Journal of Memory and Language 82 (2015) 1–17

Contents lists available at ScienceDirect

Journal of Memory and Language

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

The influence of cloze probability and item constraint on cloze task response time

Adrian Staub [a,*], Margaret Grant [b], Lori Astheimer [a], Andrew Cohen [a]

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

ARTICLE INFO

Article history:
Received 28 April 2014
revision received 19 February 2015
Available online 17 March 2015

Keywords:
Prediction
Language processing
Cloze task
Response time

ABSTRACT

In research on the role of lexical predictability in language comprehension, predictability is generally defined as the probability that a word is provided as a sentence continuation in the *cloze task* (Taylor, 1953), in which subjects are asked to guess the next word of a sentence. The present experiments investigate the process by which subjects generate a cloze response, by measuring the latency to initiate a response in a version of the task in which subjects produce a spoken continuation to a visually presented sentence fragment. Higher probability responses were produced faster than lower probability responses. The latency to produce a response was also influenced by item constraint: A response at a given level of probability was issued faster when the context was more constraining, i.e., a single response was elicited with high probability. We show that these patterns are naturally produced by an activation-based race model in which potential responses independently race towards a response threshold. Implications for the interpretation of cloze probability as a measure of lexical predictability are discussed.

Introduction

Lexical predictability plays an important role in incremental language comprehension. In reading, the eyes spend less time on a word when it is predictable in its sentence context than when it is a less predictable but plausible sentence continuation (e.g., Ehrlich & Rayner, 1981; Smith & Levy, 2013; Staub, 2011). In event-related potential (ERP) research, the amplitude of the N400 component, which is elicited by each word of a sentence in both written and spoken language comprehension, is modulated by a word's predictability (e.g., Federmeier & Kutas, 1999; Kutas & Hillyard, 1984), with N400 amplitude decreasing as a word becomes more predictable. Thus, we have evidence that lexical predictability influences overt processing behavior, and we also have direct evidence of the influence of predictability on the neural processes that presumably underlie this behavior. Recent reviews of predictability effects, and discussion of their interpretation, can be found in Federmeier (2007), Pickering and Garrod (2007), and Van Petten and Luka (2012). Lexical predictability also plays a central role in recent computational models of sentence processing that have emphasized the conditional probability of a word as a determinant of processing difficulty (Hale, 2001; Levy, 2008).

In such research, a word's predictability in a given context is almost always operationalized in terms of *cloze probability* (Taylor, 1953). This measure is simply the proportion of participants who provide the word in question as the next word of the sentence, given the preceding words. Due to the extensive interest in predictability effects, and also due to the fact that researchers often control for predictability when investigating effects of other

* Corresponding author at: Department of Psychological and Brain Sciences, University of Massachusetts, 430 Tobin Hall, Amherst, MA 01003, United States. Fax: +1 (413) 545 0996.
E-mail address: astaub@psych.umass.edu (A. Staub).

http://dx.doi.org/10.1016/j.jml.2015.02.004
0749-596X/

lexical or sentence-level variables, the cloze task has become one of the most widely used laboratory tasks in psycholinguistics. In standard practice, a fragment of a sentence is presented in written form to a group of subjects in a norming session prior to the comprehension experiment that is the researchers' main focus. Usually, two distinct groups of subjects complete the cloze task and the comprehension experiment. The subjects in the cloze task are asked to write the word that seems most likely as the next word of the sentence, though there is great variability in instructions, with some researchers asking for the most natural continuation, the most plausible continuation, the 'best' continuation, or the first word that comes to mind. In the associated comprehension experiment, researchers usually assess the effect of predictability in one of two ways: by comparing the processing of a given word in contexts in which this word has high or low cloze probability (as in 1a–b; target words italicized), or by holding the context constant, and comparing two different words that have high and low cloze probability in this context (2a–b). The first of these designs maintains control over lexical variables, at the cost of contextual variability, while the second maintains control over the context, at the cost of lexical variability. On occasion, context and target are fully crossed (3a–d).

(1) a. The athlete pulled a *muscle* in his leg during the competition.
b. Peter says that a *muscle* in his leg was bothering him during soccer practice. (Sheridan & Reingold, 2012)

(2) a. He scraped the cold food from his *plate* before washing it.
b. He scraped the cold food from his *spoon* before washing it. (Rayner & Well, 1996)

(3) a. Before warming the milk, the babysitter took the infant's *bottle* out of the travel bag.
b. To prevent a mess, the caregiver checked the baby's *bottle* before leaving.
c. Before warming the milk, the babysitter took the infant's *diaper* out of the travel bag.
d. To prevent a mess, the caregiver checked the baby's *diaper* before leaving. (Rayner, Ashby, Pollatsek, & Reichle, 2004)

The cloze task is itself a language production task. It is an off-line task, in the sense that it does not require the subject to make a rapid, or even timed, response. The use of cloze probability as a predictor variable in language comprehension research rests, then, on the assumption that the off-line production probability of a word obtained from one group of subjects predicts some aspect of on-line comprehension of that word for a different group of subjects. This broad assumption appears justified, as cloze probability is a very useful predictor variable indeed. Smith and Levy (2011) found that a word's cloze probability is a better predictor of self-paced reading time than is the word's empirically determined conditional probability, as estimated from either a book corpus or a web-based corpus. Indeed, corpus-based conditional probability explained no additional reading time variance in a model that included cloze probability. Similarly, Frisson, Rayner, and Pickering (2005) found that even small differences in cloze probability at the low end of the scale influence eye fixation durations in reading, and that corpus-based transitional probability has no additional effect when cloze probability is carefully controlled (cf. McDonald & Shillcock, 2003).

The present work is motivated by the question of what the cloze probability of a word actually represents, in cognitive terms. The answer to this question requires a theory of how subjects perform the cloze task. This is the situation for most dependent variables in laboratory tasks; for example, the probability that a particular stimulus will be identified as 'old' in a recognition memory task receives a psychological interpretation only in the context of a theory of how subjects perform old/new judgments. At the broadest level, cloze probability has been assumed to represent native speakers' estimates of a word's probability given the preceding sentential context, such that, if in (2) above *plate* has cloze probability of .8 and *spoon* has cloze probability of .2, it can be reasonably concluded that native speakers estimate the probability of *plate* in this context to be .8 and *spoon* to be .2. But what kind of implicit theory of how the task is performed links cloze probabilities to speakers' estimates of conditional probability?

There are two possible theories that researchers may have in mind. The literal meaning of cloze values of .8 for *plate* and .2 for *spoon*, in (2) above, is that *plate* was provided as a sentence continuation by 80% of the participants in a cloze norming session, while *spoon* was provided by 20% of participants. Thus, one apparently plausible interpretation is that there are two groups of subjects who are genuinely different, in terms of their linguistic experience and real-world knowledge, so that for 80% of subjects *plate* is the most expected word, and for 20% *spoon* is the most expected word, with each subject producing his or her most expected word. On this view, the cloze probability of a word represents its predictability for the community of speakers in the aggregate, but not for individual speakers. For each individual speaker there is a single most expected word. (See Van Petten & Luka, 2012, for further discussion of this idea.)

If this were the model linking cloze probabilities to predictability, a comprehension experiment with sentence (2) above would be expected to reveal two distinct groups of subjects, with approximately 80% of participants responding to *plate* as if it is the expected word, showing relatively fast reading of this word and a reduced N400 amplitude, and the other 20% of participants showing the opposite pattern, with a processing advantage for *spoon*. No such distinction between subgroups of comprehension subjects has ever been reported or, to our knowledge, seriously considered. We infer that this is not, in fact, the link between cloze probabilities and predictability that researchers have in mind.

The alternate theory assumes that though each subject in the cloze task provides only a single response (leaving aside the relatively few cloze studies in which subjects are asked to provide multiple responses, e.g., Roland,

Yun, Koenig, & Mauner, 2012), cloze probabilities of .8 and .2 for *plate* and *spoon*, respectively, mean that *plate* is quite generally more expected than *spoon*, for all speakers, or at least the vast majority. In other words, the 80/20 split in cloze probabilities does not imply an 80/20 split between subjects, but rather something like an 80/20 split within the mind of each subject. On this theory, *plate* is in some sense the more predictable response even for those subjects who do not actually produce it in the cloze task; variability in responses is due to a probabilistic aspect of the response process itself, rather than differences in subjects' actual linguistic expectations.

How might subjects' discrete cloze responses arise from a subjective probability distribution that is more-or-less similar across subjects? Smith and Levy (2011) provide a suggestion: Each subject performs the cloze task simply by sampling once from this probability distribution. Thus, even if subjects have completely identical probabilistic models of the language – indeed, even if the same subject were to provide multiple responses (at different times) to the same fragment – a range of cloze responses would be elicited, in proportion to their probabilities in the cognitive model that subjects share.

We suspect that Smith and Levy's (2011) conception of the link between cloze probabilities and predictability is the one that most researchers have in mind, even if this conception is generally inexplicit. However, Smith and Levy point out that there is little empirical support, at present, for the notion that subjects in the cloze task are successfully sampling from a subjective probability distribution:

> [P]articipants in a cloze task have some knowledge of their language, which they presumably draw on when producing continuations. But isn't clear how they use this knowledge. If they generated their cloze responses by sampling from their subjective probability distribution ('probability matching'), then cloze probabilities would be identical to subjective probabilities. But cloze norming is an offline, untimed, and rather unnatural task, which leaves ample room for conscious reflection and other strategic effects to distort this process—if participants are even probability matching in the first place (p. 1637–8).

Importantly, another result reported by Smith and Levy (2011) is that the correlation between a word's cloze probability in a given context and its conditional probability in actual corpora is modest, with certain biases appearing in cloze responses. For example, a cloze response is especially likely to be a semantic associate of words in the preceding fragment, and to be a familiar word. Considering these biases in the cloze task together with the finding that cloze probability predicts reading time better than does corpus probability, Smith and Levy suggest that these biases may reflect "genuine errors in native speakers' probabilistic models of their language" (p. 1637), which operate both in the cloze task and in the course of on-line comprehension. Thus, Smith and Levy's response to their observation of systematic discrepancies between cloze probabilities and corpus-based conditional probabilities is to preserve the assumption that subjects in the cloze task are predicting the next word of the sentence based on sampling from a subjective probability distribution, but to question the accuracy of this subjective probability distribution. Speakers' subjective probability distribution for the word following a given context over-weights familiar words and words that are semantic associates of words in the context, and as a result they produce these words in the cloze task, and they expect to encounter these words when reading or listening.

Though the cloze task is traditionally an untimed task, the process of issuing a cloze response does unfold in time, as in any timed production task such as picture naming (e.g., Schriefers, Meyer, & Levelt, 1990), sentence production (e.g., Ferreira, 1996), or selecting an agreeing verb form (e.g., Staub, 2009). In research using these timed tasks, the pattern of response times (RTs) is used in concert with response proportions to constrain theories of the process by which a response is generated. In cognitive psychology more broadly, it is regarded as critical for cognitive models to account for both response proportion data and RT data, and for the relations between them (e.g., Luce, 1986). In the present work, we collect RTs together with cloze responses in order to inform and constrain a model of the process by which subjects produce cloze responses.

Smith and Levy's (2011) conception of a cloze response as arising from a simple process of sampling once from a subjective probability distribution, if this is regarded as a genuine process model, makes a prediction regarding RT that seems *prima facie* unlikely.[1] In the absence of additional mechanisms, this view of the process predicts that there should be no relationship between the latency to produce a response and the identity of the response itself or the context in which that response was issued. That is, higher probability responses should be neither faster nor slower than lower probability responses, and responses to a high constraint context (e.g., one that elicits a single response with probability .9) should be neither faster nor slower than responses to a low constraint context (e.g., one that elicits no single response with a probability above .1). This is because the duration of a simple, one-time sampling process does not depend on the value that is actually retrieved on a given sample, or the shape of the sampled distribution.

We think most researchers share the intuition is that it is much easier, and therefore faster, to produce the expected continuation to a constraining context as in (1a) than it is to produce any continuation to an unconstraining context as in (1b). These contexts are repeated here, without the remainder of the sentences.

(1)	a. The athlete pulled a
	b. Peter says that a

In fact, there is previous evidence that this intuition is correct, though this small literature does not effectively distinguish effects of cloze probability itself from effects of

[1] We are not certain that Smith and Levy (2011) actually do intend a simple, one-time sampling model to be regarded as a process model. It is possible that it is intended as a computational level (Marr, 1982) description of the subject's task, rather than as a description of the algorithm that is used to complete the task.

item constraint. Goldman-Eisler (1958) found that speakers are more likely to pause before producing low probability continuations. Two published studies since then (Cohen & Faulkner, 1983; Nebes, Boller, & Holland, 1986) have examined cloze task latencies in some form. Cohen and Faulkner (1983) had subjects read aloud sentences with a missing final word, asking them to supply this word. They compared the time to read the sentence and supply the final word with the time taken to read the sentence when the final word was provided. This difference was greater when the final word had low cloze probability than when it had high cloze probability. Nebes et al. (1986) used a somewhat finer-grained method, closer to the one we employ in this paper, playing subjects recorded sentences that were missing a final word, and asking subjects to verbally complete the sentence so that it made sense. Subjects took longer to respond to less constraining contexts.

To anticipate the conclusions of the experiments we present here, both cloze probability and item constraint have reliable, and sizable, influences on RT: A cloze response is faster when the response is higher in probability, and is also faster, at a given level of probability, when the item is more constraining. We show that these patterns are consistent with a very simple kind of process model, according to which a cloze response is the winner of a race toward a threshold level of activation; on this model, subjects are reporting the first word to reach this threshold. For other psycholinguistic applications of the notion of a race between activated representations, see e.g. van Gompel, Pickering, and Traxler (2000) or Frauenfelder and Schreuder (1992).

The remainder of this paper proceeds as follows. We present two experiments using a version of the cloze task in which the context was presented in Rapid Serial Visual Presentation (RSVP) format, and the RT to begin speaking was recorded. This method of stimulus presentation enabled timing of the latency to initiate a response without introducing additional variability due to sentence reading time. We find that cloze probabilities in this task are very similar to the standard paper-and-pencil cloze task, warranting inferences from this task to the process of producing a cloze response in the usual untimed version of the task. We find that higher probability responses are produced faster, and that at a given level of probability, a response is produced faster in a more constraining context. Additional analyses show that this latter effect cannot be reduced to an effect of semantic relatedness between responses in constraining contexts. In the General Discussion we describe how a race model naturally captures the RT patterns obtained in the experiments, and we discuss implications of these findings for the interpretation of the cloze probability variable.

Experiment 1

Method

Subjects

Thirty-nine subjects participated in the experiment. Six subjects were excluded from analysis due to a failure to follow the experimental instructions (e.g., not producing any continuation on many trials, or producing full sentence continuations rather than single words), leaving 33 subjects. All subjects were undergraduate students at the University of Massachusetts Amherst, who earned psychology course credit for their participation, and were naïve to the purpose of the experiment. All subjects were native speakers of English with normal or corrected-to-normal vision.

Materials

Each subject was presented with 377 experimental sentence fragments. Of these, 128 were adapted from Staub (2011). The items in Staub (2011) were originally based on items from Altarriba, Kroll, Sholl, and Rayner (1996). Staub (2011) collected cloze norms for each of the items, using the standard paper-and-pencil method, from between 37 and 42 University of Massachusetts undergraduates. Several of the normed items were excluded from the eye movement experiment reported in Staub (2011), but all were included in the current experiment. The remaining 249 items were adapted from Best (2011), some of which were based on materials from Bloom and Fischler (1980). The materials from Best (2011) were previously normed in a paper-and-pencil cloze task by 22 University of Massachusetts undergraduates. For 176 of the 377 fragments, a specific word was established to have high cloze probability for the purposes of the Staub (50 items) or Best (126 items) experiments. An additional 15 filler items, which were not normed in any earlier study, were intermixed with the experimental items. These were short fragments that were included in order to increase the number of short fragments in the experiment and thereby decrease subjects' surprise at an early termination of a fragment. Presentation of five practice items preceded the experiment. The full set of items is available upon request.

Procedure

The experiment was carried out using E-Prime 2.0 experimental software (Schneider, W., Eschman, A., and Zuccolotto, A., 2012), running on a Windows PC computer with a CRT monitor. Subjects were tested individually in a quiet experimental testing room. Prior to beginning the experiment, subjects were fitted with a head-mounted microphone to record their responses. Each trial was initiated by a button press from the subject. A cross then appeared for 1000 ms in the center of the screen. The words of the sentence fragment then appeared successively, centered on the same location, for 300 ms each. At the completion of the fragment, a horizontal line was presented, serving as the prompt for the subject to provide the next word. The subject then had 3000 ms in which to say a response aloud. On each trial, the software recorded a .wav file with a three-second duration, beginning simultaneously with the presentation of the prompt. The items were presented in a random order to each subject.

To begin the experiment, subjects read the instructions "[w]hen you see the blank, please say out loud the next word that you think should be in the sentence. Please try to say the word as quickly as possible, speaking at a natural volume and avoiding saying "umm" or "aah", etc. before you say the word." The experimenter remained in the

testing room while the subject completed the five practice trials, which contained items meant to range in constraint. After completing the practice trials, subjects had an opportunity to ask questions about the experimental procedure. The experimenter then left the room for the remainder of the session, which took approximately 45 min.

Coding of responses

A coder initially listened to and transcribed each of the response recordings without access to their sentence fragment contexts. After this initial coding stage, a second coder examined the responses while inspecting the sentence fragments that preceded them. At this point, several types of corrections were made. Responses that had been transcribed as obvious homophones of the intended response were corrected (e.g., *boarder* was corrected to *border* in response to the fragment *We crossed over the* _____). Obviously mistaken segmentation of words by the initial coder was also corrected (e.g., *teach her* was corrected to *teacher* as a continuation to the fragment *He entered the classroom to ask the* _____). In addition, in cases where there were both plural and singular forms of the same response for one item, these were collapsed by changing the less common form of the word to the more common form (e.g. *floors* in response to *Mary decided to sweep the wooden* _____ was changed to *floor*, which was the more frequent response), as appears to be the usual practice in coding cloze responses, and as was done in Staub (2011) and Best (2011). When subjects responded with more than one word, only the first word of the subject's response was included unless the words formed a compound. Responses that included determiners were coded without the determiner; these cases were very rare, as syntactic constraints prohibited a determiner continuation in most items. A third coder measured the response latencies by visually inspecting the waveforms using the Praat software for speech analysis (Boersma, 2001), placing a cursor at the word onset and exporting the cursor time to a text file. The responses and their latencies were then merged for analysis.

Results

Trial exclusion

Of the 377 experimental items completed by each subject, two were eliminated due to programming errors. In addition, there were 992 trials (8.0%) on which the subject did not make a verbal response within the 3 s deadline, or in a few cases made an indecipherable response, leaving a total of 11,383 codable responses. Trials without a codable response were not included in the analyses below. The rate of failing to respond was related to item constraint; for the items that were defined as having high constraint based on the Staub (2011) and Best (2011) norms, the non-response rate was only 4%. In the analyses reported below, 33 (i.e., the maximum possible number of codable responses) serves as the denominator for all computation of response proportions. Note, however, that both the qualitative patterns and all statistical conclusions are unchanged if the number of valid responses to each item is used as the denominator.

Item constraint and agreement with previous norms

To assess each item's level of constraint, we first computed Shannon entropy, an information-theoretic measure that takes into account how many unique responses an item elicited, and these responses' probabilities; a high-entropy item is one that elicited many distinct responses, with these responses having relatively similar probability. However, in the present data, entropy was almost perfectly correlated with the probability of the item's modal response, $r = -.977$; the higher the probability of the modal response, the lower the entropy. We note that both entropy (e.g., Yun, Mauner, Roland, & Koenig, 2012) and the probability of the modal response (e.g., Schwanenflugel & LaCount, 1988) have been used as measures of item constraint in the literature. In the remainder of this paper we use the probability of the modal response as the measure of item constraint, due to its ease of interpretation, and due to the fact that it is on the same scale as cloze probability. The mean level of item constraint in this experiment, by this measure, was .454. The distribution of constraint values was somewhat bimodal, as expected based on the fact that the items were originally designed to be either very constraining or not; the median was .394, with first and third quartiles of .182 and .727 respectively.

To assess the extent to which responses elicited by the current procedure agreed with the previous paper-and-pencil norms obtained by Staub (2011) and Best (2011), we compared the modal response to each item in the present experiment and in the paper-and-pencil norms, for those items that were specifically identified as highly constraining in the earlier studies. The modal response was the same word for 161 of the 176 items (91.5%). The exceptions were generally cases in which the two distinct responses were closely semantically related, if not synonyms, e.g., *exam* vs. *test*, *cancer* vs. *illness*. In addition, the correlation across all 375 items between the item constraint values in this study and in the original paper-and-pencil norms was $r = .91$. In sum, the switch from a paper-and-pencil paradigm to a speeded spoken response paradigm had little effect on the word that was most likely to be selected, for those items that were highly constraining, and there was also a high degree of correspondence between an item's level of constraint in the paper-and-pencil paradigm and in the current paradigm. These conclusions are important, as they justify the inference from RT results in the current paradigm to conclusions about the interpretation of cloze probabilities in the usual paper-and-pencil paradigm.

Response time analysis

To analyze effects of a range of predictors on RT, we constructed linear mixed-effects models of RT using the lme4 package (version 1.1-7; Bates, Maechler, Bolker, & Walker, 2014) for the R statistical programming language (version 3.1.2; R Core Team, 2014). Several preliminary notes are in order. First, in all the models below, we included the maximal random effects structure for subjects (i.e., subject intercepts, by-subject slopes for each of the fixed effects and their interactions, and correlation parameters), except as noted. We did not include random effects for items, as in the present design item effects cannot be estimated

independently of the fixed effects of interest, especially the item constraint variable, as each item occurs at only one level of item constraint. Second, all fixed effects were centered on their mean prior to being entered into the model, except as noted below. Third, note that we report models of raw as opposed to log-transformed RT, but in all cases models with log-transformed RT as the dependent variable produce qualitatively similar results and identical patterns of significance. Fourth, to check for excessive collinearity we computed the variance inflation factor (VIF) for all predictors, using the vif.mer() function developed by Austin Frank for mixed-effects models (available at https://github.com/aufrank/R-hacks/blob/master/mer-utils.R). Except as noted, VIFs were less than 5. This is a fairly conservative criterion, as a value of 10 is often suggested as the maximum acceptable VIF (e.g., Hair, Anderson, Tatham, & Black, 1995). In the few cases in which a VIF exceeded this value, we modified the model by removing individual predictors with the highest VIFs until all VIFs were below 5, as described in the text.

Finally, we note that we explored effects of three control predictors on RT that are not included in the models we present below: the log frequency of the word that a subject produced (based on the SUBTLEX norms; Brysbaert & New, 2009), the order of the trial in the experiment sequence, and the number of words in the context fragment. Word frequency and trial order did not significantly affect RT, either on their own or in interaction with other factors. However, word frequency was indeed correlated with response probability and item constraint, and we discuss these correlations below.

The length of the context fragment did have a reliable effect on RT, as responses were faster for longer fragments. This variable was also related to item constraint, as more constraining items tended to be longer, with a correlation between the two variables of $r = .49$. However, the effect of item constraint was fully significant in models that also included fragment length, and the two effects did not interact. Moreover, we conducted Experiment 2 in part to eliminate any potential influence of differences between fragments in length or structure. Thus, we leave all three of these control predictors out of the analyses we report here.

The overall mean RT was 1329 ms ($SD = 515$ ms). Fig. 1a illustrates the relationship between cloze probability and RT. Higher cloze responses were issued much faster than lower cloze responses. The relationship appears roughly linear, though there is some indication that RT decreases faster across very low levels of cloze than at higher levels. Interestingly, Smith and Levy (2013) have suggested that there is a logarithmic relationship between predictability and self paced reading time. In Fig. 1b we show RT as a function of $\log_{10}$ cloze probability. In this experiment RT is not quite a logarithmic function of cloze, as a perfectly logarithmic function would show a straight-line relationship on this graph. The relationship appears to be somewhere between a linear and logarithmic function (see Smith & Levy, 2013, Fig. 1). Identifying the best-fitting parameterization of this relationship is far from the aim of this paper, so we leave this issue for future research. Here we assume a linear statistical model.

The results of our statistical analyses are reported in Table 1. Our first statistical model (Model 1) included cloze probability as the only fixed effect, revealing a very large and highly significant effect of this variable on RT (we treat $|t| > 2$ as statistically significant). However, it is conceivably possible that this effect of cloze probability does not reflect a difference between high and low cloze responses to the same item, but rather reflects an effect of item constraint, such that high constraint items generally elicit faster responding. To rule this out, we constructed a model (Model 2) with item constraint and a response's status as a modal or non-modal response to the item as predictors. Modal vs. non-modal response was coded with modal response taking the value .5 and non-modal response the value −.5. The effects of these two variables (with constraint split at .5 for purposes of illustration) are shown in Fig. 2. Modal responses were much faster than non-modal responses; thus, it is clear that the overall effect of response probability cannot be attributed to an effect of item constraint. In addition, there was indeed a sizable effect of item constraint, and a sizable interaction, in the direction of a larger effect of the modal vs. non-modal distinction for high constraint items. This interaction is expected, as the difference in probability between a modal and non-modal response will necessarily be larger for more constraining items.

In order to investigate whether there is an effect of item constraint that is independent of the effect of response probability, we constructed a model of RT that included only responses with cloze probability <.4. Responses with cloze >.5 are necessarily the modal response, so for these responses it is not possible to separately evaluate the influence of response probability and item constraint. In the present data there was only a single item that elicited a response with probability between .4 and .5 where this was not also the modal response, so we restricted our analysis to responses with cloze probability <.4. This model (Model 3) included cloze probability and item constraint, and their interaction, as fixed effects. In this model the cloze probability variable was centered at a value of .2, i.e., the midpoint of the range of cloze probability values. In the model, the effect of cloze probability was significant, the effect of item constraint was significant, and the interaction was not significant. The data patterns are shown in Fig. 3, where for ease of visualization the responses have been divided into two item constraint groups (at .5) and three probability bins (singleton responses with probability 1/33 = .03, non-singleton responses with cloze less than .20, and responses with cloze from .20 to .40). It is clear from the figure that at each level of cloze probability, there is a sizable effect of item constraint, with responses being issued faster when the item was more constraining.

Word frequency

As mentioned above, word frequency was not a significant predictor of RT in this experiment. It is possible, however, that interpreting the relationship between response probability, item constraint, and RT is complicated by differences between the words that were actually produced at different levels of response probability and item constraint. Specifically, it is possible that higher probability

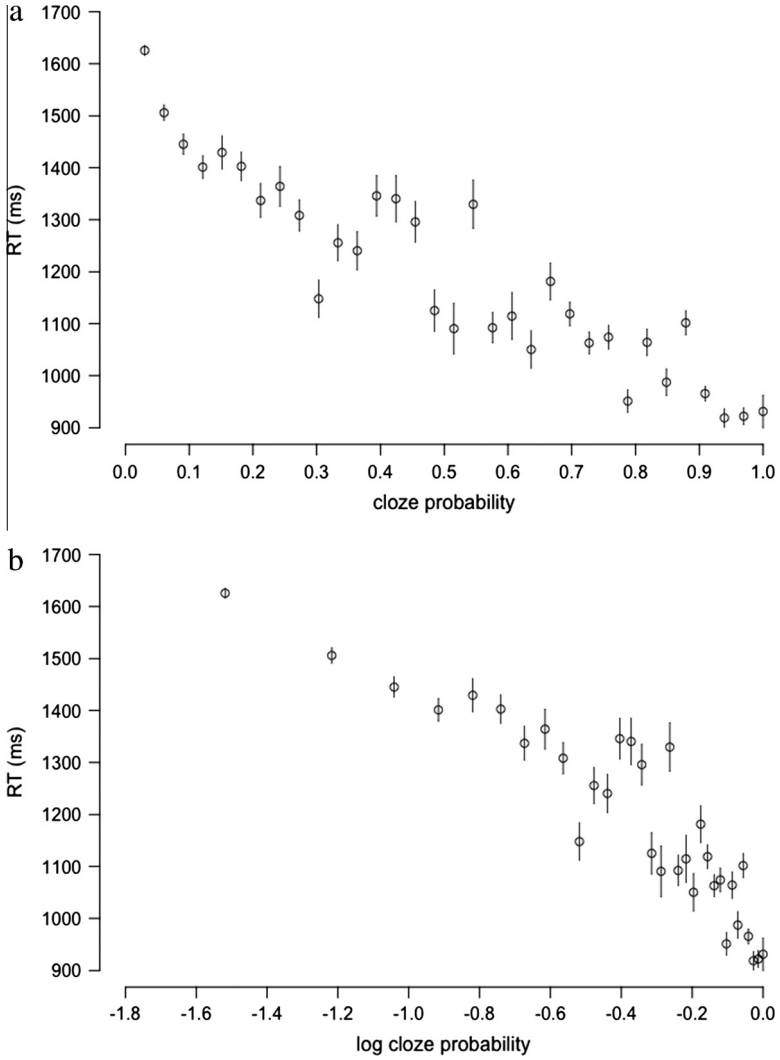

Fig. 1. Mean RT in Experiment 1 at each level of (a) cloze probability and (b) log cloze probability. Error bars represent standard error of the mean.

Table 1
Parameter estimates, standard errors, and *t* values from linear mixed-effects models of RT in Experiment 1. Models are described in the text.

	Estimate (β)	Std. Error	*t* value
Model 1			
Intercept	1338.10	32.03	41.78
Probability	−695.64	34.35	−20.25
Model 2			
Intercept	1355.44	32.98	41.10
Modal vs. Nonmodal	−205.49	11.93	−17.22
Constraint	−548.17	37.68	−14.55
Modal × Constraint	−188.56	36.73	−5.13
Model 3			
Intercept	1338.55	35.38	37.84
Probability	−1124.03	89.26	−12.59
Constraint	−331.31	73.80	−4.49
Prob × Constraint	17.50	470.08	.04
Model 4			
Intercept	1038.04	39.33	26.39
Probability	−1538.42	122.21	−12.59
Constraint	−294.75	37.11	−7.94
LSA	−147.33	37.80	−3.90

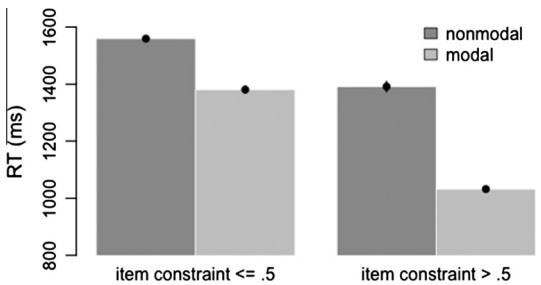

Fig. 2. Mean RT in Experiment 1 by a response's status as modal vs. non-modal response, and item's level of constraint. Error bars represent standard error of the mean.

responses, and responses in high constraint contexts, tend to be more frequent words, and more frequent words are produced faster (e.g., Griffin & Bock, 1998). In fact, the correlations between (log) word frequency and these variables are in the opposite direction from what this hypothesis predicts. Higher probability responses were

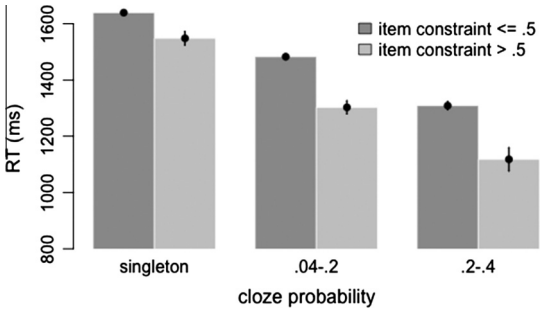

Fig. 3. Mean RT in Experiment 1 by cloze probability and item constraint, restricted to responses with cloze probability <.4. Error bars represent standard error of the mean.

slightly *less* frequent than lower probability responses ($r = -.038$, $p < .001$). Among responses with cloze probability <.4, those that were produced in high constraint contexts were *less* frequent than those that were produced in lower constraint contexts ($r = -.109$, $p < .001$). Thus, the RT patterns obtained here occurred despite, rather than because of, word frequency differences.

Semantic relatedness

Here we present an analysis to evaluate a specific explanation of the finding that a low cloze response is faster in a constraining context. When an item is constraining, non-modal responses are likely to be related in meaning to the modal response. For example, the fragment *He campaigned to win the* _____ was completed with *election* by 20 subjects, for an item constraint value of .61, but other responses were related words such as *contest*, *battle*, *award*, and *prize*. In less constraining contexts, on the other hand, the various responses often appeared to be relatively unrelated in meaning. For example, the item *He heard a faint sound of one* _____ had constraint of only .1, and elicited responses as diverse as *bird*, *girl*, *cat*, and *crying*. Thus, it is possible that the effect of item constraint on RT, at a given level of probability, may actually be an effect of the semantic relatedness of a response to other responses. One way this might come about is by means of facilitatory connections between semantically related words; responding *contest* after *He campaigned to win the* may be relatively fast, even if this response is low in probability, because *contest* benefits from its connection to the highly activated word *election* (see Roland et al., 2012, for a related hypothesis regarding comprehension).

We coded the semantic relatedness of each non-modal response to the corresponding modal response by determining the Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) cosine between these two words using the web-based interface maintained by CU Boulder (lsa.colorado.edu; specifically, LSA cosines were computed using the General Reading topic space). We note, first, that the LSA cosine is a relatively coarse measure of meaning relatedness, based on the similarity of the linguistic contexts in which two words appear. We also note that there may be advantages to a measure taking into account the relatedness of a given response to every other response to an item. However, computing such a measure raises a variety of methodological questions (e.g., whether to count response types or response tokens; how to count the relatedness of a response with itself), so for simplicity we computed the similarity to the modal response, which in any case is likely to be a semantically 'central' response. In the few cases in which there was more than one modal response to an item (i.e., a tie between two responses) we used the average of the cosines. We excluded from analysis any responses that were not included in the corpus (e.g., proper names, acronyms) as well as spaced compounds, which are treated by LSA without regard to word order (e.g., *meat loaf* and *loaf meat* are regarded as the same text). This left 5268 non-modal responses for analysis. LSA cosines range from −1 to 1, with positive values indicating meaning overlap between two terms. The great majority of values are between 0 and 1, with most falling in the range between 0 and .5. For example, the LSA cosine between *contest* and *election* is .34, while the cosine between *lottery* (another response elicited by this item) and *election* is .04.

Non-modal responses were more related in meaning to the modal response as item constraint increased, $r = .244$, $p < .001$. Higher probability non-modal responses were also more related in meaning to the modal response, $r = .328$, $p < .001$. Dividing responses into those elicited in high and low constraint items (item constraint >.5 vs. <=.5) and those with high and low response probability (responses provided by more than one subject vs. singleton responses), the mean LSA cosine for the four groups of responses was: high constraint, non-singleton (.337); high constraint, singleton (.229); low constraint, non-singleton (.251); low constraint, singleton (.171). A mixed-effects linear regression model with LSA cosine as the dependent measure, and with the item's constraint and the probability of the response as predictors (the interaction term was dropped, due to a VIF of over 50) revealed a significant effect of both predictors (item constraint: $b = .153$, $SE = .012$, $t = 13.22$; response probability: $b = .846$, $SE = .046$, $t = 18.45$).

We then constructed a linear mixed effects model of RT for these responses, including as a predictor the LSA cosine between the response and the corresponding modal response, as well as response probability and item constraint; each predictor was centered on its mean. The initial model including interactions of all fixed effects did not show a main effect of semantic similarity, but very high VIFs (>30) for several terms in this model required a retreat to a main-effects only model (Model 4 in Table 1). All three predictors had a significant effect in this model, though the effect of semantic similarity was the smallest both in absolute and standardized terms. Fig. 4 shows mean RT, for the 5268 responses included in this analysis, as a function of item constraint (coded by splitting at .5), response probability (coded as singleton vs. non-singleton response) and LSA cosine with the modal response, split into low, medium, and high semantic similarity based on a split into terciles at .11 and .25. This graph makes apparent that RT does depend on both response probability and item constraint, but does not seem to depend strongly on LSA cosine.

In sum, this analysis suggests several conclusions. It is indeed the case that non-modal responses are more closely related to the modal response when the item is more

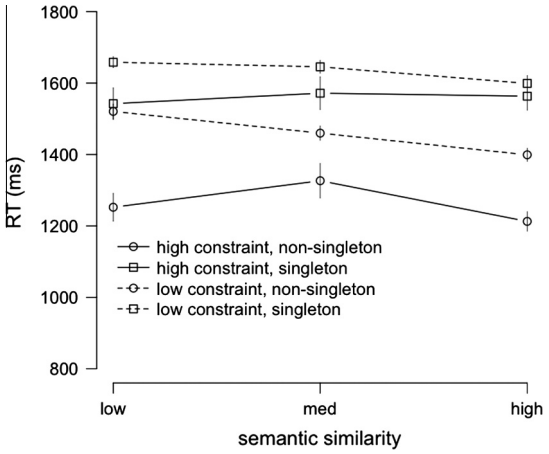

Fig. 4. Mean RT in Experiment 1 for non-modal responses by status as singleton or non-singleton response, item constraint, and LSA cosine with modal response.

constraining. In addition, the semantic similarity between a non-modal response and the corresponding modal response may have an independent effect on RT, though this effect should be regarded with some caution, due to the fact that it emerged in model in which interaction terms were eliminated, and that the effect is not readily apparent in Fig. 4. In any event this effect appears to be smaller than the effects of response probability and item constraint on RT, and the effects of both response probability and item constraint are significant in all models. There is little evidence to support the contention that the effect of item constraint on RT is due in its entirety or even in large measure to semantic relatedness.

Discussion

This experiment showed, first, a very high degree of correspondence between the responses that were produced in a timed cloze task and in the previous paper-and-pencil norms for the same items. This correspondence warrants inferences from the pattern of RTs in the present task to conclusions about the process of producing a cloze response in the typical untimed task. In addition, it suggests that cloze responses in the typical untimed version of the task are not much affected by the opportunity for slow deliberation over what should come next. Subjects produce essentially the same responses with no time limit as when they have only three seconds to respond and have no opportunity to re-read the fragment.

Cloze probability and RT are very strongly related; as Fig. 1 shows, responses with cloze probability near 1 were about 700 ms faster than singleton responses. This relationship held within items, as the modal response to an item was, averaging across levels of item constraint, over 200 ms faster than non-modal responses. In addition, low cloze probability responses that occurred in high constraint contexts were faster than similarly low cloze responses in low constraint contexts. The estimate of the constraint effect in Model 3 indicates that at a given level of response probability, an increase in item constraint from, e.g., .2 to .8 results in an RT decrease of about 200 ms. These effects are not due to differences in the frequency of the words that are actually produced, and cannot be attributed to semantic relatedness among potential responses.

In the General Discussion we take up the question of how these effects are best interpreted. However, we first present the results of a second experiment designed to rule out a potential artifactual explanation of the effect of constraint on RT, and to show that the basic patterns obtained in Experiment 1 are replicable. In Experiment 1, there were uncontrolled structural differences between the high and low constraint contexts, and there were uncontrolled differences in the syntactic and thematic position in which the target word appeared. Thus, it is possible that in high constraint contexts the point in the sentence at which the response prompt is likely to appear was more easily anticipated by subjects than in low constraint contexts. This would be the case if, for example, there are only certain syntactic environments that tend to impose constraint on the next word, and many other syntactic environments that do not. If subjects are able to use these regularities to anticipate the prompt position in high constraint contexts, this would have had the result of generally speeding RT in these contexts. In other words, the apparent effect of item constraint may actually have been an effect of the predictability of the prompt location. In addition, as noted above there was a relatively strong correlation between item constraint and the length of the fragment, and even though the effect of item constraint did not appear to be due to this confound in Experiment 1, we wanted to eliminate this confound entirely. In Experiment 2 all sentence contexts were identical in length and in structure. In these sentences, the position of the target prompt was perfectly predictable in all items, whether high or low constraint.

Experiment 2

Method

Subjects

Forty-three subjects from the same pool as Experiment 1 participated in Experiment 2, of whom three were excluded due to a high rate of failure to respond, leaving forty subjects whose data were analyzed. No subject participated in both experiments.

Materials

Each subject was presented with 338 experimental sentence fragments. There were no filler fragments in this experiment. The experimental fragments differed from those used in Experiment 1 in that all were exactly five words in length, and all used the same sentence structure, i.e., The ADJ NOUN VERB[+past] DET _____. The determiner that was the last word of the fragment was usually the definite article (e.g., *The challenging topic stumped the*_____), though possessive pronouns or quantifiers were used in several items (e.g., *The barefoot man stubbed his* _____; *The happy gardener planted some* _____). The items were designed so that based on native speaker

intuitions they varied along the full range of constraint. The experiment itself was used to verify these intuitions. The full set of items is available upon request.

Procedure

The procedure was identical to Experiment 1, except that a voice key, operated through a second microphone, was used in conjunction with the experimental software to record the latency from the onset of the response prompt until the subject began speaking.

Coding of responses

Coding was identical to Experiment 1, except that we developed software that enabled a single coder to first transcribe what he or she heard, then to immediately see the preceding fragment so as to correct any homophone transcription errors or word segmentation errors. In addition, the use of the voice key eliminated the labor-intensive process of visual wave form inspection for the purpose of identifying the response latency on each trial.

Results

Trial exclusion

No audible response was produced on 482 out of 13,520 trials (3.6%). We attribute the lower non-response rate in this experiment to the predictability of the timing of the response prompt. Of the remaining 13,038 trials, a valid RT was registered by the voice key on 12,045 trials (92.4%). Trials with a transcribed response but no valid voice key trigger were included in the analysis of cloze probabilities. As in Experiment 1, this analysis used the maximum number of potential responses in the denominator, in this case 40; as in Experiment 1, using the actual number of valid responses to each item in the denominator results in identical conclusions. The analysis of RTs includes only those trials with a valid voice key trigger.

Item constraint

The mean level of item constraint was .445, which was very similar to Experiment 1. The distribution of item constraint values was more uniform than in Experiment 1, with the median at .375, and first and third quartiles at .225 and .650, respectively.

Response time analysis

Overall, RTs were substantially faster in Experiment 2 than in Experiment 1 (mean = 1062 ms, SD = 511 ms). Again, we attribute this to subjects' preparedness for the response prompt.

Fig. 5 shows the overall relationship between cloze probability and RT. As in Experiment 1, this relationship is roughly linear, but trends toward a logarithmic relationship. Table 2 shows the results of the statistical analyses. A linear mixed-effects model of RT with cloze probability as the sole fixed effect (Model 1) reveals a highly significant effect of this factor. Fig. 6 shows that the relationship between cloze probability and RT holds within items; the modal response to an item was issued much faster than were other responses. A model with a response's status as modal vs. non-modal, item constraint, and their interaction as fixed effects (Model 2) reveals highly significant effects of both factors, and a significant interaction, in the direction of a larger difference between modal and non-modal responses when the item was more constraining.

Fig. 7 shows that, among low-probability responses, there was a substantial effect of item constraint, such that responses at a given level of response probability were faster when the item was more constraining. A model of RTs for responses with cloze probability <.4 (Model 3) reveals a highly significant effect of cloze probability, a highly significant effect of item constraint, and a non-significant interaction.

Semantic relatedness

Because the semantic relatedness analysis of Experiment 1 yielded somewhat equivocal results, we repeated it for Experiment 2. The restriction to non-modal responses that were represented in the LSA corpus yielded 6605 observations. The correlation between response probability and relatedness to the modal response was similar to Experiment 1 (r = .386, p < .001), as was the correlation between item constraint and relatedness (r = .190, p < .001). Again, both factors were significant in a mixed-effects model (item constraint: b = .107, SE = .013, t = 8.02; response probability: b = .851, SE = .031, t = 27.10). Unlike in Experiment 1, including the interaction did not inflate the VIF, and there was indeed a significant interaction between these variables (b = −.694, SE = .180, t = −3.85). As both item constraint and response probability increased, the LSA cosine increased less than would be predicted on an additive model.

We then constructed a model of RT as for Experiment 1 with LSA cosine, response probability, and item constraint as predictors. We note that in this case the VIF was not inflated when the model included interaction terms, though it was necessary to remove random interaction slopes to obtain convergence, as well as the random effect correlation parameters. The results are shown in Model 4. As in Experiment 1, all three main effects are significant, but the effect of semantic relatedness is again the smallest. In this model, two of the two-way interactions are also significant. It appears that RT may be slower than an additive model would predict when a non-modal response is both high in probability and strongly related to the modal response, and that RT may be faster than predicted by an additive model when the item is constraining and semantic relatedness is high. We note that the latter finding contrasts with a recent finding in the comprehension literature by Yun et al. (2012), who found that in self-paced reading, semantic similarity of a word to other potential continuations reduced reading time only in low constraint contexts. The effects are visualized in Fig. 8, where RT is displayed for the 6605 observations in this analysis as a function of response probability (singleton vs. non-singleton response), constraint (split at .5) and semantic similarity to the modal response (split into terciles at .10 and .26). In this figure, unlike in Fig. 4, there is indeed a clear relationship between semantic similarity and RT, though again the separate effects of response probability and item constraint are both apparent.

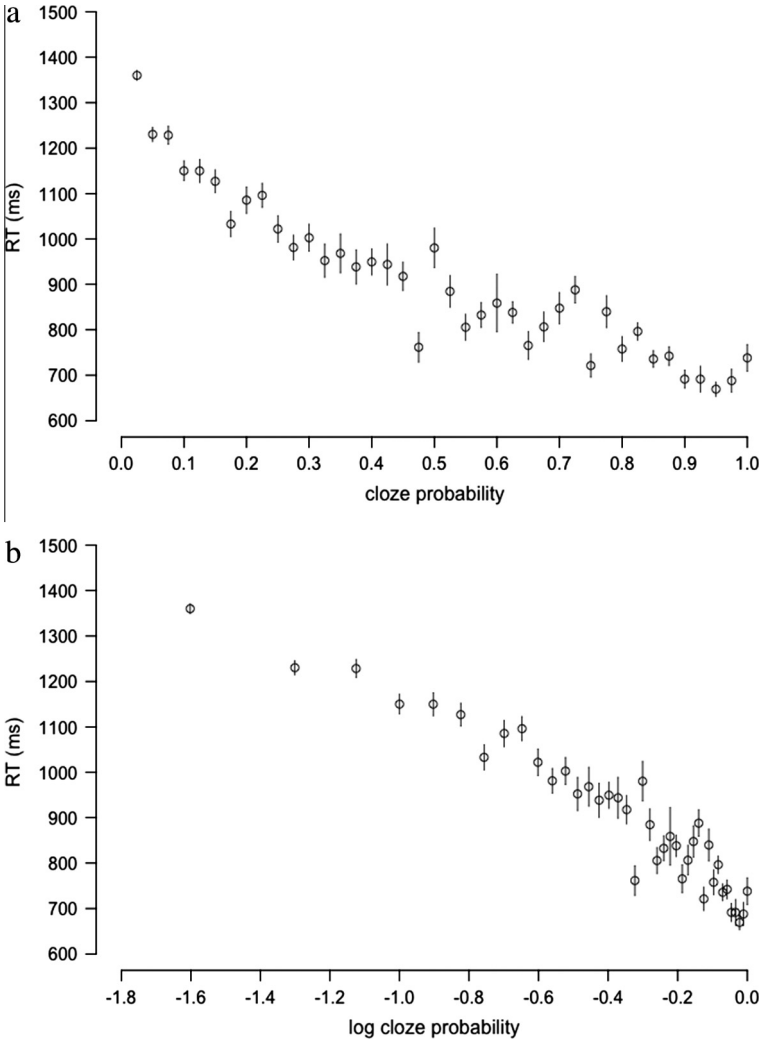

Fig. 5. Mean RT in Experiment 2 at each level of (a) cloze probability and (b) log cloze probability. Error bars represent standard error of the mean.

Discussion

This experiment provides a very straightforward replication of the patterns in Experiment 1 in a context in which the position of the target prompt was perfectly predictable. There was a very large effect of cloze probability on RT, with higher cloze responses being produced faster. This relationship held within items, as the modal response to each item was produced much faster than non-modal responses. In addition, this experiment replicated the finding that among low cloze responses, a response was produced substantially faster when it occurred in a high constraint context. We note that all effects in this experiment are similar in size to those in Experiment 1, despite the fact that responses in this experiment were faster overall. Evidently, the RT advantage for high constraint contexts cannot be attributed to the potential for subjects to anticipate the prompt location in those contexts. In this experiment the effect was clearly replicated despite the prompt occurring at the same point in all items. Indeed, in Experiment 2 the estimate of the effect of item constraint on RT for low cloze responses was numerically larger than in Experiment 1.

The semantic relatedness analysis increased somewhat the strength of the evidence that relatedness of a non-modal response to the corresponding modal response does have an effect on RT. However, this analysis confirmed that the effect of relatedness does not subsume the effect of item constraint; to the contrary, the effect of item constraint was larger than the effect of relatedness in a model that included both.

General discussion

These experiments establish two important RT effects in the cloze task. First, higher probability responses are issued faster than lower probability responses. This relationship holds in general, and it also holds within items, as the modal response to an item is issued much

Table 2
Parameter estimates, standard errors, and t values from linear mixed-effects models of RT in Experiment 2. Models are described in the text.

	Estimate (β)	Std. Error	t value
Model 1			
Intercept	1064.15	26.82	39.68
Probability	−671.49	28.68	−23.41
Model 2			
Intercept	1063.74	26.98	39.42
Modal vs. Nonmodal	−191.12	11.08	−17.25
Constraint	−536.92	36.78	−14.60
Modal × Constraint	−85.98	37.17	−2.31
Model 3			
Intercept	1037.80	28.10	36.93
Probability	−1259.18	73.93	−17.03
Constraint	−423.26	61.12	−6.93
Prob × Constraint	−426.61	423.82	−1.01
Model 4			
Intercept	1212.82	31.31	38.73
Probability	−1540.54	93.31	−16.51
Constraint	−302.03	42.59	−7.09
LSA	−160.28	33.42	−4.80
Prob × Constraint	991.75	529.69	1.87
Prob × LSA	863.58	337.47	2.56
LSA × Constraint	−575.50	164.31	−3.50
Prob × Constraint × LSA	209.94	2209.15	.10

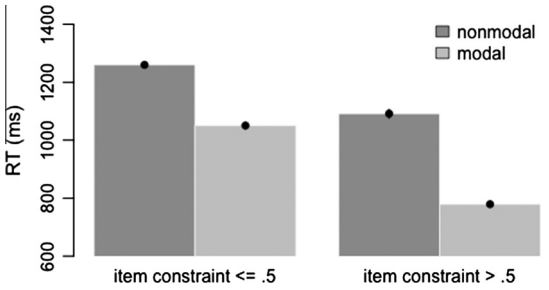

Fig. 6. Mean RT in Experiment 2 by a response's status as modal vs. non-modal response, and item's level of constraint. Error bars represent standard error of the mean.

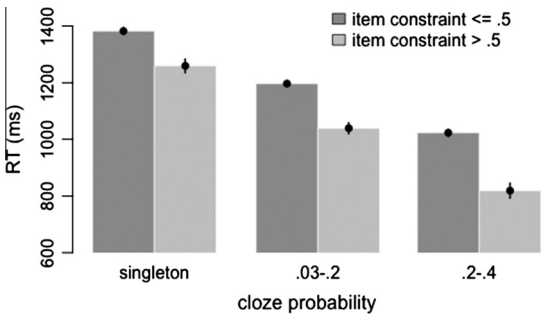

Fig. 7. Mean RT in Experiment 2 by cloze probability and item constraint, restricted to responses with cloze probability <=.4. Error bars represent standard error of the mean.

faster than non-modal responses. This is true for both high constraint and low constraint items. Second, responses at a given level of cloze probability are issued faster in high

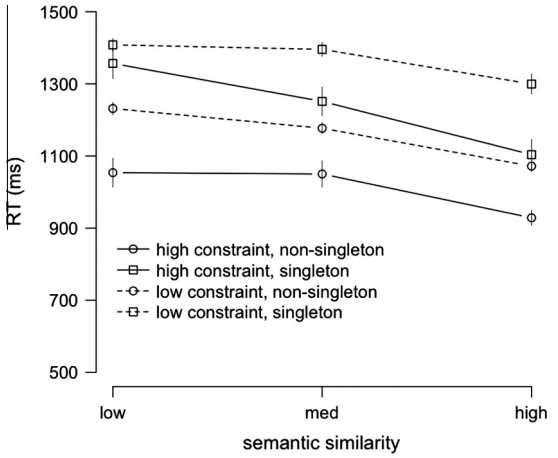

Fig. 8. Mean RT in Experiment 2 for non-modal responses by status as singleton or non-singleton response, item constraint, and LSA cosine with modal response.

constraint contexts. For example, though a response with cloze probability of .2 occurring in a context that also elicits a response with cloze probability of .8 is produced slower than the alternative high probability response, it is produced faster than a response with cloze probability of .2 when that response is the modal response to the context in which it occurs.

Though low cloze responses in high constraint contexts tend to be semantically related to the modal response, this semantic relatedness is not a main source of the RT advantage for these responses. There does appear, however, to be a relatively small effect of semantic relatedness on RT. In the following discussion we focus on the two larger effects of response probability and item constraint, though we discuss below the issue of integrating the effect of semantic relatedness into the model we propose.

The main goal of the present work was to inform our understanding of what cloze probability represents by developing a model of how cloze responses are actually produced. The RT effects constrain such a model, which should explain why more common responses to a given item are faster, and why responses are faster when an item elicits a single response with higher probability. As we noted in the Introduction, a model on which the subject selects a response by sampling once from a subjective probability distribution does not predict such RT effects, in the absence of additional mechanisms. In what follows, we show that a model in which potential responses are independently activated, and race towards a threshold level of activation, naturally predicts these RT effects. We discuss implications of this conclusion for the interpretation of the cloze probability variable.

Race model simulations

The general class of model that we use to account for the empirical data is a very familiar one in cognitive psychology. A wide variety of tasks that deliver both response proportion and RT data have been modeled as evidence

accumulation processes. The basic idea underlying this framework is that the process of selecting a response on each trial involves gradual accumulation of evidence over time, with the response that is issued on that trial being the first response to reach an evidence threshold. Prominent models include Ratcliff's diffusion model (1978; Ratcliff & McKoon, 2008), Usher and McClelland's (2001) leaky competing accumulator model, Reddi and Carpenter's (2000) LATER model, Nosofsky and Palmeri's (1997) exemplar-based random walk model (EBRW), and Brown and Heathcote's (2008) linear ballistic accumulator (LBA) model. Generally, the evidence accumulation framework is used to model data from two-alternative tasks, but it can be extended to situations where multiple response alternatives are present (e.g., Brown & Heathcote, 2008), as in the cloze task. The very simple model we present here is most similar in spirit to a poisson counter model (Smith & Van Zandt, 2000; Townsend & Ashby, 1983) in which multiple responses independently accrue evidence toward a response criterion, without any facilitatory or inhibitory interaction between them.

We simulate a race process as follows. Consider ten potential responses to a sentence fragment, each independently accruing activation. The response that is produced is the first to reach a threshold. There is both between-response and within-response variability in the rate at which activation accrues, and therefore in how long it takes a response to reach the threshold. We present a simulation in which the overall mean time for a response to reach the threshold is one second. The mean finishing time for the ten responses varies from 955 ms to 1045 ms, in 10 ms increments. Within-response finishing times are assumed to be Normally distributed with a standard deviation of 50 ms. The between-response variability may be thought of as reflecting stable differences in the relative rate with which different responses are activated by the context, while the within-response variability may be thought of as reflecting trial-to-trial variability in each response's rate of activation, due either to differences between subjects or to within-subject variability.

We simulate 100,000 races between the responses, i.e., 100,000 cloze trials. On each trial, we determine which of the ten responses finishes first and the finishing time of that response. Fig. 9 illustrates the relationship between the proportion of trials won by a given response and the mean RT on the trials won by that response. These are plotted with the points labeled *race 1*. It is evident that trials that are won by higher probability winners are completed faster. (Indeed, the plot shows a logarithmic trend in the relationship between winning proportion and mean RT, as in our experiments.) This happens for a simple reason. When a relatively fast finishing time is sampled from the distribution of finishing times for the response with the fastest average time (955 ms), this time will almost always be the winning time, as other responses very rarely have times that are as fast. The response with the second fastest average time (965 ms) will tend to beat the fastest response only when the response that is fastest, on average, does not have a particularly fast time on that trial. The response with the third fastest average time (975 ms) will beat the fastest and second fastest only if both of those do not have particularly fast times, and so on.

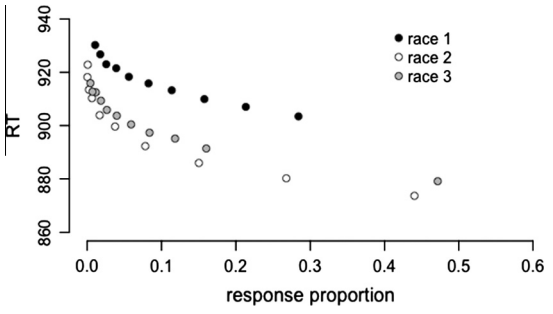

Fig. 9. Mean RT by winning probability, in three race model simulations described in the text.

It is important to note that all that is needed for this general pattern to emerge is that there is both within- and between-response variability in finishing time. The nature of this variability is not particularly important in producing this pattern. For example, the within-response variability in finishing time may be modeled as right-skewed distribution, such as a gamma distribution, which is the distribution of finishing times if activation accumulation is a poisson process (Townsend & Ashby, 1983).

We model the effect of item constraint in two ways. In *race 2*, we make the mean finishing times of the ten responses more variable. This assumes that the effect of a constraining context is to spread out the distribution of lexical activation, such that in a constraining context there is a greater activation difference between the most- and least-activated words; this is to say that a constraining context is one in which some words are especially activated as potential continuations, compared to a neutral context, while the activation of other words is decreased. In this simulation, the finishing times have the same overall mean of 1 s, but the mean finishing times for the ten responses now vary in 20 ms increments from 910 to 1090 ms, rather than in 10 ms increments as in race 1. The within-response, trial-to-trial variability in finishing time is unchanged, with each response's finishing times varying Normally with standard deviation of 50 ms. The results of 100,000 runs are shown on the plot. Because there is more difference between responses in mean finishing time, responses vary more in their winning proportions. While the response with the fastest mean finishing time won 28% of trials in race 1, the response with the fastest mean time won 44% of trials in race 2. The same relationship between winning proportion and RT observed in race 1 obtains in race 2.

In comparing race 1 and race 2, it is evident that at a given level of winning probability, RT is faster in race 2. In other words, the second critical pattern in the data has emerged: RT is faster at a given level of response probability when constraint is higher. Like the effect of response probability on RT, the effect of constraint on RT is simply explained. In race 2, the difference between responses' mean finishing time is greater than in race 1, with some responses being faster, on average, than any of the responses in race 1. Thus, in order to win the race on a given trial in race 2, a response must reach the threshold very quickly.

It is not obvious that the correct way to model item constraint is simply by making lexical activation more variable. In *race 3*, we increase constraint from race 1 in a different way. We assume here that a high constraint context is one in which there is a single response that is very rapidly activated by the context. (The patterns we describe here also hold if more than one rapidly activated response is added.) On this view, the sum of lexical activation is not fixed; a constraining context has the effect of highly activating one or more words, without reducing the activation of other words compared to a neutral context. We add an eleventh response to those competing in race 1, and this response reaches the threshold very quickly on average, with a mean finishing time of 915 ms, i.e., 40 ms faster than the response that was the fastest in race 1. This response has the same standard deviation, 50 ms. The results of 100,000 runs are shown on the plot. In race 3, the newly added, rapidly activated response wins 47% of races. Again, the same relationship between winning proportion and RT is in evidence. In addition, as in the comparison of race 2 and race 1, the comparison between race 3 and race 1 shows that RT is faster, at a given winning proportion, in the more constraining context. Thus, this pattern holds whether an increase in constraint is modeled as an increase in the variability of responses' rate of activation, or if it is modeled by adding a single rapidly activated response to the distribution present in a lower-constraint context.

Another comparison between race 1 and race 3 is of interest. Consider the second most likely winner of race 3, compared to the most likely winner of race 1. These are responses that have exactly the same underlying distribution of finishing times, with a mean finishing time of 955 ms and standard deviation of 50 ms. In race 1, this response wins 28% of the time, with a mean finishing time of 903 ms when it does win. In race 3, this response wins only 16% of the time, but with a mean finishing time of 891 ms when it wins. This demonstrates how the presence of a single very fast response in race 3 influences the dynamics of the race process overall; it makes other responses less likely to win, but ensures that their wins are accompanied by faster RTs. It also illustrates that there is not a monotonic mapping, across contexts, between a response's winning probability and its typical rate of activation (i.e., mean finishing time).

It may be noted that in the comparison of race 2 and race 3, the RT pattern under consideration does not hold. Race 3 has higher constraint, as the modal response has a higher winning proportion, but the points on the plot are slightly above those for race 2, i.e., RT is slightly slower. This shows that it is not *required* for a race model to predict the pattern in which higher constraint reduces RT. Changes in constraint can be modeled in many different ways. In race 3, there is less variability in mean finishing times than in race 2, but on the other hand there is a fast outlier. This particular way of increasing constraint would seem to have little to argue for it on theoretical grounds. What we have shown, however, is that two different plausible ways of implementing the notion of contextual constraint do result in the observed pattern.

The qualitative patterns present in the data are naturally predicted by a model on which potential responses independently accrue activation towards a response threshold. In saying that these patterns are naturally predicted by this model, we mean that they are predicted with a very minimal set of assumptions: (a) there is between-response variability in mean finishing time; (b) there is within-response trial-to-trial variability in finishing time; and (c) an increase in constraint may be due either to greater between-response variability in finishing time, or to the addition of a response that is very fast, on average, relative to other responses. Note that we do not assume any interaction, either facilitatory or inhibitory, between the potential responses. The finishing time of each response, on a given trial, is independent of the finishing time of all other responses.

Given the salience of the two critical RT effects in the data (i.e., their size and replicability), they are benchmark effects that any process model should be able to account for. Obviously, the question is not settled in favor of an activation-based race model of the type we have introduced. However, this model captures the qualitative patterns easily, using evidence accumulation assumptions that are extremely familiar in models of a wide variety of tasks in which the data include both response proportions and RTs. Our own future research in this area is focused on methods for parameterizing the model so as to provide the best possible quantitative fit to the data, and determining whether other modeling frameworks are also able to account for the critical effects.

An obvious question is how the semantic similarity effect could be captured by an activation-based race model. It is perhaps unsurprising that semantic similarity should be related to both response probability and item constraint, on the assumption that a context will tend to activate multiple related words (e.g., Roland et al., 2012). However, it is arguably more surprising that semantic similarity has an effect on RT that is independent of these other factors. In the architecture of the model we have proposed, it is not obvious how any other factor can have an influence on RT that is independent of the effects of response probability and item constraint, as any factor that influences a response's finishing time will also influence the probability that it wins the race. Though the effect of semantic similarity on RT is small, we do regard this effect as a challenge for the model. It is an also an important question for future research to determine the sources of the opposite interactions between semantic similarity and constraint obtained in Experiment 2 of the present study and in Yun et al. (2012).

Theoretical implications

If the process of producing a cloze response is best understood as an activation-based race process, what subjects are actually doing in the cloze task is reporting the first word that reaches a threshold level of activation; we may think of this more simply as reporting the first word that comes to mind. Cloze probability, then, is the probability that a word is the first to come to mind as a continuation, given the cloze prompt. What are the theoretical implications of this view of cloze probability? Specifically, what implications does this view have for the idea that cloze probability is a measure of *predictability*,

or of speakers' estimate of a word's conditional probability? This boils down to the question of the relationship between two probabilities: the probability that a word is the first to reach a threshold level of activation, given a cloze prompt, and the speaker's estimate of its conditional probability given that prompt.

These two probabilities must be related, but it is unlikely that they are simply identical. There is already evidence that factors in addition to a word's empirically determined conditional probability have an influence on cloze probability. As noted in the Introduction, Smith and Levy (2011) found that lexical properties themselves influence cloze probability, so that, e.g., familiar words are especially likely in cloze responses. This would be expected if, as in many theories of lexical access and processing (e.g., Coltheart, Rastle, Perry, Langdon, & Zeigler, 2001; Levelt, Roelofs, & Meyer, 1999; Morton, 1969), the rate with which a lexical entry is activated in the course of either comprehension or production depends on the strength of that entry in long term memory. A word that is highly likely as a sentence continuation may be relatively unlikely as a continuation in the cloze task if it is difficult to retrieve. Consider, for example, likely cloze responses to the fragment *To determine how fast his engine was revving, the race car driver checked his* _____. We assume that the very low frequency word *tachometer* is unlikely to be produced by many subjects, even though many might recognize it as the 'correct' continuation.

Smith and Levy (2011) also found that participants are biased to produce words that are lexically primed by the words in the preceding context, which is expected on the common assumption that exposure to a word causes spreading activation to the word's associates (e.g., Neely, 1977). It is also possible that other types of word–word associations increase the probability of a cloze response. Subjects may be likely to produce words that are phonologically primed by words in the fragment, and especially when the task is in written form, they may be likely to produce words that share orthographic properties with words in the fragment.

More generally, the factors that influence a word's rate of activation by a cloze prompt, and therefore its probability of winning a race to threshold, may depend on all the details of activation of semantic, syntactic, and event-related representations. A perusal of actual cloze responses reveals these complexities. Consider, from Experiment 2, the item *The confused man asked the* _____. The modal response to this item was *question*, with probability .55. But for many of the lower-probability responses, including *clerk*, *attorney*, *police officer*, *wife*, *woman*, *person*, *lady*, *girl*, and *teller*, the response was an indirect object noun, rather than a direct object. This illustrates how the response may depend on the relative activation of syntactic structures, in this case, double object vs. simple transitive. The large literature on syntactic priming in production (e.g., Bock, 1986; Pickering & Branigan, 1998) suggests that it should be possible to prime continuations that are consistent with one syntactic frame or the other. While the use of four different terms for a female human (*wife, woman, lady,* and *girl*) may be due to simple semantic priming from *man*, the use of *clerk*, *attorney*, and *police officer* may arise from activation of a specific event scenario in which a man is confused. Again, we suspect that it is possible to prime a response of this sort by activating a specific relevant scenario. Finally, this item also elicited responses of *time*, *date*, and *directions*. In these cases, the response may reflect the activation of a relatively fixed idiom (e.g., *asked the time*). Such idioms may be over-represented in cloze completions, if idioms are represented in a unitary manner (e.g., Swinney & Cutler, 1979). For example, we assume that *bucket* would be very dramatically over-represented in completions of *The man kicked the* _____.

Regardless of the details of how lexical activation in the cloze task is determined, viewing cloze probability as a measure of relative lexical activation suggests a particular perspective on the ubiquitous effects of cloze probability in comprehension experiments: What psycholinguists have called predictability effects may be parsimoniously re-described as contextual activation effects. This re-description simplifies an account of the causal mechanism by which cloze probability operates in comprehension. If cloze probability is seen as a measure of a word's predictability, it remains to be explained exactly how predictability causes a higher cloze word to be processed more easily than a lower cloze word. One way that this causal link can be elaborated is by means of the notion of activation; a predictable word may be easier to process because it is more strongly activated. Indeed, some theorists (e.g., DeLong, Urbach, & Kutas, 2005; Kutas, DeLong, & Smith, 2011) have equated prediction with 'pre-activation'. However, if cloze probability is itself seen as a measure of relative activation by context, rather than as a measure of predictability, one of the links in the proposed causal mechanism has been eliminated; we may now propose that words with higher cloze probability are easier to process in comprehension simply because cloze probability is itself a measure of the word's relative level of (pre-) activation.

Conclusion

In two experiments, RT in the cloze task was found to vary based both on the response's probability and on the item's constraint: Higher probability responses were faster, and at a given level of probability, a response was faster in a higher constraint context. These patterns are consistent with a process model of the cloze task in which potential responses independently accrue activation towards a response threshold. We argue that on this conception of the cloze task, it is not obvious that cloze probability is a measure of predictability *per se*. Instead, it may be a measure of a word's relative level of activation by the cloze prompt, which is likely to be influenced by many factors in addition to the word's conditional probability. We suggest that this conception of the cloze probability variable offers a re-interpretation of cloze probability effects in comprehension.

Authors' note

The authors thank Lap Keung, Marie DiCienzo, Anne Lee, Carol Santoro, Jacob Dustin, and Michelle Fuentes for

extensive help with data collection and coding. Thanks also to audiences at Architectures and Mechanisms of Language Processing, September 2012, Riva Del Garda, Italy; the 53rd Annual Meeting of Psychonomic Society, November 2012, Minneapolis, MN; the Haskins Laboratory Colloquium, November 2012, New Haven, CT; Mayfest 2013, University of Maryland Department of Linguistics; and the McMaster University Cognitive Science of Language Colloquium.

References

Altarriba, J., Kroll, J. F., Sholl, A., & Rayner, K. (1996). The influence of lexical and conceptual constraints on reading mixed-language sentences: Evidence from eye fixations and naming times. *Memory & Cognition, 24*, 477–492.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4_. R package version 1.1-7*. <http://CRAN.R-project.org/package=lme4>.

Best, L. A. (2011). Attentional cues during speech perception. *Electronic doctoral dissertations for UMass Amherst*. Paper AAI3482585. <http://scholarworks.umass.edu/dissertations/AAI3482585>.

Bloom, P. A., & Fischler, I. (1980). Completion norms for 329 sentence contexts. *Memory & Cognition, 8*, 631–642.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology, 18*, 355–387.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*, 341–345.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology, 57*, 153–178.

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

Cohen, D., & Faulkner, D. (1983). Word recognition: Age differences in contextual facilitation effects. *British Journal of Psychology, 74*, 239–243.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*, 204–256.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience, 8*, 1117–1121.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior, 20*, 641–655.

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology, 44*, 491–505.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language, 41*, 469–495.

Ferreira, V. S. (1996). Is it better to give than to donate? Syntactic flexibility in language production. *Journal of Memory and Language, 35*, 724–755.

Frauenfelder, U. H., & Schreuder, R. (1992). Constraining psycholinguistic models of morphological processing and representation: The role of productivity. In *Yearbook of morphology 1991* (pp. 165–183). Netherlands: Springer.

Frisson, S., Rayner, K., & Pickering, M. J. (2005). Effects of contextual predictability and transitional probability on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 862–877.

Goldman-Eisler, F. (1958). Speech production and the predictability of words in context. *Quarterly Journal of Experimental Psychology, 10*, 96–106.

Griffin, Z. M., & Bock, K. (1998). Constraint, word frequency, and the relationship between lexical processing levels in spoken word production. *Journal of Memory and Language, 38*, 313–338.

Hair, J. F., Jr., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate data analysis* (3rd ed.). New York: Macmillan.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL* (Vol. 2, pp. 159–166).

Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. In M. Bar (Ed.), *Predictions in the brain: Using our past to generate a future* (pp. 190–207). New York: Oxford University Press.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307*, 161–163.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.

Levelt, W. J. M., Roelofs, A., & Meyer, A. (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.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

Marr, D. (1982). *Vision*. Cambridge, MA: MIT Press.

McDonald, S. A., & Shillcock, R. C. (2003). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research, 43*, 1735–1751.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review, 76*, 165–178.

Nebes, R. D., Boller, F., & Holland, A. (1986). Use of semantic context by patients with Alzheimer's Disease. *Psychology and Aging, 1*, 261–269.

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

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104*, 266–300.

Pickering, M. J., & Branigan, H. P. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language, 39*, 633–651.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences, 11*, 105–110.

R Core Team (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. <http://www.R-project.org/>.

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

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

Rayner, K., Ashby, J., Pollatsek, A., & Reichle, E. D. (2004). The effects of frequency and predictability on eye fixations in reading: Implications for the E-Z Reader model. *Journal of Experimental Psychology: Human Perception and Performance, 30*, 720–732.

Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review, 3*, 504–509.

Reddi, B., & Carpenter, R. H. S. (2000). The influence of urgency on decision time. *Nature Neuroscience, 3*, 827–831.

Roland, D., Yun, H., Koenig, J.-P., & Mauner, G. (2012). Semantic similarity, predictability, and models of sentence processing. *Cognition, 122*, 267–279.

Schneider, W., Eschman, A., & Zuccolotto, A. (2012). *E-Prime User's Guide*. Pittsburgh: Psychology Software Tools Inc.

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.

Schwanenflugel, P. J., & LaCount, K. L. (1988). Semantic relatedness and the scope of facilitation for upcoming words in sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 344–354.

Sheridan, H., & Reingold, E. M. (2012). The time course of predictability effects in reading: Evidence from a survival analysis of fixation durations. *Visual Cognition, 20*, 733–745.

Smith, N. J., & Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 1637–1642).

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*, 302–319.

Smith, P. L., & Van Zandt, T. (2000). Time-dependent Poisson counter models of response latency in simple judgment. *British Journal of Mathematical and Statistical Psychology, 53*, 293–315.

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

Staub, A. (2011). The effect of lexical predictability on distributions of eye fixation durations. *Psychonomic Bulletin & Review, 18*, 371–376.

Swinney, D. A., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior, 18*, 523–534.

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly, 30*, 415–433.

Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes*. New York: Cambridge University Press.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108*, 550–592.

van Gompel, R. P. G., Pickering, M. J., & Traxler, M. J. (2000). Unrestricted race: A new model of syntactic ambiguity resolution. In A. Kennedy, R. Radach, D. Heller, & J. Pynte (Eds.), *Reading as a perceptual process* (pp. 621–648). Oxford: Elsevier.

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology, 83*, 176–190.

Yun, H., Mauner, G., Roland, D., & Koenig, J.-P. (2012). The effect of semantic similarity is a function of contextual constraint. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th annual conference of the cognitive science society* (pp. 6 pages). Austin, TX: Cognitive Science Society.