings to number agreement lies in the fact that grammatical gender agreement is heavily dependent on specific lexical relationships, much more so than number agreement. In grammatical gender languages like Slovak, there is little semantic basis for gender agreement among nouns and the gender inflection system, apart from nouns with human referents. The deep roots of number inflection in number semantics, and the evidence that number semantics is important to producing agreement, limits the viability of a number agreement process that depends on specific words.

This is not to say that cue-based lexical retrieval is fully irrelevant to producing agreement. However, the weakness of the lexical effects in Experiments 1 and 2 is inconsistent with an agreement production system that is chiefly or exclusively based on word-based connections.

The critical difference between production and comprehension, and agreement comprehension in particular, is that accurate comprehension is *by necessity* a word-centered process, with the details of an apprehended meaning dependent on the occurrence of certain words. This makes lexical retrieval essential to precise comprehension. This difference between comprehension and production was addressed in Experiments 4 and 5. Comparing results between these experiments and Experiment 1 provides evidence that comprehension relies more on lexical factors than production does. However, other than that, the broad pattern of results between the two is similar, suggesting at least that in the task at hand, the prediction necessary to do a button-pressing task, and the comprehension necessary to produce a preamble completion can both also take structural information into account.

Agreement in a diffusion framework

The analysis technique in the present experiments, diffusion modeling, has been previously used in cognitive psychology to investigate tasks that involve explicit decision-making, such as lexical decision and recognition memory. Those tasks differ from agreement in several important ways. For example, lexical decision and recognition memory tasks involve a comparison between an existing memory represention and a single presented stimulus, eliciting a binary decision — deciding whether a single item is a word or non-word, or has or has not been seen. In contrast, agreement production involves an implicit decision embedded within an ongoing process of planning and articulating sentences. Despite these differences, the tasks have an important resemblance. All of them involve a discrete two-choice decision (yes/no; singular/plural).

This underlying similarity allows the use of diffusion modeling to break down the

components of a singular/plural selection process. Applied to agreement, the diffusion parameter *v* indexes the processes of reconciling number sources in the subject noun-phrase while removing extraneous sources of variance (*Ter*) and accounting for participant- and item-level changes in strategy shifts (*a* and *z)*. The application of the model here assumes that, like other processes examined with this technique (e.g. Ratcliff et al 2004; Ratcliff, 1978; Voss et al, 2008), language production recruits sets of domain-specific and domain-general cognitive mechanisms that can be modeled in terms of a choice mechanism. The present experiments called on diffusion analysis to overcome a problem that is ubiquitous in language production: Speakers continually sacrifice speed for precision and vice-versa. Fast speech tends to be error-prone speech. The tradeoffs can occur at many levels, from the formulation of messages through the selection of words to articulation. By combining measures of speed and accuracy, we aimed to more clearly disentangle the processes behind the implementation of agreement. Doing so helps to solve the reaction-time interpretation problem outlined in Pachella (1974) and opens the way toward a reliable measure of continuous performance in cases where errors are natural outcomes of a habitual activity.

Despite its strengths, one stumbling block in the diffusion analysis of agreement performance is the lack of as clear a mapping for the parameters *a* and *z*. The existing literature has validated these parameters with examination of individual differences such as age and explicit response strategy in traditional tasks (e.g. Ratcliff et al, 2012; Ratcliff, et al, 2006; Spaniol et al, 2006). In our application, without individual-difference variables, allowing variations in these parameters by condition improved the model fits (see K-S *p* values for all experiments and Appendix C and D).

The response conservativeness parameter (*a)* was most strongly affected by the removal of a time deadline in Experiment 5, suggesting a link between willingness to make errors and *a*. In addition, response conservativeness (*a)* also tended to be affected by conflicts between notional and grammatical number. This suggests a connection between monitoring, notional agreement, and the parameter *a*.

Variations in *z (*response bias) were considerably more minimal, though allowing this parameter to vary substantially improved model fits for all experiments (see Appendix C and D). The results of Experiments 1 and 4 suggested that relative response bias was informed by grammatical and notional number. This variation is consistent with differential weighting of positive and negative evidence (Voss et al. 2008) for a non-default, plural response. The results of Experiment 3 disclosed that response bias (*z*) is affected by arrangements of objects, suggesting a possible perceptual role for determining starting point in the agreement decision, and the results of Experiment 6 hint toward a role of response bias in monitoring for agreement errors.

A particularly promising application of diffusion modeling to language processing is in comparisons between sentence comprehension and production. As noted earlier, there may be different evidence-gathering processes at work in the two modalities that cannot be easily diagnosed with speed or accuracy measures alone. To explore underlying processing differences, a diffusion model has the advantage of mathematically combining latency and outcome measures to allow a decomposition of task-specific and task-general mechanisms. We began to approach this question in Experiments 4-6, the results of which suggested that there are differences in the balance of structural versus lexical information used in comprehension. Furthering this train of thought, a diffusion

analysis could also provide a uniform scale for results from diverse measures (e.g. reading time, Dillon et al., 2013; electrophysiology, Shen et al., 2013) and diverse tasks (e.g. the sentence completion tasks in the present experiments; maze reading, Nicol, et al, 1997; and speeded acceptability judgments, Nicol et al. 1997).

Conclusion

Language production involves combining structure and words. For the production of agreement, by using a statistical model that combines both speed and accuracy measures, we found that wholistic properties of sentence subjects were more powerful than lexical-semantic relationships in the creation of variations in difficulty. Though there must be a role for lexical information in agreement, it is evidently restricted. This follows naturally from the fact that before speakers start to talk, they typically have messages with internal relationships that demand language structure in order to be conveyed. It is only within those structures that words communicate intended meanings.

REFERENCES

Allik, J., & Tuulmets, T. (1991). Occupancy model of perceived numerosity. *Perception and Psychophysics*, 49(4), 303–314.

Allik, J., Tuulmets, T., & Vos, P. G. (1991). Size invariance in visual number discrimination. *Psychological Research*, 53(4), 290–295.

Baars, B.J., Motley, M.T., & Mackay, D.G. (1975). Output editing for lexical status in artificially elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior*, 14, 382-391.

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, 65–85.

Barner, D., Chow, K., & Yang, S. J. (2009). Finding one's meaning: A test of the relation between quantifiers and integers in language development. *Cognitive Psychology, 58,* 195–219.

Barner, D., Thalwitz, D., Wood, J., Yang, S., & Carey, S. (2007). On the relation between the acquisition of singular–plural morphosyntax and the conceptual distinction between one and more than one. *Developmental Science*, 10(3), 365–373.

Barwise, J., & Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy, 4*, 159-219.

Bates, D., Maechler, M., Bolker B., and Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-6. http://CRAN.R-project.org/package=lme4

Bergen, L., Levy, R., & Gibson, E. (2012). Verb omission errors: Evidence of rational

processing of noisy language. *Proceedigns of the 34th Annual Conference of the Cognitive Science Society, pp. 1320-1325.*

Bock, J. K., Butterfield, S., Cutler, A., Cutting, J. C., Eberhard, K. M., & Humphreys, K. R. (2006). Number agreement in British and American English: Disagreeing to agree collectively. *Language, 82*, 64-113.

Bock, J. K., Carreiras, M., & Meseguer, E. (2012). Number meaning and number grammar in English and Spanish. *Journal of Memory and Language*, 66, 17–37.

Bock, J., & Ferreira, V. (2014). Syntactically speaking. In M. Goldrick, V. Ferriera, & M. Miozzo (Eds.) *The Oxford Handbook of Language Production* (pp. 21-46). Oxford: Oxford University Press.

Bock, K., & Cutting, J. C. (1993). Regulating mental energy: Performance units in language production. *Journal of Memory and Language, 31*, 99-127.

Bock, K., & Eberhard, K. (1993). Meaning, sound, and syntax in English number agreement. *Language and Cognitive Processes*, *8*, 57–99.

Bock, K., & Miller, C. A. (1991). Broken agreement, *Cognitive Psychology*, 23(1), 45-93.

Bock, K., Eberhard, K. M., & Cutting, J. C. (2004). Producing number agreement: How pronouns equal verbs. *Journal of Memory and Language*, 51(2), 251-278

Bock, K., Eberhard, K. M., Cutting, J.C., Meyer, A. S., & Schriefers, H. (2001). Some attractions of verb agreement. *Cognitive Psychology*, 43, 83–128.

Bock, J. K., Nicol, J., & Cutting, J. C. (1999). The ties that bind: Creating number agreement in speech. *Journal of Memory and Language, 40,* 330 –346.

Brady, T. F., Konkle, T., Alvarez, G. A. and Oliva, A. (2008). Visual long-term memory

has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences, USA*, 105 (38), 14325-14329.

Brehm, L., & Bock, K., (2013). What counts in grammatical number agreement? *Cognition*, 128, 149-169.

Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The bank of standardized stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PloS ONE*, *5*(5), e10773

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84, 73–111

Clark, H. H., & Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology*, 37, 201–242.

Cohen, J.D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, and Computers*, 25(2), 257-271.

Damian, M. F., Vigliocco, G., & Levelt, W. J. M. (2001). Effects of semantic context in the naming of pictures and words. *Cognition, 81*, B77–B86.

Dell, G.S, & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of The Royal Society B*. 369, 1-9.

Dell, G. S., Burger, L. K., and Svec, W. R. (1997). Language production and serial order: A functional analysis and a model. *Psychological Review,* 104, 123–147.

Dell, G.S., (1986). A spreading-activation theory of retrieval in sentence production, *Psychological Review*, 93, pp. 283–321

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*, 85–103

Eberhard, K. M. (1997). The marked effect of number on subject–verb agreement. *Journal of Memory and Language, 36,* 147–164.

Eberhard, K. M. (1999). The accessibility of conceptual number to the processes of subject–verb agreement in English. *Journal of Memory and Language*, 41, 560–578.

Eberhard, K. M., Cutting, J. C., & Bock, K. (2005). Making sense of syntax: Number agreement in sentence production. *Psychological Review*, 112, 531–559.

Fowler, H. W. (1937). *A dictionary of modern English usage*. Oxford, England: Oxford University Press.

Franck, J., Lassi, G., Frauenfelder, U. H., & Rizzi, L. (2006). Agreement and movement: A syntactic analysis of attraction. *Cognition*, *101*, 173–216.

Franck, J., Vigliocco, G., Antón-Méndez, I., Collina, S., & Frauenfelder, U. H. (2008). The interplay of syntax and form in sentence production: A cross-linguistic study of form effects on agreement. *Language and Cognitive Processes, 23*, 329-374.

Franconeri, S.L., Bemis, D.K, & Alvarez, G.A. (2009). Number estimation relies on a set of segmented objects. *Cognition*, 113, 1-13.

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

Gibson, E., Bergen, L., & Piantadosi, S.T., (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 11:8051-8056

Gillespie, M. & Pearlmutter, N. J. (2011). Hierarchy and scope of planning in subject-verb agreement production. *Cognition, 118,* 377-397.

Gillon, B.S. (1987). The readings of plural noun phrases in English. *Linguistics and Philosophy, 10,* 199-219.

Ginsberg, N., & Goldstein, S. R. (1987). Measurement of visual cluster. The American Journal of Psychology, 100, 193–203.

Greenberg, J. H. (1966). Language universals. The Hague: Mouton.

Halberda, J. Taing, L., & Lidz, J. (2008). The development of 'most' comprehension and its potential dependence on counting ability in preschoolers. *Language Learning and Development*. *4*(2), 99-121.

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*. 30, 643.

Hartsuiker, R.J. (2006). Are speech error patterns affected by a monitoring bias? *Language and Cognitive Processes, 21*, 856-891.

Hartsuiker, R. J., & Barkhuysen, P. N. (2006). Language production and working memory: The case of subject-verb agreement. Language and Cognitive Processes, 21, 181–204.

Hartsuiker, R. J., Corley, M., & Martensen, H. (2005). The lexical bias effect is modulated by context, but the standard monitoring account doesn't fly: Related Reply to Baars, Motley,and MacKay (1975). Journal of Memory and Language, 52, 58–70.

Haskell, T. R., & MacDonald, M. C. (2005). Constituent structure and linear order in language production: Evidence from subject–verb agreement. Journal of Experimental Psychology: Learning, Memory, and Cognition, 31, 891–904.

Haskell, T. R., & MacDonald, M. C. (2003). Conflicting cues and competition in subject–verb agreement. *Journal of Memory and Language, 48*, 760–778.

Howard, D., Nickels, L., Coltheart, M., & Cole-Virtue, J. (2006). Cumulative semantic inhibition in picture naming: Experimental and computational studies. *Cognition, 100,* 464–482.

Humphreys, K. R., & Bock, J. K. (2005). Notional number agreement in English. *Psychonomic Bulletin & Review*, 12, 689–695.

Hurewitz, F., Papafragou, A., Gleitman, L., & Gelman, R. (2006). Asymmetries in the acquisition of numbers and quantifiers. *Language, Learning and Development*, 2, 77–96.

Jackendoff, R. (1991). Parts and Boundaries, *Cognition* 41, 9-45.

Jackendoff, R. (1994). The proper treatment of measuring out, telicity, and perhaps even quantification in English. *Natural Language and Linguistic Theory,* 14, 305-354.

Konopka, A.E. & Bock, K., (2009). Lexical or syntactic control of sentence formulation? Structural generalization from idiom production. *Cognitive Psychology*. 58, 68-101.

Krueger, L. E. (1972). Perceived numerosity. Perception and Psychophysics, 11, 5–9.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. Cognition, 14, 41–104.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22,* 1–75.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition* 106, 1126–1177.

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye move- ment evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the*

National Academy of Sciences, *106*(50), 21086.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 2, 375–419.

Li, P., Ogura, T., Barner, D., Yang, S.J., & Carey, S. (2009) Does the conceptual distinction between singular and plural sets depend on language? *Developmental Psychology*, 45(6), 1644-1653.

Lorimor, H. (2007). Conjunctions and grammatical agreement. (Unpublished doctoral dissertation). University of Illinois, Urbana-Champaign.

MacDonald MC. 2013 How language production shapes language form and comprehension. *Frontiers in Psychology,* 4, 1-16.

McElree, B., (1996). Accessing short-term memory with semantic and phonological information: A time course analysis. *Memory and Cognition*. 24, 173-187.

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, 48, 67–91.

Mircovic, J. & MacDonald, M.C. (2013). When singular and plural are both grammatical: Semantic and morphophonological effects in agreement. *Journal of Memory and Language, 69,* 227-298.

Moreno-Martínez FJ, Montoro PR (2012) An Ecological Alternative to Snodgrass & Vanderwart: 360 High Quality Colour Images with Norms for Seven Psycholinguistic Variables. PLoS ONE 7(5): e37527.

Motley, M.T., Camden,C.T., & Baars,B.J. (1981). Toward verifying the assumptions of laboratory- induced slips of the tongue: The output-error and editing issues. Human Communication Research, 8, 3-15.

Motley, M.T., Camden, C.T., Baars, B.,J. (1982). Covert formulation and editing of anomalies in speech production: Evidence from experimentally elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior, 21*, 578-594.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. URL http://www.usf.edu/FreeAssociation/.

Nicol, J., Forster, K., & Veres, C. (1997). Subject-verb agreement processes in pomprehension. *Journal of Memory and Language*, *36*, 569–587.

Nozari, N. & Dell, G.S., (2009). More on lexical bias: How efficient can a "lexical editor" be? *Journal of Memory and Language*, 60, 291-307.

Pachella, R. G. (1974). The interpretation of reaction time in information processing research. In B. H. Kantowitz (Ed.), Human information processing: Tutorials in performance and cognition. (pp. 41-62). Hillsdale, NJ: Erlbaum.

Pearlmutter, N. J., Garnsey, S. M., & Bock, J. K. (1999). Agreement processes in sentence comprehension. Journal of Memory and Language, 41, 427–456.

Pickering, M.J. & Garrod, S., (2013). Toward an integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, 36:4, 1-18.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rahman, R.A., & Melinger, A. (2007). When bees hamper the production of honey: Lexical interference from associates in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 604-614.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85(2), 59–108.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20,* 873– 922.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, 9(3), 438–481.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review*. 111(1), 159-182.

Ratcliff, R., Love, J., Thompson, C. A., & Opfer, J. E. (2012). Children are not like older adults: A diffusion model analysis of developmental changes in speeded responses. *Child Development*, *83*(1), 367-381.

Ratcliff, R., Thapar, A. & McKoon, G. (2006). Aging, practice, and perceptual tasks: A diffusion model analysis. *Psychology & Aging*, 21(2), 353-71/

Ratcliff, R., Thapar, A., Gomez, P., & McKoon, G. (2004). A diffusion model analysis of the effects of aging in the lexical-decision task. *Psychology & Aging*, 19(2), 278–289.

Schriefers, H., Meyer, A. S., & Levelt, W. J. M. (1990). Exploring the time course of lexical access in language production—picture–word interference studies. *Journal of Memory and Language*, *29*, 86–102.

Schwarzchild, R. (1994). Plurals, presuppositions, and the sources of distributivity. *Natural Language Semantics,* 2, 201-248.

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*, 498–524.

Solomon, E. S., & Pearlmutter, N. J. (2004). Semantic integration and syntactic planning in language production. *Cognitive Psychology*, 49, 1–46.

Sophian, C., & Chu, Y. (2008). How do people apprehend large numerosities? Cognition, 107, 460–478.

Spaniol, J., Madden, D. J., & Voss, A. (2006). A diffusion model analysis of adult age differences in episodic and semantic long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 101–117.

Staub, A. (2010). Response time distributional evidence for distinct varieties of number attraction. *Cognition*. 114, 447-454.

Staub, A. (2008). The computation of subject-verb number agreement: Response time studies. Unpublished doctoral dissertation.

Staub, A. (2009). On the interpretation of the number attraction effect: Response time evidence. *Journal of Memory and Language*, 60, 308–327.

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

Thornton, R., & MacDonald, M. C. (2003). Plausibility and grammatical agreement. *Journal of Memory and Language*, 48, 740–759.

Trick, L. M., & Enns, J. T. (1997). Clusters precede shapes in perceptual organization. Psychological Science, 8(2), 124–129.

Van Dyke, J. A., & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, *49*, 285-316.

Vandekerckhove, J., Verheyen, S., & Tuerlinckx, F. (2010). A crossed random effects diffusion model for speeded semantic categorization decisions. *Acta Psychologica*. 133, 269-282.

Veenstra, A. (2014). Semantic and syntactic constraints on the production of subject-verb agreement. (Unpublished doctoral dissertation). Radboud University Nijmegen.

Veenstra, A., Acheson, D. J., Bock, K., & Meyer, A. S. (2014). Effects of semantic integration on subject–verb agreement: Evidence from Dutch. *Language, Cognition and Neuroscience, 29*, 355-380.

Vigliocco, G., & Nicol, J. (1998). Separating hierarchical relations and word order in language production: is proximity concord syntactic or linear? *Cognition*, *68*, 13–29.

Vigliocco, G., Butterworth, B., & Garrett, M. (1996). Subject-verb agreement in Spanish and English: Differences in the role of conceptual constraints. *Cognition*, *61*, 261-298.

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*, 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*, 455-478.

Vos, P. G., van Oeffelen, M. P., Tibosch, H. J., & Allik, J. (1988). Interactions between area and numerosity. Psychological Research, 50(3), 148–154.

Voss, A., & Voss, J. (2008). A fast numerical algorithm for the estimation of diffusion model parameters. *Journal of Mathematical Psychology*, *52*, 1-9.

Voss, A., & Voss, J., (2007). Fast-dm: A free program for efficient diffusion model

analysis. *Behavior Research Methods. 39*, 767-775.

Voss, A., Rothermund, K., & Brandtstädter, J. (2008). Interpreting ambiguous stimuli: Separating perceptual and judgmental biases. *Journal of Experimental Social Psychology*. 44, 1048-1056.

Wagenmakers, E.J., van der Maas, H.L.J., & Grasman, R.P.P.P., 2007. An ez-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review, 4*, 3–22.

Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language, 61*, 206–237.

Wertheimer M. (1923) Laws of Organization in Perceptual Forms. Psychologische Forschung 4, 301-350. Reprinted in W. D. Ellis (Ed. & Transl. 1938) A Source Book of Gestalt Psychology. New York: Harcourt Brace, London: Routledge & Kegan Paul. 71-88.

Wertheimer M. (1923) Principles of Perceptual Organization. In D. S. Beardslee & M. Wertheimer (Eds.) Readings in Perception. Princeton NJ: Van Nostrand-Reinhold. 115-137.

Wheeldon, L. R., & Monsell, S. (1994). Inhibition of spoken word production by priming a semantic competitor. *Journal of Memory and Language, 33*, 332-356.

APPENDIX A

Experimental Stimuli

Experiments 1 & 4-6 Stimuli

Preamble (integrated / unintegrated)	*Head match predicate*	*Local match predicate (integrated / unintegrated)*
The book with the torn page(s) / red pen(s)	fiction	yellowed / smudged
The shirt with the crazy pattern(s) / dirty towel(s)	buttoned	striped / wet
The ring with the fake diamond(s) / gold bracelet(s)	silver	sparkly / chunky
The apple with the brown spot(s) / fresh peach(es)	green	rotten / soft
The tie with the hideous stripe(s) / cotton blazer(s)	silk	dizzying / pressed
The watch with the missing hand(s) / black wallet(s)	lost	ticking / leather
The jacket with the faulty zipper(s) / wet umbrella(s)	warm	broken / soaked
The razor with the rusty blade(s) / empty can(s)	rinsed	clean / aluminum
The key with the jagged edge(s) / shiny coin(s)	brass	sharp / silver
The bed with the creaky spring(s) / tall bookcase(s)	soft	cheap / wooden
The phone with the missing button(s) / broken toaster(s)	ringing	plastic / shiny
The pillow with the nasty stain(s) / flannel sheet(s)	soft	dirty / washed
The lamp with the florescent bulb(s) / antique portrait(s)	lit	bright / painted
The magazine with the colorful ad(s) / telephone book(s)	popular	perfumed / heavy
The sweater with the tiny hole(s) / linen suit(s)	woolen	mended / formal
The receipt with the blurry price(s) / sealed package(s)	voided	illegible / mailed
The tree with the dead branch(es) / small shrub(s)	tall	pruned / thorny

The pizza with the yummy topping(s) / tasty beverage(s)	delivered	pepperoni / cold
The milk with the extra vitamin(s) / blueberry muffin(s)	pasteurized	enriched / warm
The guitar with the loose string(s) / loud drum(s)	played	plucked / hit
The blanket with the soft fringe(s) / clean skirt(s)	warm	fraying / short
The glass with the lengthy crack(s) / crystal bowl(s)	fragile	split / patterned
The bike with the bent spoke(s) / surfboard(s)	ridden	chrome / wet
The chair with the wobbly leg(s) / old table(s)	comfortable	broken / antique

Experiment 2 Stimuli

Preamble (related / unrelated)

Concrete

The ambulance and the hospital(s)/ light(s)

The tongue and the mouth(s) / teeth

His cabbage and vegetable(s) / salad(s)

The tail and the dog(s) / animal(s)

The cap and the head(s) / gun(s)

Their missile and the bomb(s) / plane(s)

The dish and the plate(s) / cat(s)

His donkey and the horse(s) / monkey(s)

The hill and the mountain(s) / valley(s)

Her leg and her arm(s) / muscle(s)

Her flask and his bottle(s) / jar(s)

His sword and her knife(knives) / weapon(s)

Her pouch and bag(s) / pipe(s)

Their raft and boat(s) / beach(es)

The rail and train(s) / fence(s)

The spine and the bone(s) / neck(s)

Abstract

The cause and the effect(s) / force(s)

His chance and her risk(s) / option(s)

The chaos and the confusion(s) / headache(s)

His compulsion and his obsession(s) / tendency(tendencies)

Her concept and his idea(s) / belief(s)

The condition and the situation(s) / issue(s)

Her curiosity and her wondering(s) / interest(s)

Their destiny and fate(s) / outcome(s)

Their domain and range(s) / rule(s)

The fantasy and the dream(s) / wish(es)

The honesty and the truth(s) / loyalty(loyalties)

The hypothesis and the theory(theories) / thought(s)

The mischief and the trouble(s) / mistake(s)

His mood and his emotion(s) / attitude(s)

The norm and the average(s) / standard(s)

Her weakness and her strength(s) / fault(s)

Experiment 3 stimuli

Root preamble	*Predicate*
candle with cotton wicks	flickering
dress with colored parts	elegant
crib for sleeping babies	secure
blender for nutritious smoothies	empty
cat with sensitive whiskers	mean
mirror with reflective surfaces	smudged
can from pantry cabinets	sealed
cd for road trips	excellent
muffin from hip bakeries	tasty
spoon with subtle carvings	stained
rubberband from supply closets	stretchy
table with rickety legs	cluttered
hat from spring lines	dorky
marker for young artists	washable
umbrella for rainy days	collapsible
shovel with tapered edges	flat
plate with subdued decorations	practical
scale from apartment bathrooms	accurate
alligator with humongous claws	hungry
basket for buttery pastries	woven
snake for reptile lovers	slithery
monkey from exotic jungles	active
pepper from green plants	spicy
anchor with giant ropes	rusty
bracelet with tasteful patterns	shiny
bee with clear wings	buzzing
match from nearby restaurants	flammable
skirt from local boutiques	short

ladder for brave firefighters	tall
skeleton for anatomy courses	spooky
rat from Psychology labs	friendly
trophy for winning runners	polished
pear from produce stands	grainy
egg for hearty breakfasts	fresh
teapot for classy parties	dainty
tomato from organic farms	squishy
truck with yellow headlights	muddy
tent for adventurous campers	musty
brush for indoor pets	broken
helmet for cautious cyclists	safe
bicycle for fit commuters	efficient
acorn from rural forests	hard
caterpillar with fringy feelers	weird
lollipop from elderly neighbors	sweet
pineapple from tropical islands	tangy
sandwich from Kosher delis	yummy
bow for special presents	decorative
cracker for afternoon snacks	crunchy
cane for injured soldiers	sturdy
stroller with cushy seats	speedy
spatula for beaten egg-whites	wide
toothbrush with pokey bristles	clean
weight with flattened sides	heavy
unicorn from made-up stories	fictional
yoyo for interested boys	fun
watermelon from farmers markets	juicy
boot with acrylic laces	stylish
arrow from outdoor stores	feathered
scorpion with beady eyes	stinging

flashlight with incandescent bulbs	blinding
bell for Christmas carols	ringing
jacket from northern climates	cozy
necklace for dressed-up ladies	fashionable
skull from respected museums	terrifying
skateboard with impressive designs	cool
screw with stripped threads	small
sweater from doting relatives	itchy
suitcase with extra pockets	locked
purse from exclusive designers	trendy
plug for electronic devices	bulky
squirrel with twitchy paws	fluffy
camel from dry hills	grumpy

APPENDIX B

Instructions for norming tasks, and Experiments 1 and 2.

Sensibility Norming: Experiments 1 and 2

Circle the number on the right that corresponds to how likely the statement on the left is.

Example: How likely is it that a planet is big? Not likely-- 1 2 3 4 5 6 **7** --Very likely

How likely is it that an ice cube is hot? Not likely-- **1** 2 3 4 5 6 7 --Very likely

Plausibility Norming: Experiment 3

You're going to be reading a number of sentences. Please judge the acceptability of these sentences on a scale from 1 to 7 (not good to very good) by picking the appropriate number.

For example:

The casserole in the oven was ready. Not good-- 1 2 3 4 5 6 **7** --Very good

This one is very acceptable, so it gets a 7.

The casserole in the thimble was ready. Not good-- **1** 2 3 4 5 6 7 --Very good

This one is not very acceptable, so it gets a 1.

Integration Norming: Experiments 1, 2, and 3

In this survey, we'd like you to read some sentences and decide how closely linked two underlined words are within each sentence. Sometimes the underlined words will be highly linked, whereas in other cases they will be more independent from each other. Your task is to circle the number that you think corresponds to how closely linked the

underlined words are within each particular sentence. One strategy that you can use when rating these items is to try to form a mental picture of each sentence To show you what we mean by "closely linked," here are two example sentences, with rating scales from 1-7 on the right:

(1) the ketchup or the mustard might be yummy Not linked-- 1 2 3 4 5 6 7 --Very linked

(2) the bracelet made of silver is nice Not linked-- 1 2 3 4 5 6 7 --Very linked

It is important that you pay close attention to the relationship between the underlined words in each individual sentence because, although words like *ketchup* and *mustard* may be related in general meaning (i.e., they are both condiments), in the first example these words are not closely connected. In (1), the only information that you have is that there are two things -- ketchup and mustard -- but you do not know anything about how these objects are related to each other. So, in the first example you would most likely circle a relatively low number like 1 or 2. In the second example, unlike the first, the underlined words are closely linked because the object bracelet is actually made from silver. So, for (2), you would probably circle a rather high number on the like 6 or 7.

Remember that we're just interested in your opinions here. Please, do not worry about right or wrong answers. Just make sure to take the time to read each sentence carefully before circling a response.

Finally, don't worry if the numbering of the items isn't in order across the pages; just go through the pages in the order you've got them. Different people have different items and different orderings in their surveys.

If you have any questions, feel free to ask the experimenter.

Thanks very much for your participation.

Notional number norming: Experiment 3

In this survey, we'd like you to read some phrases and decide whether they describe one thing or more than one thing. One strategy that you can use when rating these items is to try to form a mental picture of each phrase. To show you what we mean, here are two example phrases:

		One thing	More than one thing
(1)	the ketchup or the mustard		x
(2)	the bracelet made of silver	x	

Remember that we're just interested in your opinions here. Please, do not worry about right or wrong answers. Just make sure to take the time to read each phrase carefully before picking a response.

Thanks very much for your participation.

Experiment 1 Task Instructions

In this experiment, you will be hearing the beginnings of sentences and completing them with adjectives. The way a trial works is like this: First you will see X on the screen for about a second. This is a warning signal, telling you an adjective is about to appear. Then you will see an adjective, which you should read and remember. Then you will hear the beginning of a sentence. Then you will see "!" which is your cue to finish the sentence with the adjective you saw. For example, if you saw "Fast" and

heard "The fire engine" and then saw "!",You could say "was fast." Now let's try an example. Press the space bar to begin.

X/ Hungry / The cookie monster / !

Sometimes something a little different will happen. Instead of an exclamation point, the word "Repeat" will appear. When this happens, you should first REPEAT the phrase, and THEN complete it. For example, if you had just seen "Fast" and heard "The fire engine", and then you saw the word "Repeat", you would repeat the phrase aloud along with your completion. Like, "The fire engine was fast." Your job here is to repeat the phrase, exactly as you heard it, and make it into a complete sentence using the adjective you were given. Press the space bar for an example.

X/ Fun / The rollercoasters / Repeat

Let's review: What do you do when you see "!"? What do you do when you see "Repeat"? Any questions? When you are ready, press space bar to begin.

Experiment 2 Task Instructions

In this experiment, you will be seeing the beginnings of sentences and then completing them. For the completions, you can choose any one of four adjectives: good, bad, ready, true. The way a trial works is like this: First you will see X on the screen for about a second. This is a warning signal, telling you a phrase is about to appear. Then you will see the phrase, which you should treat as the start of a sentence. It will be followed immediately by an exclamation point. This is your signal to speak. For example, if you saw "The fire engine" and then saw "!", You could say "was ready." Now let's try an example. Press the space bar to begin.

X/ Cookie monster / !

Sometimes something a little different will happen. Instead of an exclamation point, the word "Repeat" will appear. When this happens, you should first REPEAT the phrase, and THEN complete it. For example, if you had just seen "The fire engine", and then you saw the word "Repeat", you would repeat the phrase aloud along with your completion. Like, "The fire engine was good." Your job here is to repeat the phrase, exactly as you saw it, and make it into a complete sentence using one of the four adjectives (good, bad, ready, true). Press the space bar for an example.

X/ The rollercoasters / Repeat.

Let's review: What are the 4 adjectives you can use? What do you do when you see "!"? What do you do when you see "Repeat"? Any questions? When you are ready, press space bar to begin.

Experiment 3 Task Instructions

In this experiment, you will be hearing the beginnings of sentences and completing them with adjectives. The way a trial works is like this: First you will see "+" on the screen for about a second. This is a warning signal, telling you an adjective is about to appear. Then you will see an adjective, which you should read and remember. Then, you will hear a sentence fragment and see a picture to help you visualize the fragment.

On some trials (about 15 percent), you should repeat back the beginning of the sentence and complete it with the adjective. This will be cued with the word "Repeat". For example, if you had just seen "Fast" and heard "The fire engine", and then you saw "Repeat", you would repeat the phrase aloud along with your completion. Like, "The fire

engine was fast." Please be as quick and as accurate as you can. Here is an example:

+ / TOASTED/ ‘the bagels’ / REPEAT

For the remaining sentences, you will be asked to simply add a completion to the sentence using the adjective. This will be cued with "!" instead of "Repeat." For example, if you had just seen "Fast" and heard "The fire engine", and then you saw "!", you would simply give an ending. Like, "was fast." Please be as quick and as accurate as you can. Here is an example:

+ / STICKY / ‘the bandaid’ / !

Any questions? We will start with some practice trials. Press any key to begin.

Experiment 4 Task Instructions

In this experiment, you will be completing sentences. First, you will see an adjective in the center of the screen, in red font. Then, you will see the beginning of a sentence presented one word at a time.

On some trials (about 15 percent), you should repeat back the beginning of the sentence and complete it with the adjective. Here is an example, played at a slower speed:
FAST/ The / fire / engine / REPEAT

For the remaining sentences, you will be asked to decide whether you would continue the sentence with WAS or WERE. You will indicate your decision by pressing F or J on the keyboard. Press F for WAS, and press J for WERE. Respond as quickly as you can without making mistakes. You will be given feedback if you make a mistake or if you respond too slowly. Here is an example, played at a slower speed:
FAST/ The / fire / engine / WAS WERE

Any questions? We will start with some practice trials. Press any key to begin.

Experiment 5 Task Instructions

In this experiment, you will be completing sentences. First, you will see an adjective in the center of the screen, in red font. Then, you will see the beginning of a sentence presented one word at a time.

On some trials (about 15 percent), you are asked to type out the beginning of the sentence and complete it with the adjective. Here is an example, played at a slower speed:
FAST/ The / fire / engine / Type response:

For the remaining sentences, you will be asked to decide whether you would continue the sentence with WAS or WERE. You will indicate your decision by pressing F or J on the keyboard. Press F for WAS, and press J for WERE. Respond as quickly as you can without making mistakes. You will be given feedback if you make a mistake. Here is an example, played at a slower speed:
FAST/ The / fire / engine / WAS WERE

Any questions? We will start with some practice trials. Press any key to begin.

Experiment 6 Task Instructions

In this experiment, you will be completing sentences. First, you will see an adjective in the center of the screen, in red font. Then, you will see the beginning of a

sentence presented one word at a time.

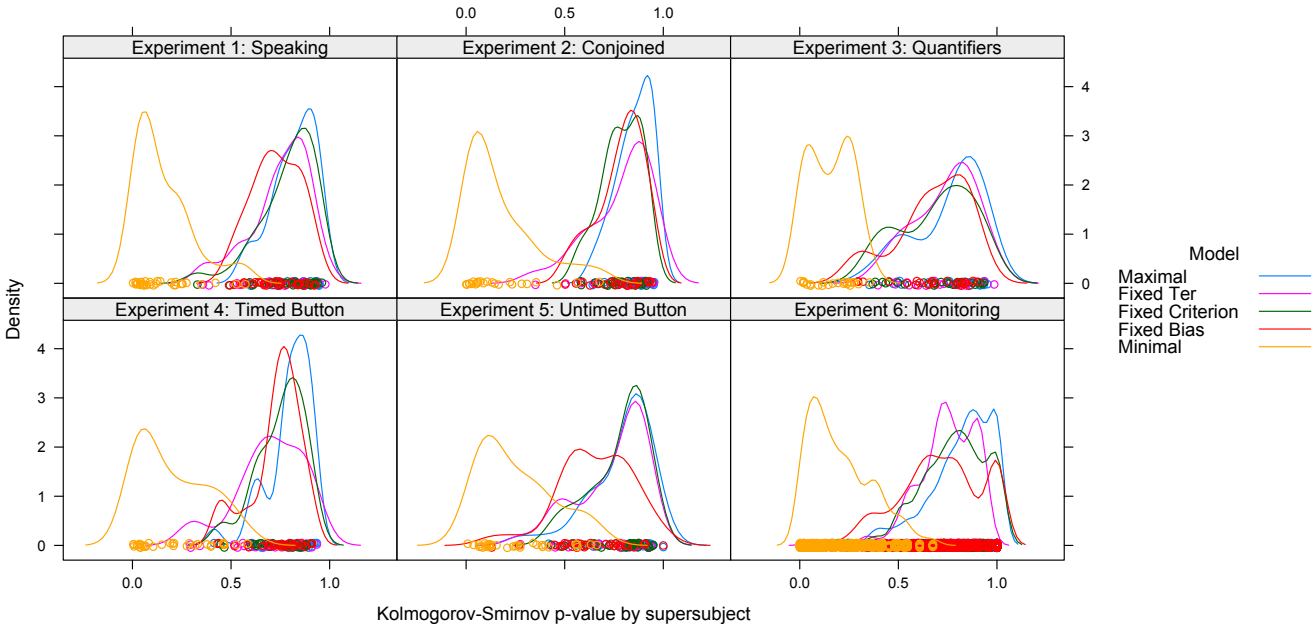
On some trials (about 15 percent), you are asked to type out the beginning of the sentence and complete it with the adjective. Here is an example, played at a slower speed:
FAST/ The / fire / engine / Type response:

For the remaining sentences, you will be asked to decide whether you would continue the sentence with WAS or WERE. You will indicate your decision by pressing F or J on the keyboard. Press F for WAS, and press J for WERE. Respond as quickly as you can without making mistakes. You will be given feedback if you make a mistake or if you respond too slowly. Here is an example, played at a slower speed:
FAST/ The / fire / engine / WAS WERE

We will start with some practice trials. Press any key to begin.

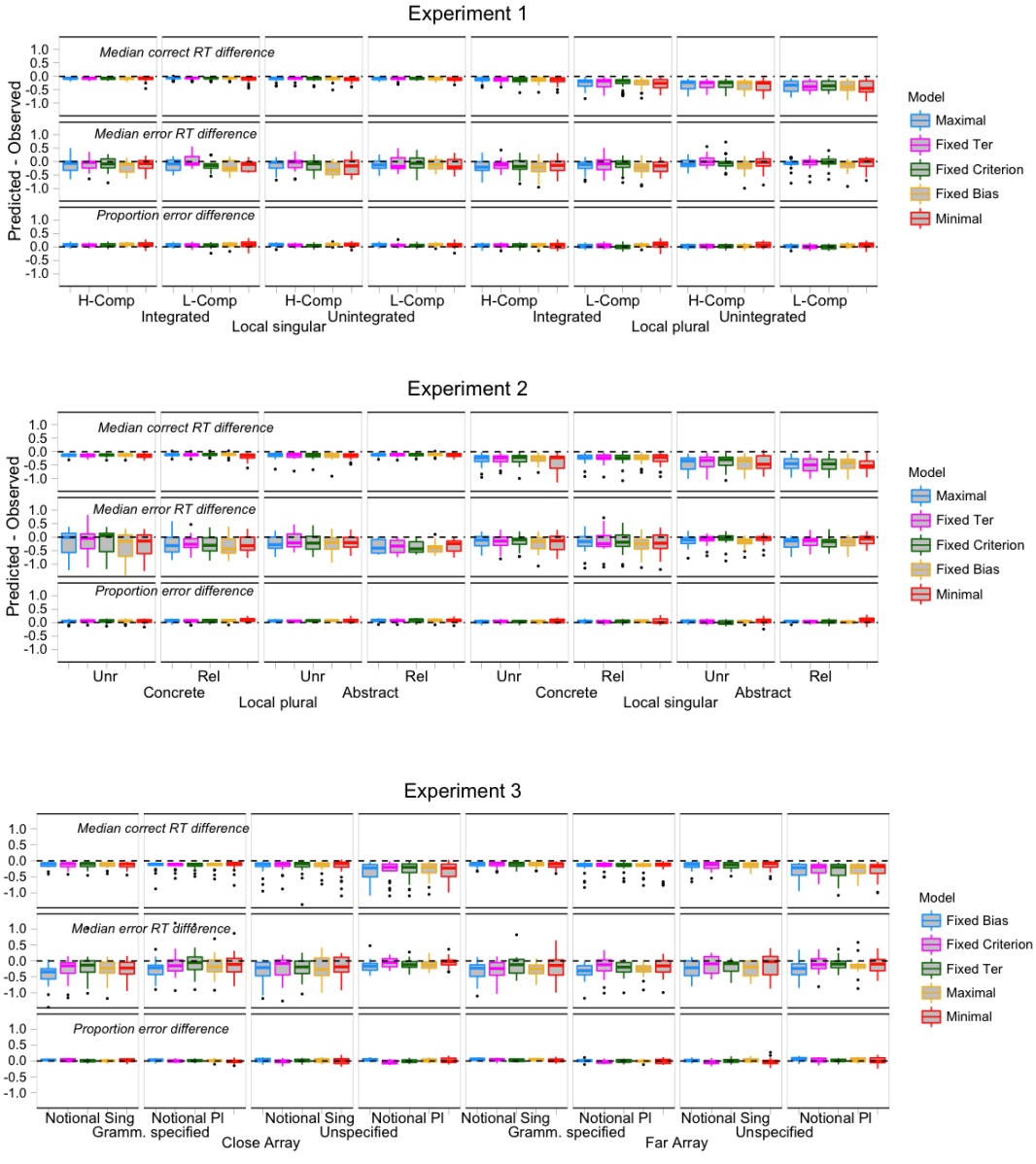
APPENDIX C

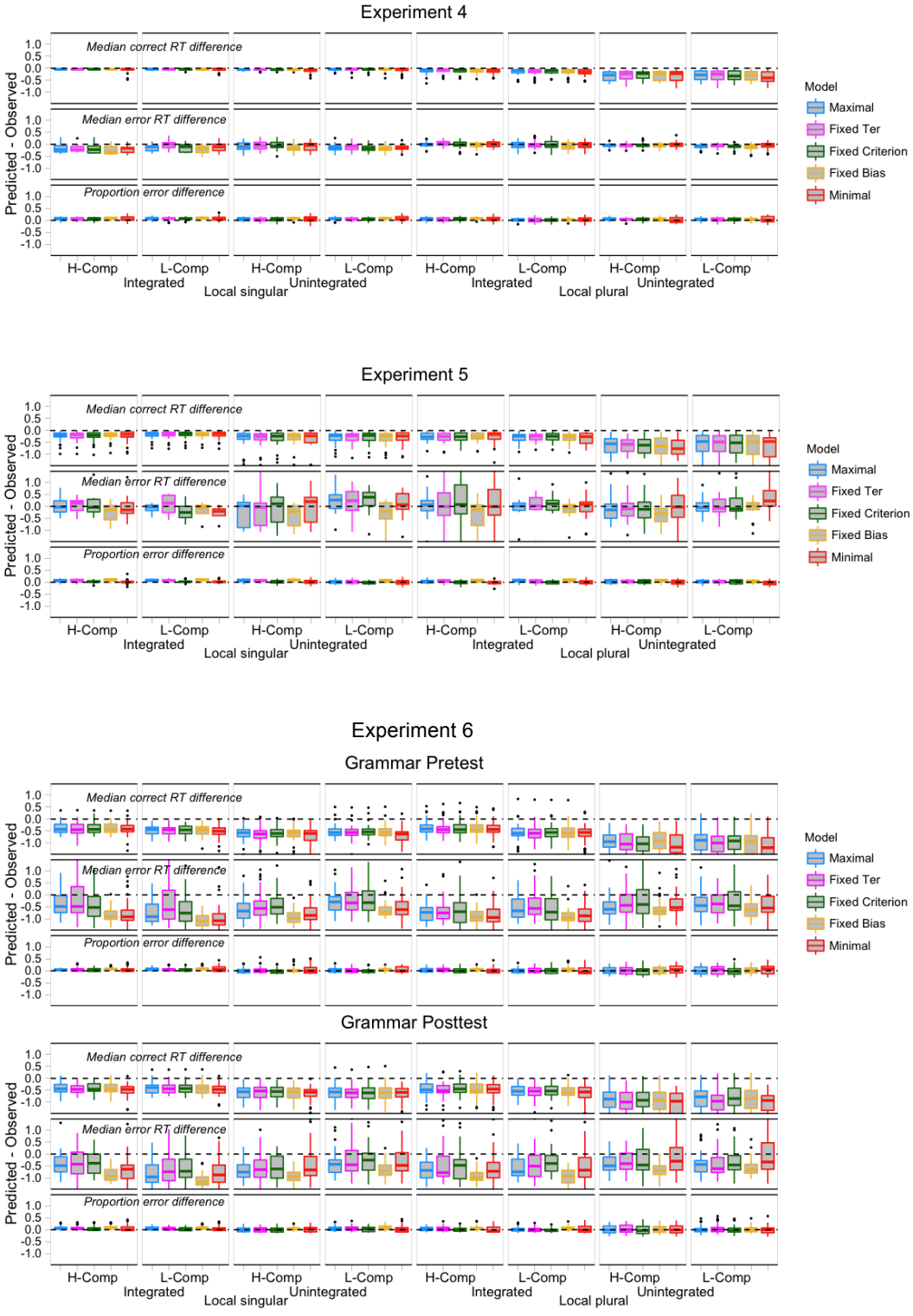
Distributions of Kolmogorov-Smirnov *p*-values across diffusion models by supersubject, by experiment.



APPENDIX D

Comparison between empirical data and predicted model fits for three points on each curve. All measures subtract predicted data from the observed data. Colors represent different models; groups of colors represent different experimental conditions.




Experiment 4
Median correct RT difference
Median error RT difference
Proportion error difference
Predicted - Observed
1.0
0.5
0.0
-0.5
-1.0
H-Comp
L-Comp
Integrated
Unintegrated
Local singular
Local plural
Model
Maximal
Fixed Ter
Fixed Criterion
Fixed Bias
Minimal
Experiment 5
Experiment 6
Grammar Pretest
Grammar Posttest

APPENDIX E

Grammar test from Experiment 6.

Part (1 or 2) of the experiment is a test about your knowledge of English grammar. Select the correct sentence in each of the pairs below by clicking the button next to the sentence.

- ○ The squirrel are often munching on walnuts
- ○ The squirrel is often munching on walnuts

- ○ The grapefruit comes with the omelet for free
- ○ The grapefruit come with the omelet for free

- ○ The dishes and the plates makes nice place settings
- ○ The dishes and the plates make nice place settings

- ○ The memo informs new employees about company policies
- ○ The memo inform new employees about company policies

- ○ The families with the expensive cameras admire beautiful scenery
- ○ The families with the expensive cameras admires beautiful scenery

- ○ Honesty and facts damages some people's self-esteem
- ○ Honesty and facts damage some people's self-esteem

- ○ The situations and the condition have received worldwide attention
- ○ The situations and the condition has received worldwide attention

- ○ A bunch of small toys was left on the floor
- ○ A bunch of small toys were left on the floor

- ○ The bridge to the islands move up to allow boats through
- ○ The bridge to the islands moves up to allow boats through

- ○ The team in the advertisements have gotten lots of publicity
- ○ The team in the advertisements has gotten lots of publicity

- ○ The books next to the computer teach programming basics
- ○ The books next to the computer teaches programming basics

- ○ The crowd at Olympic events goes wild when a point is scored
- ○ The crowd at Olympic events go wild when a point is scored

- A number of similar problems appear in the school every year
- A number of similar problems appears in the school every year

- The keys to the cabinets was rusty from many years of disuse
- The keys to the cabinets were rusty from many years of disuse

- The army with the easy-going commanders enjoy frequent breaks
- The army with the easy-going commanders enjoys frequent breaks

- The check from the stockbrokers arrive once a month
- The check from the stockbrokers arrives once a month

- The ambulance and the hospital treats wounded patients
- The ambulance and the hospital treat wounded patients

- The hypothesis and the theory give physics students trouble
- The hypothesis and the theory gives physics students trouble

- The door to the offices gets unlocked by the cleaning service
- The door to the offices get unlocked by the cleaning service

- The mountain and the hills are near the scenic lake
- The mountain and the hills is near the scenic lake

- The choir for the church services does not practice often enough
- The choir for the church services do not practice often enough

- The vase for the flowers does not match the tablecloth
- The vase for the flowers do not match the tablecloth

- The bonus bring a smile to many workers' faces
- The bonus brings a smile to many workers' faces

- The classes in the writing competition takes winning seriously
- The classes in the writing competition take winning seriously