Journal of Memory and Language 66 (2012) 123–142

Contents lists available at SciVerse ScienceDirect

Journal of Memory and Language

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

Distinguishing speed from accuracy in scalar implicatures

Lewis Bott [a,*], Todd M. Bailey [a], Daniel Grodner [b]

[a] School of Psychology, Cardiff University, United Kingdom
[b] Department of Psychology, Swarthmore College, United States

ARTICLE INFO

Article history:
Received 20 May 2011
revision received 22 September 2011
Available online 1 November 2011

Keywords:
Scalar implicatures
Inferences
Pragmatics
Language processing
Speed-accuracy trade-off

ABSTRACT

Scalar implicatures are inferences that arise when a weak expression is used instead of a stronger alternative. For example, when a speaker says, "Some of the children are in the classroom," she often implies that not all of them are. Recent processing studies of scalar implicatures have argued that generating an implicature carries a cost. In this study we investigated this cost using a sentence verification task similar to that of Bott and Noveck (2004) combined with a response deadline procedure to estimate speed and accuracy independently. Experiment 1 compared implicit upper-bound interpretations (*some* [but not all]) with lower-bound interpretations (*some* [and possibly all]). Experiment 2 compared an implicit upper-bound meaning of *some* with the explicit upper-bound meaning of *only some*. Experiment 3 compared an implicit lower-bound meaning of *some* with the explicit lower-bound meaning of *at least some*. Sentences with implicatures required additional processing time that could not be attributed to retrieval probabilities or factors relating to semantic complexity. Our results provide evidence against several different types of processing models, including verification and nonverification default implicature models and cost-free contextual models. More generally, our data are the first to provide evidence of the costs associated with deriving implicatures *per se*.

Introduction

Using a weak expression from a set of stronger alternatives often implies that the stronger alternatives are not applicable. For example using *some* instead of *all* or *many* in the sentence "John read some of Chomsky's books," can be taken to mean that John did not read all of Chomsky's books, even though the use of *some* is logically consistent with *all*. Importantly, the negation of *all* cannot be part of the literal meaning of *some* because the *not all* component of the sentence can be defeased or cancelled, as in, "In fact, he's read all of them," without any infelicity arising. These inferences have been referred to as scalar implicatures, or scalar inferences, because they involve an entailment (or *semantic*) scale (Horn, 1972), and were assumed to be derived using a form of Gricean reasoning ("if the speaker had known that the *all* was the case, and they thought it would have been informative and relevant, they would have said so"). Accounts of the linguistic environments under which scalar implicatures arise have been developed and formalized by Horn (1989), Hirschberg (1991), Gazdar (1979), Chierchia (2004), Levinson (2000), Sauerland (2004), and Sperber and Wilson (1986/1995), amongst others, but the psycholinguistic processing of these implicatures has been investigated only recently and in comparatively few studies (Bott & Noveck, 2004; Breheny, Katsos, & Williams, 2006; Huang & Snedeker, 2009; Noveck & Posada, 2003). In this article, we build on processing studies conducted by Bott and Noveck (2004, henceforth B&N) to further understand how people compute scalar implicatures.

Processing studies of scalar implicatures have focused on distinguishing between two theories of how implicatures are computed, a *default* theory (inspired by Levinson, 2000, and Chierchia, 2004) and a *contextual* theory (inspired by Sperber & Wilson, 1986/1995). The precise

* Corresponding author. Address: School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff CF10 3AT, United Kingdom.
E-mail address: BottLA@cardiff.ac.uk (L. Bott).

0749-596X
doi:10.1016/j.jml.2011.09.005

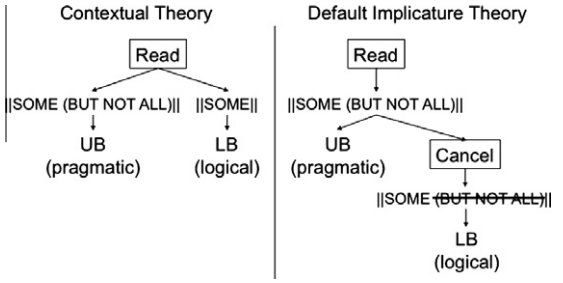

Fig. 1. The default implicature theory assumes an extra processing stage, the cancelling of the implicature, relative to the contextual account. UB and LB refer to upper- and lower-bound interpretations respectively.

instantiation of these two accounts differs across authors but B&N articulate the theories as follows. According to the default account, the implicature arises automatically and on all occasions, consistent with Neo-Gricean accounts of implicatures (Chierchia, 2004; Levinson, 2000). On hearing *some of the children are in the classroom*, for example, the first interpretation that is generated includes the upper-bound meaning,[1] *some* [but not all]. If the implicature is cancelled or defeased, either by contextual factors or by explicit statement, to arrive at the lower-bound meaning, *some* [and possibly all], the processor must first pass through a stage in which the implicature interpretation is considered and rejected. So, according to the default theory, the lower-bound meaning involves a two-stage derivation, in contrast to the one-stage derivation of the upper-bound meaning, as illustrated in the right panel of Fig. 1. According to the contextual theory, however, the implicature is not automatically incorporated into sentence representation, but depends on the contextual situation (Carston, 1998; Sperber & Wilson, 1995), as in the left panel of Fig. 1. Under this account, the processor does not necessarily have to consider the pragmatic, upper-bound interpretation before generating the literal, lower-bound meaning. Instead, at least under some circumstances, the literal meaning can be generated directly. As B&N noted, these theories make testable predictions about the time course of implicature and literal interpretations. If the default theory is correct, more processing time should be required for the lower-bound interpretation than the upper-bound interpretation, and the lower-bound interpretation should never be quicker than the upper-bound interpretation. In contrast, the contextual account does not predict that more time should be required to interpret a lower-bound meaning because it does not require that the implicature be considered and rejected before the lower-bound meaning is computed.

[1] We refer to scalar sentences as having *upper-bound* meanings when they have *some but not all* interpretations, and *lower-bound* meanings when they have the literal meaning, or *some and possibly all* interpretations, consistent with Breheny et al. (2006). The terminology refers to the scale having a bounded meaning at the upper end of the semantic scale in the implicature case (something less than *all*). When we refer to experimental conditions, however, we use the *logical* and *pragmatic* terminology used by B&N, with *logical* referring to the *some and possibly all* interpretation ($\exists x$) and *pragmatic* as the *some but not all* interpretation. This is because of the similarity between our experiments and those of B&N.

This hypothesis has been tested using a number of techniques, including reaction time studies (B&N), self-paced reading (Breheny et al., 2006) and visual world paradigms (Huang & Snedeker, 2009). These initial studies seem to have found evidence against a default view of implicature generation. For example, B&N measured reaction times in a sentence verification task involving sentences like *Some elephants are mammals*. The pragmatic, upper-bound interpretation (*some* [but not all]) required more time than the logical interpretation (*some* [and possibly all]), whether participants were explicitly instructed to interpret *some* logically or pragmatically, or whether participants made their own interpretations. This is opposite to the pattern predicted by the default theory. Similarly, Breheny et al. found that reading times on scalar quantifiers were longer in contexts for which an upper-bound meaning was likely compared to contexts in which it was not, and Huang and Snedeker found that eye movements to referent targets were comparatively slow to upper-bound *some* compared to a quantifier without an implicature (e.g., *all*).

Nonetheless, questions remain regarding the generality of these findings and how participants compute implicatures under different conditions. First, more recent investigations have suggested alternative explanations for the apparent delayed processing of upper-bound interpretations. Hartshorne and Snedeker (submitted for publication) argued that longer reading times in the upper-bound contexts used by Breheny et al. (2006) was due to a repeated noun penalty on the quantifier and not to deriving the implicature. When Hartshorne and Snedeker repeated the experiments without the confounding context they failed to observe any effects on the quantifier. Similarly, Grodner, Klein, Carbary, and Tanenhaus (2010) suggested that the delayed referential resolution for upper-bound *some* compared to *all* in Huang and Snedeker (2009) was attributable to the salience of more apt amount descriptors (*two* and *three*) in filler items and to the use of the partitive form, *some of*, as the implicature trigger. No delays were observed when numerals were omitted from the study design and when the point of disambiguation was equated. Finally, Feeney, Scafton, Duckworth, and Handley (2004) failed to find longer response times to pragmatic interpretations of underinformative sentences similar to those used by B&N. They suggested that the differences might be because of language differences across studies (B&N was conducted in French whereas Feeney et al. was conducted in English). Different findings across researchers suggest further work is required to establish why implicatures are fast in some situations and not others.

Second, it is not clear how to interpret longer processing times to upper-bound sentences relative to lower-bound sentences. The suggestion in the literature (e.g., B&N, Breheny et al., 2006) is that longer processing times reflect delayed processing of implicatures relative to literal meanings. An alternative, however, is that longer processing times are a result of greater difficulty in understanding upper-bound sentences. If participants have difficulty integrating upper-bound sentences into existing knowledge or discourse structures, they may delay committing to an interpretation until they are more confident of their response; trading off speed for an improvement in accuracy.

Speed-accuracy trade-off effects have been found to explain longer processing times in several paradigms. For example, McElree and Nordlie (1999) demonstrated that observed differences in judgment times for metaphoric and literal sentences were likely due to the *probability* of retrieving accurate interpretations rather than the *time* needed to retrieve and process the information. Similarly, McElree (1993) demonstrated that reading time differences for a verb in a frequent vs. an infrequent syntactic environment were due to probabilities of correctly retrieving the appropriate entry for the verb from the lexicon and not due to processing costs associated with serially retrieving each entry for the verb in turn. Implicature costs caused by retrieval probabilities or speed-accuracy trade-offs could selectively inflate response times for upper bound interpretations, and mask the cancellation cost predicted by the default implicature theory for lower bound interpretations.

Finally, even if the costs associated with implicatures are not due to speed-accuracy trade-off issues, there are a number of reasons why verifying a scalar implicature might be more costly than a literal meaning. For example, in B&N, participants might have been slower to verify upper-bound sentences (*some* [but not all]) because the upper-bound sentences involve a negation (*not all*) and negation is difficult to process (Clark & Chase, 1972), or perhaps extra time was needed to identify the semantic scale (*some* < *many* < *all*) and compute implications (*not many*, *not all*) in upper-bound sentences but not lower-bound sentences. It is important to establish the contributions of these factors because they inform us about how pragmatic principles are instantiated into the processor.

In this article, we report three experiments that investigate scalar implicatures in sentences similar to those used by B&N, but using a paradigm that precludes trade-offs between speed and accuracy. In Experiment 1, we test the default implicature model using a response deadline speed-accuracy trade-off (SAT) procedure that separates speed from accuracy (Reed, 1976; and developed for psycholinguistics by McElree and colleagues, e.g., Martin & McElree, 2008; McElree, 1993; McElree & Griffith, 1995, 1998; McElree & Nordlie, 1999). In Experiments 2 and 3, we consider an alternative hypothesis to the default model described above and test whether generating an implicature carries a cost above that associated with the additional semantic complexity of upper-bound sentences. Experiment 2 compares an implicit upper-bound meaning of *some* with the explicit upper-bound meaning of *only some*. Experiment 3 compares an implicit lower-bound meaning of *some* with the explicit lower-bound meaning of *at least some*.

Overview of experiments

In all of the experiments presented in this article participants read categorical sentences (e.g., *All elephants are mammals*) and responded "true" or "false" according to whether the sentences were consistent with their general knowledge. Response time was controlled experimentally by requiring responses to be made immediately after an auditory response prompt, which occurred at one of several time lags following the presentation of a sentence. Each participant made responses over a large range of intervals so that we could measure the growth of response accuracy as a function of time, starting with chance responding at very short time lags and increasing up to asymptotic accuracy at lags of several seconds (see, e.g., McElree & Nordlie, 1999). The principle advantage of the SAT technique is that it measures the temporal characteristics of response accuracy separately from asymptotic accuracy, whereas they are confounded in conventional reaction time studies. Thus, in SAT studies predictions

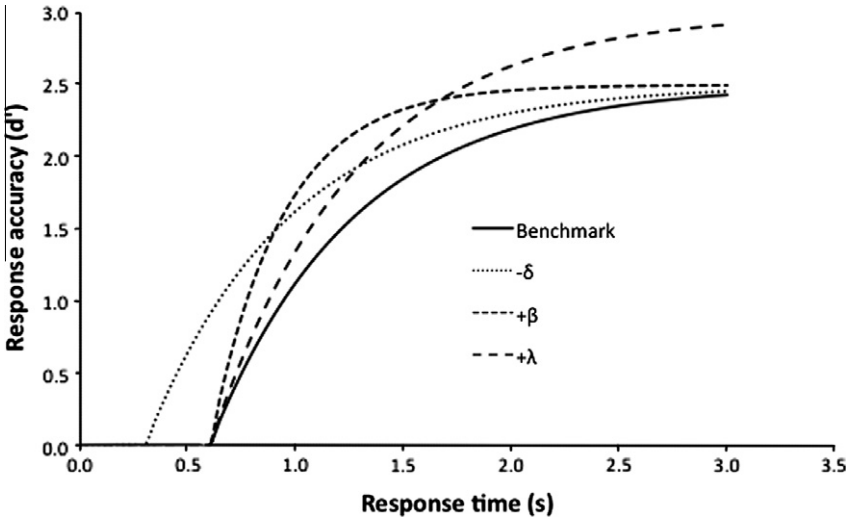

Fig. 2. Hypothetical curves illustrating the individual effects of changes to the three SAT parameters relative to a baseline response process (Benchmark). The SAT curve shows response accuracy as a function of response time. Smaller values of the intercept parameter correspond to earlier initiation of information retrieval and above-chance responses ($-\delta$). Larger values of the information accrual parameter correspond to steeper increases in response accuracy ($+\beta$). Larger values of the asymptotic accuracy parameter correspond to greater response accuracy at long response times ($+\lambda$).

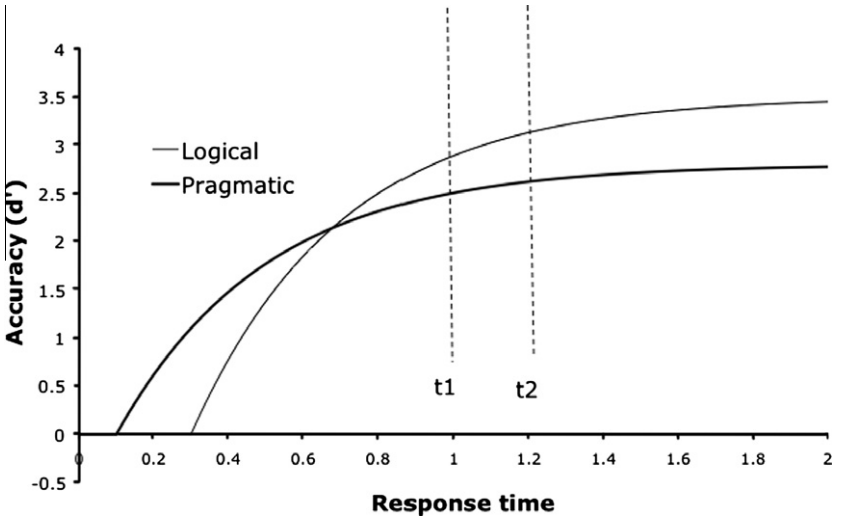

Fig. 3. Illustration of a hypothetical time course function for logical and pragmatic interpretations to underinformative sentences such as *some elephants are mammals*. A correct answer for participants who were responding logically would be "true" and a correct answer for participants responding pragmatically would be "false." The figure shows a situation in which logical interpretations have a higher asymptotic accuracy than pragmatic interpretations, yet the earlier intercept of the pragmatic interpretation is consistent with a default implicature account. Participants in B&N could have chosen to respond at t1 in the logical condition but delayed responding until t2 in the pragmatic condition in order obtain higher accuracy.

about processing time can be evaluated independently of retrieval probability.

If response accuracy is measured in terms of d' from signal detection theory, then accuracy can be modeled by Eq. (1) below. Accuracy rises as an exponential function of the response time, t, approaching asymptotic performance at long response times.

$$d' = \lambda(1 - e^{\beta(t-\delta)}), \quad \text{for } t > \delta, \text{ else } 0. \tag{1}$$

In addition to the asymptotic level of accuracy, λ, the SAT function is characterized by two additional parameters related to the speed of processing. The intercept, δ, identifies the earliest point at which accuracy departs from chance. The rate, β, determines the steepness of the accuracy curve and indexes the rate at which task-relevant information is accrued. Individual effects of λ, δ and β parameters are illustrated in Fig. 2. Detailed predictions regarding the implicature models and SAT dynamics are described below.

Experiment 1

B&N found that upper-bound interpretations of scalar sentences (*some* [but not all]) had longer sentence verification response times than lower-bound interpretations (*some* [and possibly all]). They interpreted this as evidence against a default view of scalar implicatures. An alternative explanation, however, is that there could be an inherent difference in the difficulty of verifying the truth of upper-bound sentences compared to lower-bound sentences, and that participants were choosing to spend longer on the more difficult upper-bound sentences. For example, in *some elephants are mammals*, participants might find it more difficult to seek and fail to find an elephant that was not a mammal, as in *some* [but not all], compared to merely finding overlap between elephants and mammals, as in *some* [and possibly all]. They would consequently allocate more time to verifying *some* [but not all]. In that case, even if upper-bound interpretations were generated quickly and by default, participants might spend longer assessing their truth compared to the truth of lower-bound interpretations, trading off speed to achieve an improvement in accuracy. This possibility is illustrated in Fig. 3. In the figure, the asymptotic accuracy differs across conditions so that when participants are given an unlimited amount of time to make their response, the pragmatic interpretations are more difficult to evaluate than the lower-bound interpretations. However, in spite of the lower asymptotic accuracy for pragmatic sentences, the intercept occurs earlier, that is response accuracy for pragmatic sentences departs from chance at an earlier point in processing than for lower-bound sentences. A pair of time-course functions like those in Fig. 3 is consistent with the default theory because the pragmatic interpretation is generated before the lower-bound one. The time-course functions are also consistent with B&N if participants delayed responding to the upper-bound interpretations. The first experiment in this paper tests whether the time course functions of scalar implicatures display earlier speed dynamics for upper-bound interpretations, as in Fig. 3.

Table 1
Example stimuli for Experiment 1.

Sentence type	Example	Correct response
S1	Some elephants are mammals	T or F
S2	Some elephants are insects	F
S3	Some elephants are Indian	T
S4	Some mammals are elephants	T
S5	All elephants are insects	F
S6	All elephants are mammals	T

Participants judged similar types of sentences to B&N. The six types of sentences are shown in Table 1. We used a variety of exemplars, subcategories and super-categories. The critical sentences, such as *some elephants are mammals*, were true under a lower-bound interpretation but false under an upper-bound interpretation. For example, *some* [and possibly all] *elephants are mammals* is true, but *some* [but not all] *elephants are mammals* is false. Four of the remaining five control sentences followed the same format as B&N (sentences 2, 4, 5, 6 in Table 1) but we added a new type of control sentence, S3, with a subcategory as the final word, such as *some elephants are Indian*. Without this type of sentence, participants interpreting the critical sentence type under an upper-bound (i.e., with a false response) would discover that all sentences of the form *some* [exemplar] *are X* was false, and therefore be able to anticipate false responses to the experimental sentence. Although including sentence type S3 resulted in a majority of the experimental sentences to be true, the use of d' as our measure of accuracy discounted any response bias that this may have caused.

Adopting an approach similar to B&N (Experiment 1), participants were biased through instructions and practice to be either logical or pragmatic, that is, to interpret underinformative sentences as being either lower-bound or upper-bound (see Rips, 1975). Instructions said that sentences like *some elephants are mammals* were true/false. Participants then completed a training session in which they verified sentences and received correct/incorrect feedback on their responses, although they were given unlimited time to respond (i.e., the deadline procedure was not used). Thus, participants in the *pragmatic* condition received feedback indicating that the correct response to experimental sentences like *some elephants are mammals* was false (upper-bound), whereas participants in the *logical* condition received feedback indicating that the correct response was true (lower-bound).

We analyzed correct responses to underinformative sentences using model fitting procedures described by McElree and colleagues (e.g., Martin & McElree, 2008; McElree, 1993; McElree & Griffith, 1995, 1998; McElree & Nordlie, 1999). Time course functions like those shown in Fig. 2 were fitted to individual participants and to group data and we compared parameter values across conditions. Consider what the SAT functions might look like for upper and lower-bound interpretations. We assume that the default implicature view predicts that there are at least two separate processes involved in understanding quantifier sentences, one associated with the upper-bound meaning, *P*, and one associated with the lower-bound meaning, *R*. Process *P* computes the meaning of all the expressions in the sentence and generates the associated implicatures. *P* is always triggered regardless of whether the implicature is (later) cancelled. Process *R* involves rejecting the result of *P* (i.e., cancelling the implicature) and computing the meaning of the sentence without the implicature. Deriving the upper-bound interpretation therefore involves executing only a single stage, *P*, whereas deriving the lower-bound interpretation involves executing multiple processing stages, $P + R$. In SAT terms, upper-bound interpretations should have earlier intercepts than lower-bound interpretations.[2] Fig. 3 illustrates a hypothetical set of response curves that are consistent with these predictions. The default view does not make predictions about the probability of successfully computing the upper-bound or lower-bound meanings and consequently, no predictions can be generated about asymptotic accuracy (although a speed-accuracy trade-off explanation of B&N would require lower asymptotic accuracy to be observed for upper-bound interpretations). Importantly, participants cannot trade speed for accuracy in this paradigm because they are forced to respond at the deadline.

We also examined the time course of incorrect responses using the pseudo-d' measure described in McElree (1998) and McElree & Dosher (1989). These miss rates can inform the interpretation of any differences that emerge across conditions in the SAT functions discussed above (see McElree, 1998). In particular, nonmonotonic changes across time are symptomatic of two-stage retrieval mechanisms, such as the default account shown in Fig. 1. For example, if the miss rate in the logical condition starts to increase before decreasing (a nonmonotonic function), this would suggest that participants initially consider the logical sentences to be false before eventually rejecting this interpretation and deriving the correct (true) response. The initial increase in false responding would correspond to the computation of the upper-bound meaning (process *P*, as described above) before cancellation of the implicature (process *R*).

Method

Participants

Thirty students from Cardiff University completed the experiment, and were paid for participation. One participant was removed because more than 50% of their responses were out-of-time (see the "Data treatment" section). There were consequently 14 participants in the logical condition and 15 participants in the pragmatic condition.

Materials

Test sentences were 240 sets each comprised of six related sentences. An additional 70 sets of sentences were used as practice sentences. Experimental (S1) sentences were of the form *Some* <items> *are* <category> (e.g., *Some elephants are mammals*). There were 26 different categories (e.g., fish, trees, cars) with varying numbers of items in each category (e.g., tuna, oaks, Ferraris). For each S1 sentence there were five corresponding foil sentences. S2 foil sentences were of the form *Some* <items> *are* <foil category> (e.g., *Some elephants are insects*). The items in the foil sentences were the same as in the S1 sentences, but the categories were randomly re-assigned to items belonging

[2] The difference could also be reflected in faster rates for the upper-bound interpretations. Whether differences are observed in the rate or intercept depends on assumptions about how much variability is inherent to each of the suggested processes (see McElree, 1993, for a detailed treatment of the statistics underlying serial hypotheses and McElree & Nordlie, 1999, for an accessible discussion about this point).

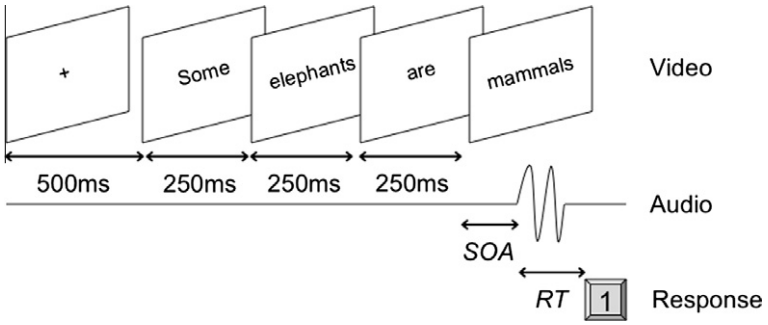

Fig. 4. Structure of an SAT trial, including fixation cross, series of words making up a categorical sentence, auditory response prompt, and participant key press. Stimulus onset asynchrony (SOA) was 27, 100, 200, 300, 400, 600, 800 or 2500 ms. Response time (RT) and key pressed ('1' or '3') were recorded.

to other categories. S3 foil sentences were of the form *Some* <items> *are* <subtype> (e.g., *Some elephants are Indian*). S4 foil sentences were of the form *Some* <category> *are* <items> (e.g., *Some mammals are elephants*). S5 foil sentences were of the form *All* <items> *are* <foil category> (*All elephants are insects*), using the same foil category as the corresponding S2 sentence. S6 foil sentences were of the form *All* <items> *are* <category> (*All elephants are mammals*).

Procedure

Trials for the speed-accuracy trade-off (SAT) procedure were modelled on McElree and Nordlie (1999), and were structured as shown in Fig. 4. Each trial began with a central fixation cross which was displayed for 500 ms, followed one at a time by the words making up the stimulus sentence, which were each displayed for 250 ms except that the final word remained on the screen until the participant responded. An auditory response prompt tone (1000 Hz, 50 ms) cued the participant's response at one of eight lag times following the onset of the sentence-final word. Participants pressed '1' or '3' to indicate whether the sentence was true or false, respectively. The experiment was programmed using DMDX (Forster & Forster, 2003).

Participants were randomly assigned to one of two conditions. Participants in the logical condition judged whether sentences were true under a lower-bound interpretation; participants in the pragmatic condition judged whether the same sentences were true under an upper-bound interpretation. Each participant completed two sessions of the assigned type, lasting for about 75 and 60 min, respectively.

Participants were instructed to read each sentence carefully, decide whether it was true or false, and respond quickly but accurately by pressing '1' on the number pad of a standard keyboard if true, or '3' if false. Instructions across the logical and pragmatic sessions were identical except for the final line, which read, "People sometimes have difficulty responding to sentences like *some elephants are mammals*. We would like you to say that these types of sentences are true [false] because elephants are indeed mammals [all elephants are mammals and not just some of them]." After receiving these instructions participants proceeded onto task training in which they were given feedback on whether their response was correct or incorrect.

A random selection of 78 of the 240 stimulus sets were used for task training, using one sentence from each set to obtain 39 sentences of type S1, and approximately equal numbers of each of the other sentence types (i.e., 7 or 8 of each). The task training sentences were used in a different random order for each participant. Task training trials omitted the auditory response prompt. Participants received feedback about the correctness of their responses to be sure they were reliably judging the lower or upper-bound meaning of the sentences as appropriate to the particular session, particularly for the experimental (S1) sentences which required different responses for logical vs. pragmatic interpretations. Feedback messages "Correct!" or "Incorrect!" were displayed for 1 s before the next trial began.

After the task training, participants were told that they would hear a beep during subsequent sentences indicating when responses should be made. The beeps occurred at one of eight lags following the onset of the final word in each sentence. The lags were as close as possible to 27, 100, 200, 300, 400, 600, 800 or 2500 ms, synchronized with a 75 Hz video display frame rate. Lags were chosen randomly for each trial, subject to counter-balancing to ensure equal frequencies across all lags for each sentence type. The instructions asked participants to respond "as soon as the beep sounds even if you are not yet sure." On all subsequent trials, participants received feedback about their response time on each trial, but not about the correctness of their responses. Responses within the target window of 100–300 ms after the onset of the beep elicited the feedback, "Good!" Responses earlier than 100 ms elicited "Wait for the beep!" and responses later than 300 ms elicited "Too slow". Feedback messages were displayed for 1 s before the next trial began.

In session 1 there were six blocks of 70 practice trials using the full SAT task with the auditory response prompt, as shown in Fig. 4. Each block used a random sentence from each set of practice sentences. Counter-balancing ensured that all six sentences in each set were used across blocks, with roughly equal numbers of each sentence type in each block. After blocks 2, 4 and 6, the video displayed the participant's percentage of on-time responses from the preceding 140 trials. Participants with more than 75%

on-time responses were told "Well done! Try to keep this up by responding as soon as the beep sounds even if you are not yet sure." Participants with a lower percentage of on-time responses were told "Try to improve on this by responding as soon as the beep sounds even if you are not yet sure." All participants were then offered a brief rest break and pressed a key to continue. In session 2 there was a single block of 70 practice trials, using a random sentence from each set of practice sentences. The video then displayed the percentage of on-time responses and offered a rest break, as above.

After the practice blocks, there were six blocks of 240 test trials. Each block used one sentence from each set of test sentences, chosen and ordered randomly for each participant. Counter-balancing ensured that all six sentences in each set were used across blocks, with equal numbers of each sentence type in each block. In case task performance suffered briefly after a rest break, each test block began with 12 filler trials using sentences from the practice set, before proceeding seamlessly into the 240 test trials. After each block the video displayed the participant's percentage of on-time responses for that block and offered a rest break.

Results

Data treatment

Participants occasionally neglected to perform both the verification task and the deadline task at the same time; instead concentrating on one or other. Failure to perform the verification task resulted in low d' scores; failure to conform to response deadlines resulted in responses outside the target response window. Throughout the experiments reported in this article we removed participants who obtained a d' of less than 1.0 on the longest time lag. We also removed participants who made less than 50% responses within the response window (i.e., 300 ms after the deadline signal).

Computing d'

We computed d' as the difference between the z-score of proportion hits and the z-score of proportion false alarms, where hits were correct responses to the experimental S1 sentences and false alarms were incorrect responses to S5 and S6 sentences, as described below. Proportion hits were calculated as the total number of correct S1 trials plus 0.5, divided by the total number of S1 trials plus 1, ignoring responses outside the target response window. Proportion false alarms were calculated in a similar way. This adjustment ensured that z-scores were always finite.

When participants were instructed to derive the lower-bound meaning (the logical condition), true responses to the experimental sentences were counted as hits, whereas when participants were instructed to derive the upper-bound meaning (the pragmatic condition), false responses were hits. Throughout the article we report analyses using control sentences beginning with *all* as the false alarm sentences. We reasoned that *all* sentences are the most theory-neutral measure of response bias because they do not involve an expression that would generate a scalar implicature. In any event, we found no qualitative differences in results when we used S2 and S4 *some* sentences as false alarm controls.

For Experiment 1, we report analyses based on two types of false alarms. First, we used a discriminant false alarm measure that was designed to take account of biases towards true responding (or against false responding). Because the correct logical and pragmatic responses were in opposition (true and false respectively), we had to use different false alarm sentences across different conditions. For the logical condition, true responses to the experimental sentences were hits and true responses to S5 sentences, e.g., *All elephants are insects*, were false alarms. In the pragmatic condition, false responses to the experimental sentences were hits and false responses to S6 sentences, e.g., *All elephants are mammals*, were (nominal) false alarms.

Second, we used a joint false alarm measure that was computed in the same way for both groups of participants. The joint false alarm analysis guarded against the possibility that effects observed using the discriminant false alarm above could be an artifact of using different false alarms in the two experimental conditions. The joint false alarm rate was based on total true (incorrect) responses to S5 sentences and false (incorrect) responses to S6 sentences.

Pseudo d'

We also analyzed misses (incorrect responses to the experimental sentences), using a form of the pseudo-d' prime measure described in McElree (1998) and McElree & Dosher (1989). The aim of this analysis was to compare incorrect responses to experimental sentences with incorrect responses to control sentences across the various response lags. Unambiguous foils like S5 or S6 sentences are assumed to involve a single-stage processing mechanism that produces a monotonic decrease in incorrect responding at longer lag times. If the experimental sentences in a particular condition also involve a single process mechanism, it might be faster or slower than in S5 or S6 sentences, but the difference should be monotonic. In contrast, if the experimental sentences involve a two-process mechanism, such as the default implicature theory shown in Fig. 1, then miss rates might initially increase during the first process, only to eventually subside in favor of correct responses (see McElree, 1998). Pseudo-d' was the difference between the z-score of proportion misses and the z-score of the joint false alarm rate described above.

Analysis

We conducted model-based analyses on the averaged data and the individual participant data. For the averaged analysis, we combined the data from each group of participants to form a single set of logical data and single set of pragmatic data. We then minimized the squared error between the averaged data and the SAT function, and conducted a hierarchical model fitting analysis. We used the lag plus the averaged latency per deadline signal as the time coordinate, as we did throughout the experiments presented in this paper. For the individual participant analysis, we optimized SAT functions for each participant separately. We then compared parameter values between the logical and the pragmatic conditions using nonparametric

inferential statistics (we used nonparametric statistics to avoid the distorting influence of outlying parameter values). In both cases, our primary goal was to determine whether there were differences in the processing dynamics across interpretations, that is, the rate of information retrieval (β parameters) or the intercept (δ parameters), or whether the differences observed in B&N could be due to asymptotic accuracy differences (λ parameters). Faster rates or earlier intercepts in the logical condition would be evidence against a default implicature account. Differences in only asymptotic accuracy would mean that the differences observed by B&N could be explained by speed-accuracy-trade-off effects and would therefore not discriminate between the default implicature and contextual theories.

The raw accuracy rates for all three experiments are shown in Appendix A. The average d' data is shown in Fig. 5 together with the fully parameterized model (parameter values are shown in the legend). The fully parameterized SAT model for the dual logical and pragmatic conditions includes independent parameters for each task, so we designated this 2λ–2β–2δ model to indicate its six parameters. Of primary interest, the intercept was earlier in the logical than the pragmatic condition, and the retrieval rate was higher for the logical condition than the pragmatic condition. Also, the asymptotic accuracy was somewhat higher in the logical than the pragmatic condition, but the statistical analyses below focus on comparisons relevant to the hypothesis of interest.

We compared the fit of the full 2λ–2β–2δ model against a four-parameter 2λ–1β–1δ model, in which the time course β and δ parameters were assumed to be the same for both logical and pragmatic conditions. The six-parameter 2λ–2β–2δ model resulted in a significantly lower residual error than the two λ restricted model, 2λ–1β–1δ, $\chi^2(2) = 22.35$, $p < .001$. Thus, after differences in asymptotic accuracy across conditions have been factored out, allowing the rate and intercept to vary across conditions resulted in a significantly better fitting model. Specifically, an earlier intercept in the logical condition, 320 ms (logical) vs. 390 ms (pragmatic) and a faster rate, $1/\beta$(logical) = 268 ms vs. $1/\beta$(pragmatic) = 602 ms, accounted for significantly greater variability than assuming rate and intercept were constant across conditions. Faster dynamics for the logical condition mean that the results of B&N could not be explained by a speed-accuracy trade-off.

Analysis of five-parameter models revealed that the 2λ–2β–1δ model significantly reduced residual error when compared to the 2λ–1β–1δ model, $\chi^2(1) = 19.81$, $p < .001$, such that β was higher in the logical condition than the pragmatic condition. Similarly, allowing δ to vary across conditions resulted in a significantly earlier intercept in the logical condition, $\chi^2(1) = 14.76$, $p < .001$. Both of the five-parameter model comparisons therefore provide evidence against the default model.

In addition to analyzing the averaged data, we fitted three parameter models to each individual and compared parameter values across conditions. Mann–Whitney tests with $N = 29$ revealed that the intercept was significantly earlier in the logical condition compared to the pragmatic condition, $m = 0.33$ vs. $m = 0.47$, $U = 42$, $p = 0.006$, and the rate was marginally faster in the logical condition compared to the pragmatic condition, $m = 4.76$ vs. $m = 2.24$, $U = 65.5$, $p = .085$. Confirming the conclusions of the averaged data, the individual participant patterns are inconsistent with a default implicature model: Participants had earlier intercepts in the logical condition.

The results using the joint false alarm measure were qualitatively similar to the analysis using discriminant false alarms. For the averaged data, the six-parameter 2λ–2β–2δ model resulted in a significantly lower residual error than the 2λ restricted model, 2λ–1δ–1β, $\chi^2(2) = 24.49$, $p < .001$, and comparisons of the five parameter models against the four parameter models produced significantly earlier intercepts and faster rates for the logical condition, $\chi^2(1)$'s > 16.66, p's < .001. For the individual participant analysis, intercepts were significantly earlier in the logical condition, $m = 0.18$ vs. $m = 0.39$, $U = 26$, $p < 0.001$, but the rate was not, $m = 2.81$ vs. $m = 2.28$, $U = 94$, $p = 0.65$. The analysis using the joint false alarm measure

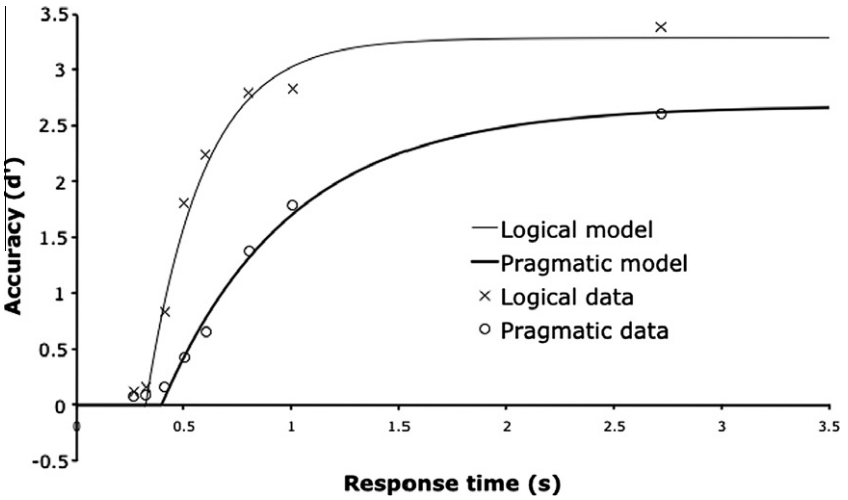

Fig. 5. Averaged data from Experiment 1 together with the fully parameterized model, 2λ–2β–2δ, λ(logical) = 3.29, β(logical) = 3.73, δ(logical) = 0.32, λ(pragmatic) = 2.68, β(pragmatic) = 1.66, δ(pragmatic) = 0.39.

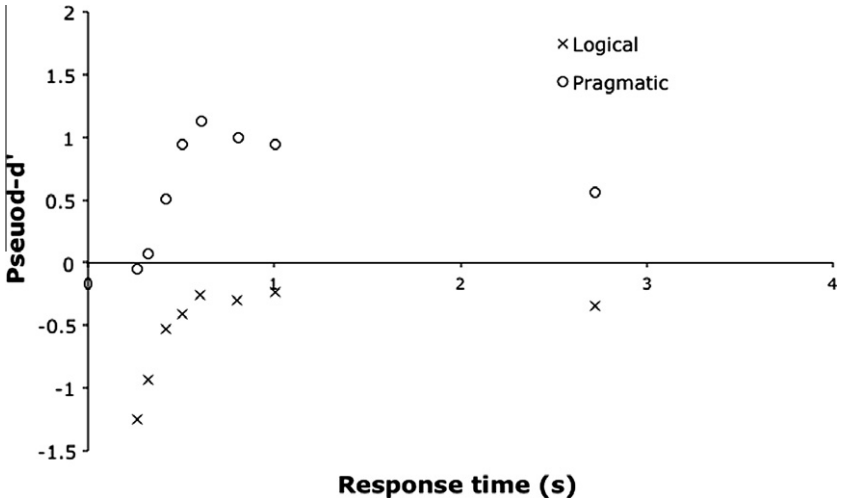

Fig. 6. Average pseudo-d' rates for Experiment 1. A nonmonotonic pattern is shown for pragmatic responses but not for logical responses.

confirms that the earlier intercepts were observe for the logical condition cannot be because of differential performance across conditions on the false alarm sentences.

Finally, we analysed the miss rates across conditions, that is, incorrect response rates to experimental sentences relative to S5 and S6 foils. Fig. 6 shows pseudo-d' scores as a function of time. For logical participants, pseudo-d' was initially quite negative (because of a response bias towards true responding) and then increased monotonically towards zero at longer response lags. For pragmatic participants, however, pseudo-d' increased at first to a maximum at around 700 ms before falling back towards zero. The pseudo-d' appears to be nonmonotonic for pragmatic participants but not for logical participants. While an overall progression towards zero would be a necessary component of an increase in standard d' over time (as in Fig. 5), a progression away from zero (i.e., an increase in the error rate on experimental sentences) would be evidence of a two-process interpretation mechanism. We therefore report analyses that test whether there was a significant progression away from zero as time increases. We compared the difference between pseudo-d' at each time point and each later time point. For pragmatic participants, a difference occurred between the first and the fifth time points $t(14) = 9.45$, $p < .0001$, where there was an increase in pseudo-d' away from zero, but for logical participants there was no significant progression away from zero for any pair of time points, all t's < 1. In summary, analysis of the miss rates revealed the reverse pattern to the predictions of a two-stage default processing account (as shown in the right panel of Fig. 1): the miss rate of participants in the pragmatic condition increased relative to errors in the control conditions before finally decreasing, but there is no similar pattern for participants in the logical condition.

Discussion

Experiment 1 tested whether participants interpreted upper-bound sentences faster than lower-bound sentences when they were not able to trade off speed for accuracy. The default implicature account predicts that participants should respond faster to the upper-bound sentences because deriving the lower-bound interpretation should involve cancelling an implicature. In contrast, we found faster rates of retrieval and earlier departures from chance in the logical condition than the pragmatic condition.

More generally, our results are qualitatively consistent with those of B&N, who found longer response times to correct upper-bound interpretations. However, whereas that pattern of reaction times could be explained by the default implicature theory under the assumption that upper-bound meanings are harder to verify than lower-bound meanings, the earlier and faster processing we observed in the SAT task for lower-bound interpretations is not. In B&N, participants could have been delaying their response to the upper-bound sentences in order to maximize accuracy, whereas in our study they were prevented from using such a strategy by the deadline procedure.

We also note that our experiment widens the range of contexts in which an implicature is costly by demonstrating that the effect is robust across different control sentences (B&N did not use sentences with subtypes, i.e., S3 in our design), a different language (B&N conducted their experiments in French whereas our experiment was in English), a different treatment of response options (we analysed the data with d' as the dependent measure), and a different type of implicature instruction (we primarily relied on correct/incorrect feedback to encourage participants to compute interpretations whereas B&N used a more detailed explanation of the different meanings of *some*).

Our goal in Experiment 1 was to test the standard default implicature model identified in B&N. In the remainder of the article we consider why the upper-bound interpretations were slower to verify than the lower-bound interpretations. There are many potential explanations for this effect but we break these possibilities down into two groups. First, the underlying form of the upper-bound interpretation is more complex and this complexity might entail more lengthy processing. For example, when a

sentence involves a scalar implicature the subject set is broken down into a reference and a complement set, for example, *some* [but not all] *elephants are mammals* requires the hearer to derive a set of elephants that are claimed to be mammals and a set that are not. In contrast, a sentence without a scalar implicature only requires a reference set. Dividing the subject set may take processing time (see Grodner, Gibson, & Watson, 2005, for evidence of additional processing cost when a reader must instantiate a complement set in addition to a reference set). Similarly, the memory search necessary to verify upper-bound sentences may be disproportionally more complex than the memory search necessary for the lower-bound interpretation. For example, *some* [but not all] *elephants are mammals* involves testing whether there are elephants that are mammals and also whether there are elephants that are not mammals, whereas only the former procedure is required for *some* [and possibly all] *elephants are mammals*. If the procedure that verifies these sentences is at least partially serial, pragmatic sentences may take longer to verify because of their propositional complexity. Second, instead of sentence complexity causing the delay, there could be extra work for the processor due to inferencing *per se*. That is, there could be additional work required to go from the linguistic input to the speaker's intended meaning. This could arise because there is a heavier dependence on context, because alternative expressions might need be considered, because the literal interpretation might be computed and rejected, because more propositions must be computed inferentially rather than explicitly decoded, or because of some other aspect of the inferential process.

Understanding which of these possibilities caused the delay is important because it can narrow down the range of theories that explain how implicatures are generated. In particular, we consider an alternative version of the default implicature theory to that tested above. According to this model, schematized in Fig. 7, the implicature and cancelling process take place prior to verification of the sentence, as opposed to the standard default model considered above which assumes cancellation occurs after, or during, verification. The inference could be automatically derived and defeased independently of the processes that verify the propositional content of the sentence. For example, the *not all* inference in *some elephants are mammals* would always be derived but, crucially, cancellation would occur independently of the processes that verify whether elephants are mammals. This account predicts that deriving a lower-bound interpretation requires more computation than deriving an upper-bound interpretation, just as in the standard (verification) default model, but that this cost could by obscured by the potentially greater cost of verifying the more complex semantic structure of the upper-bound sentences. This situation is illustrated in Fig. 7. In Experiment 2, we test this account by determining whether there is a cost to generating an implicature over and above the cost of verifying the upper-bound interpretation.

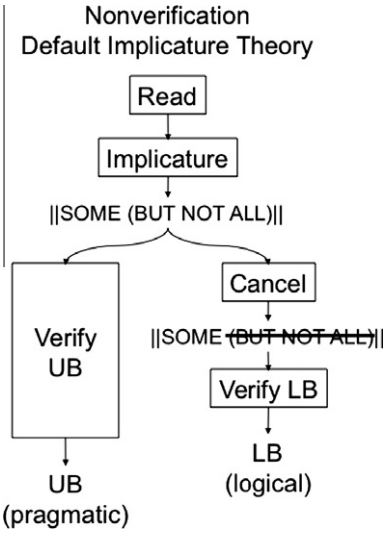

Fig. 7. Deriving lower-bound interpretation requires cancelling the implicature but this occurs at a stage prior to verification. Longer processing times could occur for the upper-bound interpretation because the longer verification time could obscure the cost of cancellation.

Experiment 2

In Experiment 1 we investigated implicature processing by comparing upper and lower-bound interpretations of the same sentences. A potential limitation of this comparison, however, was that differences in semantic complexity might have obscured the workings of a nonverificational default implicature model. In Experiment 2 we compared implicature sentences against sentences that had the same meaning but which were explicit and did not involve implicature. We compared the upper-bound *some* sentences used in Experiment 1, as in *some* [but not all] *elephants are mammals* (the pragmatic-*some* condition), with explicit literal equivalents generated using the *only* operator, as in *only some elephants are mammals* (the *only-some* condition; see Breheny et al., 2006, for a similar use of *only*). Both sentence forms generate an upper-bound interpretation and so both sentences result in equally complex sentences, but it is only the bare *some* form in which the reader has a choice about whether to derive an upper-bound interpretation. This is illustrated by the defeasibility of the bare *some* form, for example, *some elephants are mammals; in fact all of them are*, compared to the non-defeasibility of the *only* form, (∗) *only some elephants are mammals, in fact all of them are*. A comparison between the two forms can therefore provide a direct test of whether there is an inferential cost of deriving the implicature independently of semantic complexity.

According to the default model, the upper-bound interpretation is obligatorily triggered by the scalar term so that whenever *some* is read, the implicature would always be derived. This should apply whether *some* is embedded under *only* or whether it is read as a bare quantifier. Because the upper-bound interpretation is derived in both cases, and because a lower-bound interpretation is not considered in either case, the most straightforward prediction from such a model would be that no rate or intercept differences will be observed between the pragmatic and *only some* conditions. It is possible, however, that there are

Table 2
Example stimuli for Experiment 2.

Sentence type	Example	Correct response
S1	Some elephants are mammals	F
S2	Only some elephants are mammals	F
S3	All elephants are mammals	T
S4	Some elephants are Indian	T
S5	Only some elephants are Indian	T
S6	All elephants are Indian	F

extra meaning components associated with the use of *only* that need to be incorporated into the sentence representation. For example, that the upper-bound interpretation cannot be cancelled or that the complement set is a particularly important part of what is communicated. If so, there would be processing delays associated with the *only* condition, i.e., later intercepts or slower rates (alternatively, extra processing costs associated with *only* might be incorporated into the sentence representation early on in processing and might not be observable in the response window at the end of the sentence). A default implicature model therefore predicts either no differences between pragmatic-*some* and *only some* or that *only some* should involve slower processing than pragmatic-*some*.

Default accounts do not make predictions about asymptotic accuracy in pragmatic-*some* compared to *only some* sentences. The asymptotic accuracy for pragmatic-*some* is sensitive to the consistency with which participants choose the upper-bound interpretation (amongst other things). If they erroneously fail to derive the implicature (or if they cancel it) asymptotic accuracy will be lower. Because the upper-bound interpretation is semantically forced in the explicit *only some* condition but merely encouraged in the pragmatic-*some* condition, lower accuracy might be expected in the pragmatic-*some* condition.

Participants classified categorical sentences as true or false using a deadline procedure, just as they did in Experiment 1. The different sentence types are shown in Table 2. The crucial comparison is between sentence types S1, the pragmatic-*some* sentences, and S2, the *only some* sentences. All participants were given pragmatic-*some* training, just as they were in the pragmatic condition of Experiment 1, so S1 and S2 sentence types were both false.

Method

Participants

Twenty-two Cardiff University students participated for payment. Three participants were removed because their asymptotic accuracy was less than 1.0 d'. One participant was removed because more than 50% of their responses were out-of-time.

Materials

Test sentences were 240 sets each comprised of six related sentences, designed so that participants could not predict the correct answer prior to the onset of the final word. Experimental sentences began with 'Some' or 'Only some' followed by an item and its supervening category (e.g., *Some elephants are mammals; Only some elephants are mammals*). The categories and items were similar to those used in Experiment 1. There were four additional foil sentences in each set of related sentences. S3 foil sentences were of the form *All* <items> *are* <category> (*All elephants are mammals*). S4 foil sentences were of the form *Some* <items> *are* <subtype> (e.g., *Some elephants are Indian*). S5 and S6 foil sentences were the same as S3 and S4, respectively, but began with 'Only some' or 'All' (e.g., *Only some elephants are Indian*; *All elephants are Indian*).

An additional 70 sets of sentences were used as practice sentences.

Procedure

The procedure was the same as in Experiment 1, except that since the shortest lag (27 ms) seemed to be superfluous it was dropped and a lag of 500 ms was included instead. Participants were instructed to interpret *some* pragmatically. *Only some* was presented as a unit in a single presentation window of 250 ms.

Results

Hits to the pragmatic-*some* and the explicit *only some* conditions were calculated as false (correct) responses to the pragmatic-*some* and the explicit sentences respectively. False alarms were false (incorrect) responses to S3 sentences, such as *all elephants are mammals*, for both conditions. We used the same false alarm in both conditions to avoid the possibility that observed effects might be due to differences in performance on the false alarm sentences. S3 sentences were chosen because the subject and the predicate were identical to the experimental sentences (S1 and S2). However, as in Experiment 1, the effects replicate whichever false alarm sentence is chosen (S3, S4, or S5). d' was then calculated as in Experiment 1.

We analysed the data by optimizing parameters to individual participants' data and to the average data, just as we did in Experiment 1, and proceeded with hierarchical model fitting. Fig. 8 shows the averaged data together with the fitted model. Accurate performance in the *only* condition appears to start earlier but performance equates as more time is available. We compared nested models to test for differences between conditions. The fully parameterized model, 2λ–2β–2δ, resulted in significantly lower residual error than the four-parameter 2λ restricted model, 2λ–1β–1δ, $\chi^2(2) = 13.82$, $p < .001$. indicating that there were speed differences across conditions independently of asymptotic accuracy.

We fit five-parameter models (2λ–2β–1δ and 2λ–1β–2δ) to examine simple and unique effects of the intercept and rate parameters. An earlier intercept in the *only* condition significantly reduced error compared to the restricted 2λ–1β–1δ model, $\chi^2(1) = 13.18$, $p < .001$, and a higher β parameter marginally so, $\chi^2(1) = 3.56$, $p = .059$. Further comparisons demonstrated that varying the intercept parameter across conditions additionally reduced error compared to varying β, 2λ–2β–2δ vs. 2λ–2β–1δ model, $\chi^2(1) = 10.27$, $p < .001$, but the reverse was not true, 2λ–2β–2δ vs. 2λ–1β–2δ model, $\chi^2 < 1$, suggesting that the intercept was primarily responsible for differences across

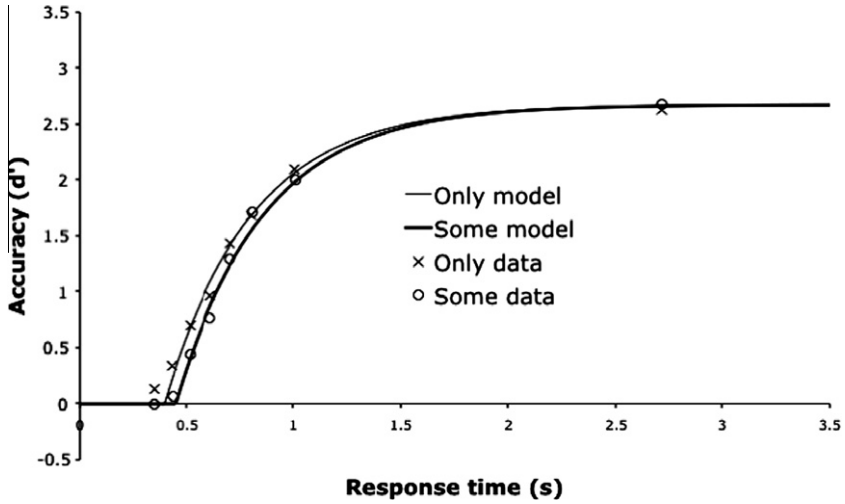

Fig. 8. Averaged data for Experiment 2 shown together with the optimum model, 1λ–1β–2δ, in which $\lambda = 2.67$, $\beta = 2.42$, $\delta(\text{only}) = 0.39$, and $\delta(\text{some}) = 0.45$.

conditions. Indeed, the four-parameter model with an earlier intercept in the *only* condition, 1λ–1β–2δ, was sufficient to account for any significant variability across conditions, χ^2's < 1. In summary, the average data is better fit with a model that assumes participants respond above chance earlier in the *only-some* condition than in the pragmatic-*some* condition.

We also analyzed the data by optimizing parameters for each individual participant. When four-parameter models were fitted to each participant's data, rate parameters were significantly faster in the *only* condition, $m = 2.48$ vs. $m = 2.47$, $Z = 2.37$, $p = .018$, and intercept parameters were significantly earlier, $m = 0.42$ vs. $m = 0.48$, $Z = 2.98$, $p = .003$. For five-parameter models in which λ varied, i.e., 2λ–2β–1δ and 2λ–1β–2δ, the results were similar: rate parameters were significantly higher in the *only* condition, $m = 2.98$ vs. $m = 2.43$, $Z = 2.28$, $p = .022$, and δ parameters were significantly earlier, $m = 0.42$ vs. $m = 0.49$, $Z = 3.07$, $p = .002$. Finally, for the six-parameter model fits, δ parameters were significantly earlier, $m = 0.36$ vs. $m = 0.48$, $Z = 2.20$, $p = .028$, but there were no differences between β parameters. Results were therefore entirely consistent with the averaged data in demonstrating faster processing of the *only some* sentences.

Discussion

Experiment 2 compared interpretations of implicit upper-bound sentences, pragmatic-*some*, against explicit upper-bound sentences, *only some*. The sentences were equally complex but an implicature was only required in the implicit version. Our results were that correct responding was delayed in the pragmatic-*some* condition relative to the *only* condition, i.e., a later intercept was needed to accurately model the pragmatic-*some* condition. Interestingly, we did not observe significant asymptotic accuracy differences across conditions. It appears that participants were equally successful at verifying the upper-bound interpretation whether it was semantically forced or whether it was derived using implicature procedures.

In the introduction we suggested that there were two types of cost that could have contributed to the delayed upper-bound interpretations seen in Experiment 1. The first referred to semantic complexity differences across interpretations, such as the extra memory search necessary to verify the upper-bound interpretation, and the second to the inferential process of deriving the implicature, such as assessing speaker's intentions. Our results indicate that at least some of the cost of interpreting pragmatic sentences in Experiment 1 was due to the extra time needed to derive an implicature, and not purely to differences in sentence complexity. Specifically, in Experiment 1, accurate responses in the pragmatic condition began 140 ms later than accurate responses in the logical condition (470 ms vs. 330 ms, respectively) and in Experiment 2, accurate responses in the pragmatic condition began 120 ms later than accurate responses in the *only-some* condition (480 ms vs. 360 ms).[3] The inferential component therefore makes a substantial contribution to the delay of implicit upper-bound interpretation relative to the lower-bound interpretation.[4]

The observed inferential cost is inconsistent with the nonverification default theory that we described in the introduction. We suggested that while the standard, verification default theory was unable to account for the results

[3] Estimates of the differences depend on the model chosen to represent the data in the experiments. We reasoned that the individual participant analysis reported in Experiment 1 was comparable to a 2λ–2β–2δ model because three parameters were optimized for the logical-*some* condition and three for the pragmatic-*some* condition. We therefore compared intercept differences from Experiment 1 with intercept differences from the 2λ–2β–2δ model in Experiment 2.

[4] There are also likely to be costs due to semantic complexity reflected in the rate parameters. Assuming the six parameter models used to derive the intercept differences, there was a $1/\beta$ delay of 237 ms for the pragmatic interpretations relative to the logical interpretations in Experiment 1 (447 ms vs. 210 ms) but a 70 ms advantage for the pragmatic *some* condition relative to the *only-some* condition in Experiment 2 (423 ms vs. 354 ms). Differences across rate parameter values must be treated with caution, however, because the individual participant analysis revealed only a marginally significant difference in Experiment 1, $p = .085$ and no difference in Experiment 2, $p = .18$.

of Experiment 1, a nonverification default model might be consistent with the data. This model assumed that cancellation of the lower-bound meaning occurred quickly but that complexity differences between lower and upper-bound sentences might have obscured the cancellation process. In this experiment we removed the differences in sentence complexity across conditions but there remained a delay in responding to implicit upper-bound interpretations. If an implicature is automatic and occurs on every occasion, as a default implicature theory predicts, there should be no additional cost associated with deriving upper-bound implicitly (pragmatic-*some*) compared to deriving the upper-bound explicitly (*only-some*), in contrast to our results.

The results of Experiments 1 and 2 suggest that the most straightforward versions of a default implicature theory cannot account for implicature processing. We now turn to other types of models that may explain our results.

Experiment 3

The contextual account considered in the introduction assumed only that implicatures were not obligatorily derived on reading a scalar term; sometimes the literal meaning could be derived directly. There are therefore potentially several versions of such a theory depending on how context determines the choice. One possible model is that scalar terms are systematically ambiguous between literal and pragmatic readings and the interpretive mechanism could probabilistically select between them (modulated by contextual constraints). There would be no costs associated with rejecting or cancelling the inappropriate readings of the scalar term; instead, context determines which is the appropriate reading and this reading is directly incorporated into the sentence representation. We have in mind something similar to the unrestricted race model proposed to account for resolving syntactic ambiguities (e.g., Traxler, Pickering, & Clifton, 1998; Van Gompel, Pickering, Pearson, & Liversedge, 2005). Delayed pragmatic interpretations in Experiment 1 would be explained by differences in sentence complexity across interpretations. However, a model that assumed this to be the only cost to deriving pragmatic-*some* interpretations could not explain the results of Experiment 2, in which complexity differences were equated across conditions. Additional mechanisms or alternative contextual accounts are needed to fully explain the data.

One possibility is a system that checks context each time a scalar term is read. For example, context might need to be assessed to determine whether the semantic environment was an implicature licensing context (e.g., that the environment was not negative, or downward entailing), or speaker's politeness intentions might need to be assessed (e.g., Bonnefon, Feeney, & Villejoubert, 2009), or whether it would be more informative to know that the stronger statement is true (Breheny et al., 2006). When the quantifier is preceded by *only*, as in Experiment 2, there would be no need to assess the context because the speaker had explicitly specified an upper-bound interpretation. This account therefore predicts that deriving pragmatic-*some* would be delayed relative to *only-some*. Sentence complexity differences could explain why pragmatic responses were delayed relative to logical responses in Experiment 1. One prediction that follows from such a context-checking mechanism is that the same context-checking cost should be incurred with logical-*some* as with pragmatic-*some*. If context needs to be checked to determine whether an implicature is appropriate, the cost should arise regardless of the outcome.

Table 3
Example stimuli for Experiment 3.

Sentence type	Example	Correct response
S1	Some elephants are mammals	T
S2	At least some elephants are mammals	T
S3	All elephants are mammals	T
S4	Some elephants are insects	F
S5	At least some elephants are insects	F
S6	All elephants are Indian	F

In Experiment 3 we tested this prediction by comparing lower-bound *some* (logical-*some*) with *at least some*. The effect of modifying *some* with *at least* is to semantically force a lower-bound interpretation, just as *only* semantically forces an upper-bound interpretation. Participants classified sentences as true or false under deadline conditions, just as they did in Experiments 1 and 2, and received feedback instructing them to interpret *some* to mean *some and possibly all*, as in the logical condition of Experiment 1. We predicted that if there is a general effect of context assessment we would observe later intercepts or slower rates of growth of the SAT curves for logical-*some* compared to *at least some*.

The six sentence types are shown in Table 3. The analysis compared S1 with S2 using a repeated measures design. Both S1 and S2 required *true* as the correct answer. To be consistent with Experiments 1 and 2, we report the d' analysis using the *all* sentence, S6, for the false alarms (although using S4 and S5 result in the same qualitative conclusions).

Method

Participants

Twenty-four Cardiff University students participated for payment. Two participants were removed because their asymptotic accuracy was less than 1.0 d' and two participants were removed because they claimed to have used an abstract strategy to perform the task, as we describe below.

Materials

Test sentences were 240 sets each comprised of six related sentences. The categories and items were the same as those used in Experiment 2. Experimental sentence types (S1 and S2) began with 'Some' or 'At least some' followed by an item and its supervening category (e.g., *Some elephants are mammals*; *At least some elephants are mammals*). There were four additional foil sentences in each set of related sentences. S3 foils were of the form

All <items> *are* <category> (e.g., *All elephants are mammals*). S4 foil sentences were of the form *Some* <items> *are* <foil category> (e.g., *Some elephants are insects*). S5 foil sentences were the same but began with 'At least some' (e.g., *At least some elephants are insects*). S6 foil sentences were of the form *All* <items> *are* <subtype> (*All elephants are Indian*).

An additional 70 sets of sentences were used as practice sentences.

Procedure

The procedure was the same as in Experiment 1, except that participants were instructed to interpret *some* with the lower-bound. *At least some* was presented in a single 250 ms presentation window, just as *only some* in Experiment 2.

Results

Two participants told us that when the final word in the sentence was a subordinate category (S6 sentences) they knew that the sentence was false without having to consider the rest of the sentence. While responses to other control sentences would not affect the results, S6 sentences form part of the d' calculation and unusual performance on these sentences would distort our conclusions. We therefore checked the data for these participants and found that the SAT function intercept was negative for these participants, suggesting that their strategy was effective. We therefore removed these participants. No other participants mentioned a similar strategy when queried or had a negative intercept, and even if there were a slight effect on S6 sentences, performance in both logical-*some* and *at least some* would be equally affected.

We analyzed the remaining participants' data in a similar way to Experiment 2. As in previous experiments, we compared the six-parameter model with the restricted four-parameter model, 2λ–1β–1δ. In this case however, there was no advantage of the more complex model, $\chi^2(2) = 2.00$, $p = 0.37$. Indeed, there was no model that had a significant reduction in error compared to the three-parameter model, 1λ–1β–1δ, χ's < 1.8, p's > 0.37. Performance was therefore equivalent across conditions. Fig. 9 shows the average data together with the best-fitting three parameter model.

Comparison of individual participant parameter values were generally consistent with average data but one model revealed earlier intercepts in the *at least* condition. For four-parameter models, neither the intercept nor the rates were significantly different across conditions, $Z = 1.76$, $p = .080$, and $Z = 1.20$, $p = .23$, but in the five-parameter models with asymptotic accuracy allowed to vary, the intercept was significantly earlier in the *at least* condition, $m = 0.27$ vs. $m = 0.29$, $Z = 2.40$, $p = .016$, and the rate was marginally higher, $m = 3.33$ vs. $m = 3.14$, $Z = 1.83$, $p = .067$. No significant differences were observed for the six-parameter model fits, $Z = 0.91$, $p = .36$, and $Z = 1.43$, $p = .15$. Because there were marginal differences between *at least some* and logical-*some* for the five-parameter models, we compared the five-parameter differences in Experiment 3 with those of Experiment 2 (while participants were not randomly allocated to experiments, they were nonetheless from the same population, Cardiff University students, and each experiment involved comparisons within subjects, so risk of biased sample effects are minimal). A greater difference between *only some* and pragmatic-*some* than between *at least some* and logical-*some* would argue against a general context assessment account and in favor of an implicature specific cost. For each participant we computed the difference between the parameter values for both conditions and compared the resulting differences using a Wilcoxon test. This revealed that the intercept differences in the 2λ–1β–2δ models were larger between *only some* and pragmatic-*some* than between *at least* and logical-*some*, $Z = 2.59$, $p = .01$, but rate differences in the 2λ–2β–1δ models were not, $Z < 1$.

As a further test of whether there were differences between the relative difficulty of processing *some*, we compared d' differences between *only some* and pragmatic-*some* and between *at least some* and logical-*some* using a

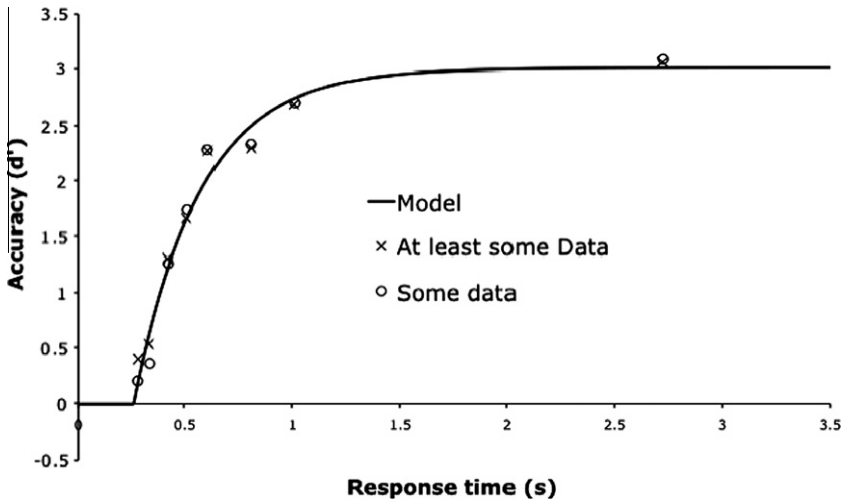

Fig. 9. Averaged data for Experiment 3 shown together with the optimum model, 1λ–1β–1δ, $\lambda = 3.02$, $\beta = 3.23$, $\delta = 0.26$.

repeated measures ANOVA with time lag (1–8) as a factor. We found that d' differences were larger between *only some* and pragmatic-*some* than between *at least some* and logical-*some*, $F(1,36) = 2.23$, $p = .028$, confirming the results of the model fitting analysis. We can thus conclude that even if there is a small delay in processing logical-*some* compared to *at least some* sentences, the delay in processing pragmatic-*some* compared to *only some* is significantly larger.

Discussion

Experiment 3 tested whether logical-*some* was slower to process than explicit lower-bound meanings (*at least some*) in the same way that pragmatic-*some* was slower than explicit upper-bound meanings (*only some*) in Experiment 2. While some of the analyses we conducted suggested the logical-*some* was delayed relative to *at least some*, many of the analyses failed to show a significant difference. We therefore cannot be confident that there is a general slowdown associated with *some*. More conclusively, comparison of the difference between the implicit and explicit interpretations across Experiments 2 and 3 revealed that the size of the difference in upper-bound interpretations of Experiment 2 was significantly larger than the lower-bound differences of Experiment 3, i.e., pragmatic-*some* was relatively more difficult to interpret than logical-*some*. This result is inconsistent with a contextual account that assumes that the only explanation for delayed pragmatic-*some* interpretations is a general cost in resolving the meaning of *some* combined with sentence complexity differences between upper and lower-bound sentences. Instead, there must be an additional cost associated with deriving the scalar implicature that does not occur when the lower-bound interpretation is derived. In the General Discussion we consider what these costs might be.

General discussion

Our aim in these experiments was to investigate the processing of scalar implicatures without confounding accuracy with processing speed. To this end, we employed an SAT version of the paradigm developed by B&N. In Experiment 1 we found that the interpretation of scalar sentences with an implicature was delayed relative to the same sentences without an implicature. In Experiments 2 and 3 we explored why the implicature could have been a costly interpretation. We therefore factored out semantic complexity differences and compared pragmatic and logical *some* sentences against their explicit equivalents, *only some* and *at least some*, respectively. Participants were delayed in their interpretation of pragmatic-*some* relative to the explicit control more than in their interpretation of logical-*some*. Importantly, our results cannot be explained by speed-accuracy trade-offs, because of the deadline procedure and model-based analysis that we used (adapted from McElree & Nordlie, 1999). Before discussing the theoretical implications of our findings in more detail we consider alternative explanations of our results.

In Experiments 2 and 3, the explicit forms of the sentences used more words than the implicit forms. For example, *only some elephants are mammals* has one more word than *some elephants are mammals*. Our approach to this was to include *only*/*at least* and *some* in the same 250 ms window and force all of the sentences to have the four windows regardless of the number of words. Could this design choice explain the differences across conditions? At least three considerations argue against this possibility. First, we might expect participants to take longer to understand sentences with more words whereas the results were in the opposite direction: participants responded more quickly to longer sentences. If anything then, our results may have been an underestimate of the differences across conditions. Second, because there were at most three quantifiers, participants only had to glean a small amount of information (1.6 bits) from the first window of any sentence. Across hundreds of trials, it is likely that participants rapidly learned to efficiently extract the critical identifying information (e.g., whether the quantifier was *only some*, *some*, or *all* in Experiment 2). Finally, if length conferred an advantage in processing *only some* vs. *some* in Experiment 2, it should have similarly facilitated processing of *at least some* vs. *some* in Experiment 3. However, the advantage for *only some* was reliably greater than that for *at least some*. In short, it is unlikely that the number of words in each quantifier affected the qualitative results of the study.

Another potential concern is that directly manipulating the interpretation of *some* may have generated artefactual effects. Perhaps instructions to interpret *some* pragmatically were unclear, say, or perhaps the instructions forced participants to bypass the normal pragmatic procedures. We think that in these experiments the interpretation manipulation was unlikely to have caused irregularities in processing for the following reasons. First, verbal instructions to interpret *some* were short and appeared indirectly related to the experiment (c.f. B&N) wherever possible. The instructions concluded with, "One other thing,..." and participants were simply told that other participants sometimes experienced difficulty with the underinformative sentences but that the correct response to those types of sentences was "true" (or "false"). Furthermore, participants were only given one interpretation of *some* sentences (participants in B&N were told that there were several interpretations of *some*). Being told how to interpret particular questions in an experiment is no different to many other instructions the participant might receive about experimental procedure. By disguising the goal of the experiment with respect to *some* we avoided metalinguistic effects of directly manipulating interpretations. Second, while the verbal instructions might have helped to push a participant towards a particular interpretation, we feel that the feedback provided in the 70 practice trials (three times as many as B&N) was the major determiner of the interpretation. It is doubtful that participants remembered what we had told them in the instructions after they had received the "correct" vs. "incorrect" feedback across a large number of trials. Indeed, we believe that the feedback manipulation was so strong that the results of the experiments would be

identical without the verbal instructions (the difference is that perhaps more participants would be eliminated from the study for low overall performance without the verbal instructions). Finally, there was no evidence that participants had difficulty understanding how they were supposed to respond in either the logical or the pragmatic conditions: we observed very little difference in asymptotic accuracy performance between the instructed *some* conditions and the semantically forced interpretations (*at least some* and *only some*).

Sentence verification

Our study involved explicitly verifying sentences. Here we consider limitations of this paradigm (see e.g., Feeney et al., 2004; Grodner et al., 2010; Huang & Snedeker, 2009; Nieuwland, Ditman, & Kuperberg, 2010; for criticisms of the sentence verification task).

First, because people do not explicitly verify statements in everyday communication, is it possible that our results are an artifact of the verification procedure? Our view is that asking for explicit judgments does not change the underlying linguistic mechanisms. One reason for this is that while people clearly do not explicitly verify statements, they are likely to implicitly verify statements on a regular basis. Understanding an assertion involves the comparison between a proposition and one's representation of the world, that is, an implicit verification (and indeed, model theoretic semantics assumes that interpretation amounts to deriving the truth conditions of a sentence). For example, if someone says, "Some of the students are in the classroom," and you happen to know that none of them are in the classroom, the inconsistency between the speaker's statement and your own knowledge will immediately become apparent even though you have not been asked to make an explicit verification judgment (and even when there is no inconsistency between the new assertion and the knowledge of the listener the statement must be verified in order to be judged acceptable, i.e., not inconsistent with known information). There is also experimental evidence that people verify information as they are reading (e.g., Rapp, 2008) and, more pertinent to this study, evidence that people are sensitive to underinformativeness even when they are not explicitly asked to verify statements (Nieuwland et al., 2010).

Our experiments have only investigated assertions, however, for which verification is an intuitively natural procedure. There are other speech acts for which it is less clear what role verification plays, such as requests or commands, and it is an empirical question as to whether our results generalize to these situations. While we cannot say what would happen, we feel that implicatures may only be derived in circumstances where the implicature is necessary to maintain conversational coherence. Verification of underinformative sentences is one context in which the deriving the implicature becomes necessary. Other studies that have investigated implicatures have also used contexts that make the distinction between lower and upper-bound meanings very salient, and arguably necessary to maintain discourse coherence (e.g., B&N; Breheny et al., 2006; Grodner et al., 2010; Huang & Snedeker, 2009). In Breheny et al., participants needed to derive a complement set when they read, "the others." Without the deriving the implicature, the text would have been incoherent, and so the pronoun may have been causing the implicature to be derived, just as requiring a verification response forced participants to derive the implicature. Similarly, in Huang and Snedeker, participants needed to derive the implicature in order to determine whether to click on the referent with *some*, as opposed to *all*, of the objects. Some way of making the implicature a necessary part of maintaining conversational coherence may be always be needed, whether this is a verification task or any other form of contextual manipulation (consistent with "good enough" processing, see Ferreira, Ferraro, & Bailey, 2002).

A final issue with our paradigm is that it is difficult to know why participants rejected the experimental sentences. Throughout the article we have assumed that participants correctly rejected the pragmatic-*some* sentences because they derived scalar implicatures and then rejected the resulting proposition as being inconsistent with their world knowledge. For example, we assumed that they derived the interpretation *some* [but not all] *elephants are mammals* to the experimental sentences and then rejected them because this was incorrect relative to their representation of elephants and mammals. However, participants may also have been rejecting the sentences because they were underinformative relative to what the speaker could have said (e.g., *all elephants are mammals*). Under a classical Gricean account (Grice, 1975), there are at least two steps to deriving an implicature. The first step involves determining whether the speaker could have made a more informative statement than the expression used (e.g., *all elephants are mammals* instead of *some elephants are mammals*), and the second step involves negating the more informative statement to explain why the speaker chose not to utter it (e.g., the speaker must have meant that *some but not all elephants are mammals*). Participants in our study could have been rejecting the experimental sentences because the sentences were not maximally informative (after the first step), rather than rejecting them because the implicated meaning was false (a consequence of the second step). A similar point was made by Katsos and Bishop (2011) with respect to truth-value judgment tasks used in the developmental literature on scalar implicatures (e.g., Noveck, 2001).

While it is possible that participants were rejecting sentences entirely on the underinformativeness, some evidence against this is shown in the comparison between *only-some* and pragmatic-*some* in Experiment 2. Since the *only-some* sentences explicitly specify the upperbound meaning, rejection of these sentences is presumably based on inconsistency with world knowledge, rather than underinformativeness. If participants were rejecting the pragmatic-*some* sentences using other criteria than those used to evaluate the *only-some* sentences, accuracy rates might be expected to differ. In contrast to this prediction we found that the asymptotic accuracy for the *only some* sentences was almost identical to that of the pragmatic-*some* sentences (see Fig. 8). The high degree of similarity between accuracy on the explicit and the implicit

upperbound sentences is due to the strength of the biasing context (primarily the feedback). This is quite different to developmental studies in which participants are typically not given a biasing context (e.g., Katsos & Bishop, 2011; Noveck, 2001) and participants might therefore be rejecting sentences for different reasons in those studies.

We cannot say for sure why participants rejected the experimental sentences. Most important, however, is that participants were engaged in some form of Gricean reasoning under either account. This means that the central hypotheses under investigation make the same predictions regardless of which rejection account turns out to be correct. Both the standard default implicature hypothesis and the nonverification hypothesis assume that the implicature is computed on every occasion. That is, the epistemic step should always be computed provided that the epistemic conditions are met.

Implications for theories of scalar implicatures

The results of our study are strong evidence against processing theories in which scalar implicatures are derived on every occasion and sometimes cancelled. Included in this set are processing theories motivated by so-called Neo-Gricean theories (e.g., Levinson, 2000), which assume that implicatures are derived using Gricean machinery combined with lexical specification of the semantic scale, e.g., <some, many, all>. Processing accounts of Neo-Gricean theories predict that upper-bound *some* should be processed more quickly than lower-bound *some*, but this is the opposite of what we found (Experiment 1). Furthermore, an extension of the standard default model to a nonverification model also failed to make the correct predictions. In Experiment 2, we compared an explicitly formed upper-bound interpretation (using *only*) against an implicitly formed upper-bound interpretation. If implicatures are computed by default and the necessary computations occur lexically, there is no reason to predict that there should be differences between the implicit and explicit formations of the upper-bound interpretation. In contrast to these predictions, however, we observed delayed sentence interpretations in the implicit condition. This cannot be due to a general cost in resolving the bare quantifier because we observed a greater cost of interpreting pragmatic-*some* relative to the upper-bound explicit form (*only some*) than logical-*some* relative to the lower-bound explicit form (*at least some*).

Our results also speak to processing models that assume scalar implicatures are computed using the grammatical properties of the sentence. The linguistic models from which these accounts might be derived are typified by Chierchia (2004, 2006; and see Gazdar, 1979; Landman, 1998) (although these authors never intended their accounts to predict processing data). Chierchia's approach assumes that scalar implicatures are computed recursively and compositionally, in a similar to way to ordinary meaning. The implicatures are computed locally with the alternatives generated using a focus operator. Indeed, Chierchia suggests that implicature computations are tantamount to a silent "only" on the basic meaning of the scalar term (Chierchia, 2006, p11). In our view there are several potential processing instantiations of such a model. One version might be that the focus operator is always applied so that the implicature occurs on every occasion. The implicature can be cancelled, but just as in a Neo-Gricean account, cancellation can only occur after the implicature has been generated. Under this view grammatical models could be seen as default implicature models, and with similar predictions. Our results argue against such a model. An alternative is that context and other factors might prevent the focus operator from being applied on every occasion so that the lower-bound interpretation could be directly incorporated into the sentence representation. The predictions for this processing model are less clear but, intuitively, a model that assumes scalar implicatures are generated using a silent "only" operator would not predict interpretation time differences between sentences presented with an implicit *only* (the pragmatic-*some* condition in Experiment 2) and an explicit *only* (the *only* condition of Experiment 2). Therefore, such a model offers no explanation for the difference we observed in Experiment 2.

Finally, we consider contextual models, in which the implicature is not derived automatically on every occasion, but instead is determined by the context. As we argued in Experiment 3, cost-free contextual models with semantic complexity explanations cannot account for the difference between *only-some* and pragmatic-*some*, and contextual models that assume only a general cost to resolving *some* cannot explain why pragmatic-*some* is more costly to derive than logical-*some*, relative to their respective literal equivalents. Consistent with the data are therefore those models that predict a cost to deriving the implicature, but that predict a smaller or nonexistent cost to understanding the lower-bound interpretation. In fact, a processing version of the standard Gricean model accounts for our data quite well, under the assumption that people are able to block, or prevent, the implicature from arising in appropriate contextual circumstances. If implicatures were not considered in the logical-*some* conditions of Experiments 1 and 3 there would be no reason for logical-*some* responses to be delayed relative to the explicit equivalent. What mechanisms might prevent the implicature from arising? There are at least two possibilities in a Gricean account. Participants might have either "cancelled" the scalar implicature prior to verifying the sentences, or they might not have considered the scalar alternative (*all*) relevant to the task. Cancelling the implicatures prior to uttering the scalar term is conceptually problematic (see Geurts, 2010, for an in depth discussion of this point), but it seems plausible that participants would never consider the scalar alternatives relevant under the logical training in our task. In the logical-*some* conditions, participants were never given any indication that they should derive the implicature, so there was no need for them to consider the alternatives important. Conversely, in the pragmatic-*some* conditions, the training and feedback ensured that they must do. The extra cost of pragmatic-*some* compared to *only some* would arise because of the standard Gricean stages involved in computing the scalar implicature (which we discuss in the next section).

Another contextual model that would be consistent with our data is relevance theory (Carston, 1998; Sperber & Wilson, 1986/1995). Relevance theorists argue that going beyond the encoded content of an utterance is costly, which explains why pragmatic-*some* was delayed relative to *only-some* in Experiment 2. Noveck and Sperber (2007) have also explicitly stated that the lowerbound meaning involves going beyond the encoded content. This could explain why there may have been a cost to logical-*some* in Experiment 3. Nonetheless, there are no relevance theory specific mechanisms that might explain precisely why a cost should arise.

Why implicatures are costly

Deriving an upper-bound interpretation involves several extra computations relative to deriving a lower-bound interpretation, as described above and by B&N. These might include deriving a complement set, processing negation (e.g., *not all*) and completing a more complex memory search to integrate the meaning of the upper-bound sentence with knowledge schemas. All of these are part of what it means to understand a scalar implicature. Our experiments, however, have suggested that there are costs of implicatures linked to the inferential mechanism *per se* over and above whatever costs might be associated with deriving or verifying the resulting (upper-bound) sentence structure (Experiments 2 and 3). Here we consider some of the possible explanations for the inferential cost.

First, extra time might be required to execute the epistemic step when generating an implicature, that is, the inference from *the speaker had no evidence that p*, to *the speaker has evidence that not p* (part of Step 2 of the standard Gricean account, as discussed above; see Sauerland, 2004). Experimental evidence that addressees take into account the speaker's knowledge is shown by experiments using unreliable or unknowledgeable speakers, in which the pragmatic interpretation is derived only when the, participant has a reasonable expectation that the speaker is reliable (Bergen & Grodner, 2010; Grodner & Sedivy, in press). The epistemic step and other speaker-knowledge effects would not be necessary in *only some* sentences because *only* communicates that the speaker has the knowledge to deny the stronger elements in the scale. Of course, if participants were not making the epistemic step at all – merely rejecting pragmatic-*some* sentences on the basis of their underinformativeness (as considered above) – the epistemic step cannot be the cause of the slow pragmatic-*some* responses.

A second possibility is that *only* forces costly upper-bound computations to occur earlier in the sentence than implicatures. Whereas implicatures might be generated at a sentence level, *only* could force the upper-bound interpretation to occur on or around the quantifier. Since with *only* there is no doubt that the speaker intends the *not all* inference to be drawn, there is no need to postpone computation of the implicature until pragmatic factors, such as speaker's intentions, can be incorporated or assessed at a sentence level. Under this account, the cost of implying would be to delay processing rather than extend the time required (although because there are extra words in the explicit upper-bound the overall comprehension time may be faster in the implicit upper-bound). While this account is intuitively appealing, there is very little evidence in previous work that *only* affects processing (more than implicatures) early in sentence comprehension. Using self-paced reading tasks, Breheny et al. (2006) found a significant difference between *only some* and lower-bound *some* on the quantifier region, but no difference between *only some* and upper-bound *some*, and Hartshorne and Snedeker (submitted for publication) found no differences between *only some*, lower-bound *some*, and upper-bound *some*. In contrast, effects of *only* have been observed late in processing. Filik, Paterson, and Liversedge (2009) investigated the difference between *only* and *even* in relation to focus. They found longer reading times for *only* than *even* on the region in which the complete proposition was available but not before (although the study has limited applicability to scalar processing because quantifiers were not used and because results were reported with respect to *even*). Further studies using techniques better suited to identifying mid-sentence, lexical effects, such as eyetracking, could be used to test the effects of *only* on quantifiers.

A further explanation for the *only some* advantage derives from underinformative nature of the experimental sentences. As we discussed above, *some elephants are mammals* is underinformative relative to what participants know to be correct, *all elephants are mammals*. While participants were given training to reject the experimental sentences in the pragmatic-*some* condition, it is possible that some infelicity remained and caused participants to slow down their responses. The semantic properties of *only* would prevent the pragmatic infelicity from arising because the *only some* sentences were not underinformative, merely false. This account is possible but in order to explain the greater delay for *only some* compared to *at least some*, a greater infelicity would have to exist in pragmatic-*some* sentences than logical-*some* sentences (while receiving equal amounts of contextual training), even though the sentences were identical. Furthermore, if there was any infelicity remaining after the training phase, it seems likely that this would have been reflected in the asymptotic accuracy measure in addition to the speed dynamics; yet there were no differences between implicit and explicit *some*.

Another potential explanation for the delayed pragmatic-*some* interpretation is that the lower-bound meaning might first need to be rejected before derivation of the upper-bound meaning can take place (a default logical interpretation model, as opposed to a default implicature model). While this explanation would be at odds with most psycholinguistic investigations of pragmatics that have not found evidence of the need to reject the literal interpretation of conventional expressions, such as indirect speech (e.g., Shapiro & Murphy, 1993) or metaphor (McElree & Nordlie, 1999), scalar implicatures display many distributional properties that are different to standard inferences (see Geurts, 2010) and they might therefore invoke quite different processing mechanisms. For example, upper-bound scalar sentences entail their lower-bound

counterparts, by definition (whenever *some but not all X are Y* is true, *some X are Y* must be true), so that it is efficient to always consider (and possibly reject) the lower-bound interpretation on every occasion, whereas the figurative meaning of metaphors, say, does not entail their literal meaning.

The explanations for the implicature cost discussed above are separate and empirically distinguishable using the right techniques. Whatever mechanism turns out to be causing the cost, however, it is apparently very persistent. Even after corrective feedback and hundreds of trials participants were still unable to adjust their interpretative mechanism so that when they read "some" they understood the upper-bound interpretation directly. It would seem relatively simple to remap *some* to *only some* for example, but the effects we observed suggest otherwise. The cost of deriving a scalar implicature – relative to an explicit equivalent – seems to be obligatory in environments such as ours.

Conclusion

The experiments presented in this study investigated what might cause scalar implicatures to be processed more slowly relative to equivalent sentences without an implicature. Our data makes two major contributions to establishing what these costs might be. First, we have shown that delayed upper-bound interpretations are not due to simple speed-accuracy trade-off strategies. Second, we made a distinction between sentence complexity costs, that is, costs due to differences in propositional content, and inference costs, that is, costs specifically related to deriving the implicature. While we cannot say for sure what proportion of the costs are due to each of these components, we can say that there is a significant and persistent cost from the inference mechanism *per se*. We look forward to the results of other studies that are able to test which of several potential explanations are responsible for this inference cost.

Acknowledgments

This work was funded by ESRC Award RES-062-23-2410 to L. Bott, T.M. Bailey & D. Grodner. The authors are grateful for helpful comments from Brian McElree, Napoleon Katsos, and an anonymous reviewer.

Appendix: Raw accuracy proportions averaged across participants

See Tables A1–A3.

Table A1
Experiment 1 accuracy proportions as a function of training condition, deadline and sentence type.

	Deadline							
Sentence	1	2	3	4	5	6	7	8
Pragmatic condition								
S1	0.499	0.492	0.438	0.409	0.432	0.608	0.680	0.835
S2	0.515	0.574	0.660	0.757	0.862	0.897	0.923	0.928
S3	0.449	0.435	0.471	0.550	0.706	0.858	0.893	0.911
S4	0.514	0.485	0.563	0.729	0.761	0.834	0.866	0.915
S5	0.430	0.462	0.588	0.757	0.826	0.894	0.906	0.952
S6	0.528	0.581	0.667	0.748	0.802	0.876	0.930	0.937
Logical condition								
S1	0.854	0.856	0.849	0.892	0.909	0.944	0.949	0.985
S2	0.144	0.187	0.393	0.620	0.747	0.838	0.818	0.877
S3	0.853	0.840	0.777	0.815	0.873	0.895	0.922	0.897
S4	0.847	0.835	0.859	0.881	0.913	0.916	0.950	0.929
S5	0.168	0.201	0.438	0.632	0.759	0.800	0.826	0.876
S6	0.840	0.888	0.883	0.924	0.929	0.945	0.947	0.961

Table A2
Experiment 2 accuracy proportions as a function of deadline and sentence type.

	Deadline							
Sentence	1	2	3	4	5	6	7	8
S1	0.305	0.330	0.401	0.517	0.624	0.689	0.758	0.868
S2	0.377	0.429	0.503	0.603	0.689	0.693	0.776	0.855
S3	0.802	0.820	0.842	0.860	0.882	0.899	0.914	0.934
S4	0.641	0.616	0.708	0.748	0.760	0.810	0.868	0.902
S5	0.584	0.641	0.669	0.682	0.767	0.788	0.815	0.903
S6	0.637	0.682	0.722	0.767	0.797	0.827	0.859	0.877

Table A3
Experiment 3 accuracy proportions as a function of deadline and sentence type.

	Deadline							
Sentence	1	2	3	4	5	6	7	8
S1	0.561	0.597	0.704	0.752	0.823	0.848	0.877	0.926
S2	0.627	0.654	0.722	0.738	0.840	0.837	0.883	0.925
S3	0.615	0.647	0.729	0.743	0.807	0.856	0.874	0.911
S4	0.418	0.493	0.719	0.844	0.889	0.915	0.905	0.958
S5	0.388	0.494	0.697	0.823	0.876	0.923	0.905	0.925
S6	0.511	0.529	0.737	0.824	0.868	0.871	0.907	0.934

References

Bergen, L., & Grodner, D. (2010). Scalar implicatures are sensitive to the speaker's epistemic state. In *Poster presented at the 23rd CUNY conference on human sentence processing*. New York, NY.

Bonnefon, J. F., Feeney, A., & Villejoubert, G. (2009). When some is actually all: Scalar inferences in face-threatening contexts. *Cognition, 112*, 249–258.

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language, 51*, 437–457.

Breheny, R., Katsos, N., & Williams, J. (2006). Are generalized scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition, 100*, 434–463.

Carston, R. (1998). Informativeness, relevance and scalar implicature. In R. Carston & S. Uchida (Eds.), *Relevance Theory: Applications and implications* (pp. 179–236). Amsterdam: John Benjamins.

Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. In A. Belletti (Ed.), *Structures and beyond*.

Chierchia, G. (2006). Broaden your views: Implicatures of domain widening and the logicality of language. *Linguistic Inquiry, 37*, 535–590.

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology, 3*, 472–517.

Feeney, A., Scafton, S., Duckworth, A., & Handley, S. J. (2004). The story of some: Everyday pragmatic inferences by children and adults. *Canadian Journal of Experimental Psychology, 58*, 121–132.

Ferreira, F., Ferraro, V., & Bailey, K. G. D. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science, 11*, 11–15.

Filik, R., Paterson, K. B., & Liversedge, S. P. (2009). The influence of only and even on online semantic interpretation. *Psychonomic Bulletin & Review, 16*, 678–683.

Forster, K., & Forster, J. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers, 35*, 116.

Gazdar, G. (1979). *Pragmatics: Implicature, presupposition, and logical form*. New York: Academic Press.

Geurts, B. (2010). *Quantity implicatures*. Cambridge University Press.

Grice, H. P. (1975). Logic and conversation. In P. Cole, & J. Morgan (Eds.) *Syntax and semantics 3: Speech acts* (pp. 41–58).

Grodner, D., & Sedivy, J. (in press). The effect of speaker-specific information on pragmatic inferences. In N. Pearlmutter, & E. Gibson (Eds.), *The processing and acquisition of reference*. Cambridge, MA: MIT Press.

Grodner, D., Gibson, E., & Watson, D. (2005). The influence of contextual contrast on syntactic processing: Evidence for strong interaction in sentence comprehension. *Cognition, 95*, 276–296.

Grodner, D., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). "Some", and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition, 116*, 42–55.

Hartshorne, J. K., & Snedeker, J. (submitted for publication). The speed of inference: Evidence against rapid use of context in calculation of scalar implicatures.

Hirschberg, J. (1991). *A theory of scalar implicature*. New York: Garland Publishing Company.

Horn, L. R. (1972). *On the semantic properties of logical operators in English*. Ph.D. thesis, University of California, Los Angeles.

Horn, L. R. (1989). *A Natural History of Negation*. Chicago, Ill: University of Xcago Press.

Huang, Y. T., & Snedeker, J. (2009). On-line interpretation of scalar quantifiers: Insight into the semantic-pragmatics interface. *Cognitive Psychology, 58*, 376–415.

Katsos, N., & Bishop, D. V. M. (2011). Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition, 120*, 67–81.

Landman, F. (1998). Plurals and maximalization. In S. Rothstein (Ed.), *Events and grammar*, Dordrecht: Kluwer.

Levinson, S. C. (2000). *Presumptive meanings*. Cambridge, Mass: MIT Press.

Martin, A. E., & McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *Journal of Memory and Language, 58*, 879–906.

McElree, B. (1993). The locus of lexical preference effects in sentence comprehension: A time-course analysis. *Journal of Memory and Language, 32*, 536–571.

McElree, B. (1998). Attended and non-attended states in working memory: Accessing categorized structures. *Journal of Memory and Language, 38*, 225–252.

McElree, B., & Dosher, B. A. (1989). Serial position and set size in short-term-memory – the time course of recognition. *Journal of Experimental Psychology – General, 118*, 346–373.

McElree, B., & Griffith, T. (1995). Syntactic and thematic processing in sentence comprehension – evidence for a temporal dissociation. *Journal of Experimental Psychology – Learning Memory and Cognition, 21*, 134–157.

McElree, B., & Griffith, T. (1998). Structural and lexical constraints on filling gaps during sentence comprehension: A time-course analysis. *Journal of Experimental Psychology – Learning Memory and Cognition, 24*, 432–460.

McElree, B., & Nordlie, J. (1999). Literal and figurative interpretations are computed in equal time. *Psychonomic Bulletin & Review, 6*, 486–494.

Nieuwland, M. S., Ditman, T., & Kuperberg, G. R. (2010). On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. *Journal of Memory and Language, 63*, 324–346.

Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition, 78*, 165–188.

Noveck, I., & Posada, A. (2003). Characterizing the time course of an implicature: an evoked potentials study. *Brain and Language, 85*, 203–210.

Noveck, I., & Sperber, D. (2007). The why and how of experimental pragmatics: The case of scalar inferences. In Noel Roberts (Ed.), *Advances in pragmatics*. Palgrave.

Rapp, D. N. (2008). How do readers handle incorrect information during reading? *Memory & Cognition, 36*, 688–701.

Reed, A. V. (1976). List length and time course of recognition in immediate memory. *Memory & Cognition, 4*, 16–30.

Rips, L. (1975). Quantification and semantic memory. *Cognitive Psychology, 7*, 307–340.

Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy, 27*, 367–391.

Shapiro, A. M., & Murphy, G. L. (1993). Can you answer a question for me? Models of processing indirect speech acts. *Journal of Memory and Language, 32*, 211–229.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition*. Oxford: Blackwell.

Traxler, M. J., Pickering, M. J., & Clifton, C. Jr., (1998). Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language, 39*, 558–592.

Van Gompel, R. P. G., Pickering, M. J., Pearson, J., & Liversedge, S. P. (2005). Evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language, 52*, 284–307.