COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 41 (2017) 590–622

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12353

A Bootstrapping Model of Frequency and Context Effects in Word Learning

George Kachergis,[a] Chen Yu,[b] Richard M. Shiffrin[b]

[a]*Department of Psychology, New York University*
[b]*Department of Psychological and Brain Sciences/Cognitive Science Program, Indiana University*

Received 9 December 2013; received in revised form 8 December 2015; accepted 9 December 2015

Abstract

Prior research has shown that people can learn many nouns (i.e., word–object mappings) from a short series of ambiguous situations containing multiple words and objects. For successful cross-situational learning, people must approximately track which words and referents co-occur most frequently. This study investigates the effects of allowing some word-referent pairs to appear more frequently than others, as is true in real-world learning environments. Surprisingly, high-frequency pairs are not always learned better, but can also boost learning of other pairs. Using a recent associative model (Kachergis, Yu, & Shiffrin, 2012), we explain how mixing pairs of different frequencies can bootstrap late learning of the low-frequency pairs based on early learning of higher frequency pairs. We also manipulate contextual diversity, the number of pairs a given pair appears with across training, since it is naturalistically confounded with frequency. The associative model has competing familiarity and uncertainty biases, and their interaction is able to capture the individual and combined effects of frequency and contextual diversity on human learning. Two other recent word-learning models do not account for the behavioral findings.

Keywords: Statistical learning; Language acquisition; Cross-situational learning; Contextual diversity; Word frequency

1. Introduction

Despite the high degree of referential uncertainty in the world, infants learn nouns with astonishing speed. Assuming that caregivers sometimes refer to visible objects, a learner who can remember some of the co-occurring words and referents can gradually learn the intended word-referent mappings after experiencing a variety of situations. Cross-situational learning based on cross-modal memory and the statistics of the

Correspondence should be sent to George Kachergis, Department of Psychology, New York University, 6 Washington Place, New York, NY 10003. E-mail: george.kachergis@nyu.edu

language environment may be an important way for infants to acquire nouns (Gleitman, 1990; Smith, 2000). Cross-situational learning has been demonstrated by infants (Smith & Yu, 2008) and by adults (Yu & Smith, 2007). As an ability that is likely key to acquiring language—perhaps humanity's most defining trait, cross-situational word learning also offers an enticing glimpse into the interlocking fundamental mechanisms of human cognition, as it likely relies on domain-general attention, memory, and learning processes (Kachergis, 2012; Smith, 2001).

In adult cross-situational learning studies, participants are instructed to learn which word goes with which object and then study a series of training trials. On each trial, an array of several novel objects is displayed while pseudowords are successively heard. Although each pseudoword refers to a particular onscreen object, the correct referent for each pseudoword is not indicated, thus making meanings ambiguous on individual trials. For example, you might see objects $\{o_1, o_2\}$ on the first trial, while hearing words {*manu*, *bosa*}. You cannot know if *manu* refers to o_1, o_2, both, or neither; the same is true of *bosa*. On a later trial you see $\{o_3, o_1\}$ while hearing {*bosa*, *stigson*}. If you have any memory of *bosa* having appeared with o_1 previously, you may prefer to strengthen that pairing (i.e., *bosa*-o_1) rather than storing *bosa*-o_3. If you assume that words are mapped 1-to-1 to objects, you might also focus on the *stigson*-o_3 association, rather than considering the possibility that *stigson* also refers to o_1. Yurovsky and Yu (2008) and Kachergis, Yu, and Shiffrin (2012a) have shown that this bias for mutually exclusive pairings, a bias observed in 2-year-olds (Markman & Wachtel, 1988; Merriman & Bowman, 1989), is present in adults and can be succinctly explained using an associative model with competing biases for strengthening prior knowledge and for attending to stimuli with uncertain associates (Kachergis et al., 2012a). It is not unreasonable to assume that adults and infants share the same basic kinds of mechanisms for language acquisition, though they undoubtedly differ in degree. Because adults can endure longer duration studies allowing more complex designs, the data from such studies can produce additional insights beyond those available from studies of infants (e.g., Gillette, Gleitman, Gleitman, & Lederer, 1999; Kachergis et al., 2012a; Smith, Smith, & Blythe, 2011; Suanda & Namy, 2012; Yurovsky, Yu, & Smith, 2013).

In the original study reported in Yu and Smith (2007) and many follow-up studies (e.g., Kachergis, Yu, & Shiffrin, 2010; Suanda & Namy, 2012; Yurovsky, Yu, et al., 2013), language learners are exposed to a set of to-be-learned word-object pairs with equal frequency. This study asks how varied word-object pair frequency affects the course of learning. Word frequency varies greatly in natural language (Zipf, 1949), and higher frequency words are more likely to be learned faster by infants (Hills, Maouene, Riordan, & Smith, 2010). Intuitively, it seems that more frequently appearing pairs will be learned far more easily than less frequent pairs, given the greater number of opportunities for disambiguation and storage. It seems reasonable that once high-frequency pairs are well-known, attention should shift from these pairs to lower frequency pairs. Continuing the earlier example, if you later experience a trial with objects $\{o_1, o_4\}$ and words {*bosa*, *fimi*}, you may focus only on storing *fimi*-o_4, since *bosa*-o_1 is already quite certain. If learners indeed bootstrap the learning of low-frequency words using prior knowledge of high-frequency pairs, they may be able to learn more of both the high- and low-frequency mappings.

However, it is not only a given pair's frequency and knowledge state that might influence attention, but also those of the pairs that co-occur with it. It seems reasonable that a pair will be learned better if it appears in a set of trials with sufficiently diverse contents (i.e., *contexts*). In the extreme, if two words and objects always occur together, even many times, the correct pairings for these stimuli would remain ambiguous, regardless of the number of occurrences of these trials. Thus, a stimulus pair that appears with only a few other specific stimuli (i.e., has low contextual diversity) might be difficult to learn. Conversely, the more diverse the contexts in which a pair appears, the more likely may be the acquisition of that pair. Indeed, it has been suggested that word frequency effects on lexical decision times (i.e., for words in the adult mental lexicon) can be explained by contextual diversity (Adelman & Brown, 2008). Thus, this study focuses on two potentially influential factors in word learning: (a) *frequency*: repetitions per word-referent pair and (b) *contextual diversity*: the number of other pairs each pair appears with over time. The role of each individual factor in the context of cross-situational learning has not been systematically studied. Moreover, the potential interactions among these factors, as illustrated in the above examples, remain unexplored.

In addition, a third related factor is within-trial ambiguity: how many words and objects co-occur together in a learning situation. Until a pair has appeared with all other pairs in the vocabulary, increasing within-trial ambiguity can yield greater contextual diversity. Will the toll of increased ambiguity outweigh the advantages of increased contextual diversity? Similarly, greater pair frequency can yield greater contextual diversity until that pair has been seen with all other pairs. Are repetitions solely crucial as learning opportunities, or as a means to increase contextual diversity? The current studies systematically investigate these three factors—both individually and in combination—and measure their effects on word learning. More specifically, Experiment 1 will focus on frequency alone, whereas Experiment 2 will explore contextual diversity and within-trial ambiguity. Experiment 3 will explore the interaction of contextual diversity and frequency. By manipulating the learning input and measuring what is learned, we can discover factors that predicate successful acquisition and shed light on the underlying learning, memory, and attention mechanisms. Toward this end, we compare human performance to two computational word-learning models that have previously accounted for other word-learning behaviors: the incremental probabilistic model (Fazly, Alishahi, & Stevenson, 2010a) and the familiarity- and uncertainty-biased model (Kachergis et al., 2012a). Finally, we consider the propose-but-verify model (Trueswell, Medina, Hafri, & Gleitman, 2013), which assumes that learners store a single meaning hypothesis for each word and replace the hypothesis if it is disconfirmed.

2. Experiment 1

Participants were asked to simultaneously learn many word-referent pairs from a series of individually ambiguous training trials using the cross-situational word-learning

paradigm (Yu & Smith, 2007). Each training trial is comprised of a display of four novel objects with four spoken pseudowords. With no indication of which word refers to which object, learners have a small chance of guessing the four correct word-referent pairings from the 16 possible ones. However, since words always appear on trials with their intended referents, the correct pairings may be learned over the series of trials because the present design (like most) produces a statistical accumulation of pair counts that is highest for a single pairing.

The key manipulation of Experiment 1 is to repeat some pairs more often than others within the same set of trials. As discussed above, the more often a word-object pair is repeated, the more opportunities there are to deduce and rehearse that pairing. In addition, more frequent pairs appear with more other pairs, and thus have greater contextual diversity. We created two training conditions with subsets of pairs that appear with different frequency. In both conditions, training consisted of 27 trials containing 18 word-referent pairs, four of which were displayed on each trial. In the *two frequency subsets* condition (Fig. 1, left), nine of the stimulus pairs appeared three times (lower right), and nine of the pairs appeared nine times (upper left). In the *three frequency subsets* condition, six pairs appeared three times, six pairs appeared six times, and six pairs appeared nine times. A dramatic frequency effect was predicted: The more frequent pairings would be learned more often, and pairs with a mere three repetitions may not be learned at all. Importantly, the same pair was never allowed to appear in neighboring trials, as this would enable learners to selectively attend to the repeated (or unrepeated) stimuli and learn significantly more, as we have shown elsewhere (Kachergis, Yu, & Shiffrin, 2009b, 2013).

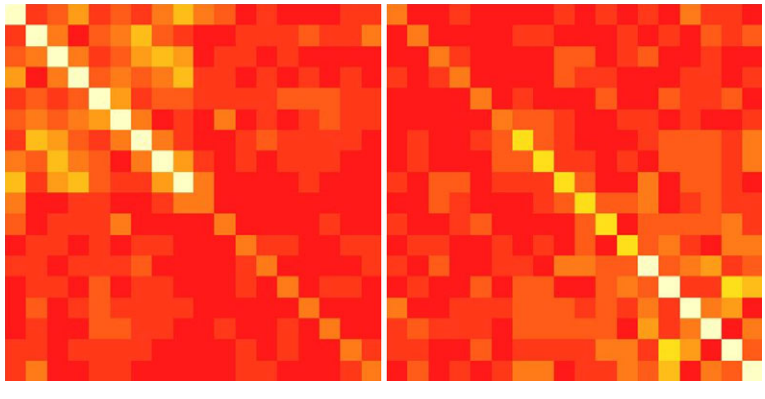

Fig. 1. Word-referent co-occurrence matrices for the two learning conditions in Experiment 1. Each cell represents the co-occurring frequency of a specific word-referent pair. The 18 correct pairs are on the diagonal. The other cells show spurious co-occurrences of incorrect word-referent pairs. Co-occurrences range from 0 (red) to 9 (white). Left: in the two frequency condition, 18 pairs form two frequency groups: nine repetitions (the top 9 pairs) and three repetitions (the bottom 9). Right: in the three frequency condition, 18 pairs appear at three different frequencies: 3, 6, and 9 (the top, middle, and bottom 6 pairs, respectively).

2.1. Subjects

Participants were 33 undergraduates at Indiana University who received course credit for participating. None had participated in other cross-situational experiments.

2.2. Stimuli

Each training trial consisted of four uncommon objects (e.g., strange tools) concurrently shown while four pseudowords were spoken sequentially. The 36 pseudowords generated by computer are phonotactically probable in English (e.g., "bosa") and were spoken by a monotone, synthetic female voice. These 36 arbitrary objects and 36 words were randomly assigned to two sets of 18 word-object pairings, one set for each learning condition.

Training for each condition consisted of 27 trials. Each training trial began with the appearance of four objects, which remained visible for the entire trial. After 2 s of initial silence, each 1 s word was heard followed by two additional seconds of silence, for a total duration of 14 s per trial. Words were heard in a random order for each participant, and condition order was counterbalanced.

After each training phase was completed, participants were tested for knowledge of word meanings. A single word was played on each test trial, and all 18 referents were displayed in locations that changed trial-to-trial. Participants were instructed to click on the correct referent for the word (i.e., 18AFC; 18-alternative forced choice). Each of the 18 words was presented once, and the test trials were randomly ordered.

2.3. Procedure

Participants were informed that they would see a series of trials with four objects and four alien words, and that their knowledge of which words belong with which objects would be tested at the end. After training, their knowledge was assessed using 18AFC testing: On each test trial a single word was played, and the participant was instructed to choose the appropriate object from a display of all 18. Condition order was counterbalanced.

2.4. Results and discussion

Fig. 2 displays the learning performance[1] for the subsets of pairs in both training conditions. To test the reliability of the differences between the means shown in Fig. 2, we fit a logistic mixed-effects regression model to the trial-level accuracy data using the lme4 package in R (Bates, Maechler, Bolker, & Walker, 2015; R Development Core Team, 2010). Mixed logit models are more appropriate for forced-choice data than ANOVAS, especially when different conditions yield different amounts of data, as in the present experiment (Jaeger, 2008). The model included random intercepts for subjects with random by-subjects slopes for Frequency, and Condition and Frequency as fixed factors (i.e., model syntax: Correct ~ Cond × Freq + (Freq|Subject)). Condition was coded as a main effect and

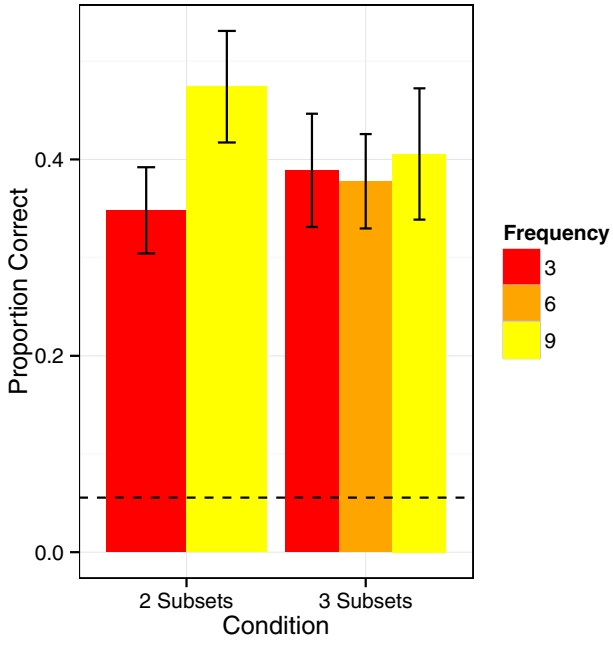

Fig. 2. Accuracy for subsets of pairs with different frequency in two training conditions. Learning was well above chance (dashed line; 18AFC chance = .056) in every condition. High-frequency pairs were learned better than low-frequency pairs in the two subsets condition, but there was no frequency advantage evident in the three subsets condition. Error bars show ±SE.

Frequency, a continuous predictor (3, 6, 9), was centered and scaled to [-1, 1]. There was a significant negative intercept, showing that participants were less likely to choose the correct answer than an incorrect answer ($b = -.47$, $OR^2 = 0.63$, Wald's $Z = -2.38$, $p < .05$). There was no significant main effect of Condition ($b = -0.11$, $Z = -0.80$, $p = 0.43$), with participants learning a mean proportion correct of .40 (7.2 pairs) per condition. There was a significant effect of frequency ($b = 0.17$, $OR = 1.18$, $Z = 1.98$, $p < .05$), with participants learning more nine-frequency pairs ($M_9 = .45$) than six- or three-frequency pairs ($M_6 = .38$, $M_3 = .36$). There was also a marginally significant interaction of Frequency and Condition ($b = -0.28$, $OR = 0.76$, $Z = -1.84$, $p = .07$). In the two-frequency subset condition, participants were significantly more likely to learn nine-frequency pairs ($M_9 = .47$) than three-frequency pairs ($M_3 = .35$, paired $t(29) = 3.08$, $p < .01$), in accord with the hypothesis that greater frequency aids statistical learning. However, this frequency advantage was barely evident in the three subsets condition, in which the subsets were learned nearly equally well ($M_3 = .39$, $M_6 = .38$, $M_9 = .41$).

Why did increased frequency aid learning in one condition, but not the other? How can it be explained that pairs of frequency 3, 6, and 9 are learned at equal rates? One plausible explanation is that once a pair is learned, future trials containing that pair effectively have reduced within-trial ambiguity. For example, if a learner sees (*A B*; *a b*) and has already learned *A*-*a*, then *B*-*b* may be inferred through one exposure where

it would not otherwise be certain. In this way, high-frequency pairs may be learned first and then used to effectively reduce the degree of ambiguity in later trials, and by doing so, they increase the learning of low-frequency pairs appearing in the same trials. If this is true, the contexts in which high- and low-frequency pairs co-occur should play a critical role in effective statistical learning. More generally, the context in which a word-object pair appears—whether with high-frequency (i.e., likely already-known) or low-frequency (i.e., likely not-yet-learned) words—may greatly affect learning. Indeed, the frequency effect in the two subsets condition could be due to limited opportunities for effective bootstrapping: with relatively more low-frequency pairs than the three subsets condition, trials with only one low-frequency pair (a good bootstrapping scenario) may be relatively rare. The smaller number of low-frequency pairs in the three subsets condition would make this type of trial more common, thus smoothing out frequency's effect on performance. In the next experiment, contextual diversity is varied to understand the counterintuitive finding in Experiment 1 and to directly measure the role of contextual diversity.

3. Experiment 2

Experiment 1 showed that higher frequency can result in greater learning, but does not necessarily do so. In Experiment 2, we hold word-referent frequency constant and vary the contexts in which each pair appears to measure how the learning of a given pair can be affected by the other pairs it co-occurs with during training. The contextual regularities for each word-referent pair can be captured by two factors: (a) the number of co-occurring words and referents within a trial, namely, within-trial ambiguity; and (b) the number of *different* co-occurring words and referents over all the training trials, namely, contextual diversity (CD). The three conditions in this experiment manipulated both factors. In the *low/medium CD* condition, 18 pairs were divided into two groups. Six word-referent pairs in the low CD group were constrained to appear only with other pairs in this group during training. Likewise, the 12 pairs in the medium CD group only co-occurred with each other, and never with the six low CD pairs (Fig. 3, left). Thus, whenever a low CD pair appeared, the other stimuli on that trial had to be selected from the five remaining low CD pairs. In contrast, a given medium CD pair could appear with any of the 11 other medium CD pairs. Note that frequency was held constant—each of the 18 pairs was seen six times during training—and within-trial ambiguity was the same (three words and three referents per trial). Only contextual diversity varied between these two groups. In each of the other two conditions in this experiment, all 18 pairs were randomly distributed to co-occur without constraint. To explicitly test the role of within-trial ambiguity, we implemented two versions of this design: the *uniform CD/3 pairs* condition with three words and three referents per trial, and the *uniform CD/4 pairs* condition with four words and four referents per trial (Fig. 3, middle and right, respectively).

Table 1 shows two metrics describing contextual diversity in this experiment: the mean number of other pairs that each pair co-occurs with during training, and the mean

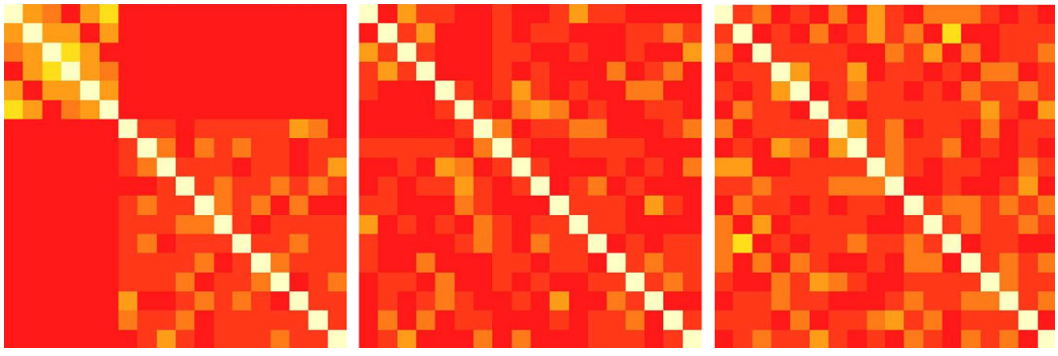

Fig. 3. Word-referent co-occurrences for Experiment 2 (0 = red, 6 = white). Left: in the low/medium CD condition, each group's pairs co-occur only with other pairs within that group. Middle and Right: in the uniform CD/3 pairs and the uniform CD/4 pairs conditions, each pair randomly co-occurs with any of 17 other pairs.

Table 1
Contextual diversity by condition in Experiment 2

Condition\CD	Low/Med		Uniform/3	Uniform/4
Pairs per CD Group	6	12	18	18
Mean # of different co-occurring pairs	4.0	9.2	8.8	12.2
Mean frequency of co-occurring pair	3.0	1.3	1.4	1.5

frequency of those co-occurring pairs. These two metrics are inversely related: If a given pair is made to co-occur with more other pairs, it must occur with each of these other pairs fewer times, on average. For example, if pair *A-a* always appears with pair *B-b*, the incorrect associations *A-b* and *B-a* may be learned as they appear equally frequently as *A-a* and *B-b*. However, if *A-a* appears with many other pairs, it is unlikely to occur very often with any one of them (e.g., *B-b*). This is an example of how contextual diversity may be important for learning.

Greater within-trial ambiguity not only creates more possible associations on each trial, but also influences CD: In the three pairs/trial conditions, each pair appears on six trials, and thus appears with 12 other pairs during training (unique or not). In the four pairs/trial condition, each pair appears with 18 other pairs during training, as it occurs on six trials with three other pairs. Thus, pairs in the four pairs/trial condition appeared with more diverse pairs than pairs in the three pairs/trial conditions. Moreover, note in Table 1 that the 12 medium CD group pairs have very similar CD—by both metrics—to the uniform/three pairs condition, since pairs in both these groups appeared with only 12 other pairs.

3.1. Subjects

Undergraduates at Indiana University received course credit for participating. The low/medium CD condition had 63 participants, and uniform three pairs/trial condition had 38

participants, and the uniform four pairs/trial had 77 participants.[3] None had previously participated in cross-situational experiments.

3.2. Stimuli and procedure

The sets of pseudowords and referents for Experiment 2 were identical to those used in Experiment 1, but several new trial orderings were constructed to vary contextual diversity and within-trial ambiguity. The 27-trial, four pairs/trial conditions had the same timing as Experiment 1. The 36-trial, three pairs/trial conditions also had 3 s per stimulus pair, with 2 s of initial silence, making a total of 11 s. Knowledge was assessed after the completion of each condition using 18AFC testing, as in Experiment 1.

3.3. Results and discussion

Fig. 4 displays the mean number of pairs learned in Experiment 2. We fit a logistic mixed-effects regression model (N = 3,132) to the trial-level accuracy data with subject as a random factor and with Condition and CD group, a continuous predictor (18, 12, or 6 pairs) centered and scaled to [−1, 1], as fixed factors, with by-subject random slopes for CD (model syntax: Correct ~ Cond + CD + (CD|Subject)). There was a significant negative intercept, showing that participants were less likely to choose the correct answer than an incorrect answer (b = −1.10, OR = 0.33, Z = −6.16, p < .001). Using the three

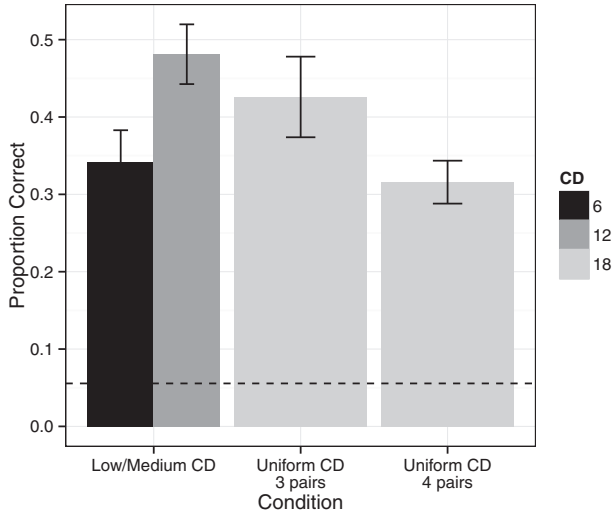

Fig. 4. Proportion correct by CD group for the three conditions of Experiment 2. The uniform CD four pairs/trial condition had lower performance than the three pairs/trial conditions. Within the Low/Medium CD condition, the twelve medium CD pairs had better performance than the six low CD pairs. In total, the number of pairs learned in the Uniform CD three pairs/trial conditions and the Low/Medium CD condition were nearly equal, and greater than the number learned in the four pairs condition. Error bars show ±SE.

pairs/trial uniform CD condition as baseline, there was a significant negative effect for the four pairs/trial condition ($b = -.78$, $OR = 0.46$, $Z = -5.43$, $p < .001$), showing that greater within-trial ambiguity leads to lower performance (3 pairs/trial $M = .43$, 4 pairs/trial $M = .32$). However, there was no significant effect of being in the low/medium CD condition ($b = -0.13$, $Z = -0.95$, $p = .34$), showing that this condition was overall no different than the three pair/trial condition (varied CD $M = .43$). As discussed, these two conditions have nearly the same degree of CD (see Table 1) along with the same level of within-trial ambiguity, which may explain their equal difficulty. There was also a significant effect of CD group ($b = 0.89$, $OR = 2.42$, $Z = 5.15$, $p < .001$), indicating that being in a larger CD group results in improved learning. Another model, this time using by-item CD (centered but not normalized), found similar results, with a CD odds ratio of 1.15 ($b = .14$, $Z = 4.62$, $p < .001$).

Within the low/medium CD condition, the 12 medium CD pairs were learned significantly better than the 6 low CD pairs (12 pairs $M = .47$, 6 pairs $M = .34$, paired $t(62) = 4.11$, $p < .001$), demonstrating a clear advantage for greater contextual diversity. Moreover, incorrect responses in the low/medium CD condition were largely chosen from the subset of pairs within the same group (thus co-occurring with the target pair): 56% of incorrect answers for low CD words were chosen from the 6 low CD referents (chance = 33%, $t(55) = 5.48$, $p < .001$), and 76% of incorrect answers for medium CD words were chosen from the 12 medium CD referents (chance = 66%, $t(55) = 3.72$, $p < .001$). Thus, even incorrect answers reflected co-occurrences encountered during training, rather than arbitrary guesses.

In summary, this experiment demonstrated that with the same frequency and degree of within-trial ambiguity, greater CD alone improves learning. However, the cost of greater within-trial ambiguity in the four pairs/trial condition outweighs any benefit conferred by greater CD in this condition (mean CD of 12.2 vs. mean CD of 8.8 in the three pairs/trial uniform condition; see Table 1). In Experiment 3, frequency and contextual diversity are manipulated within several conditions to elucidate the interaction of these factors.

4. Experiment 3

Experiment 2 showed that greater contextual diversity results in greater learning of those pairings. In Experiment 3, within-trial ambiguity was held constant, and frequency and contextual diversity were varied within four training conditions. Each condition had 18 pairs divided into three subsets of six pairs occurring at three frequencies: 3, 6, and 9. In the *low CD* condition, the pairs in each of the three frequency subsets appeared on trials only with pairs in the same group—never with pairs in other groups (Fig. 5a). That is, a three-repetition pair would only be seen with other three-repetition pairs, and similarly for six- and nine-repetition pairs. In this way, learning a three-repetition pair could help disambiguate only other three-repetition pairs, etc. In the *high CD* condition, pairs of different frequencies co-occurred randomly throughout training (Fig. 5b). In this condition, learning a given pair may help participants learn any pairs it co-occurred with in the

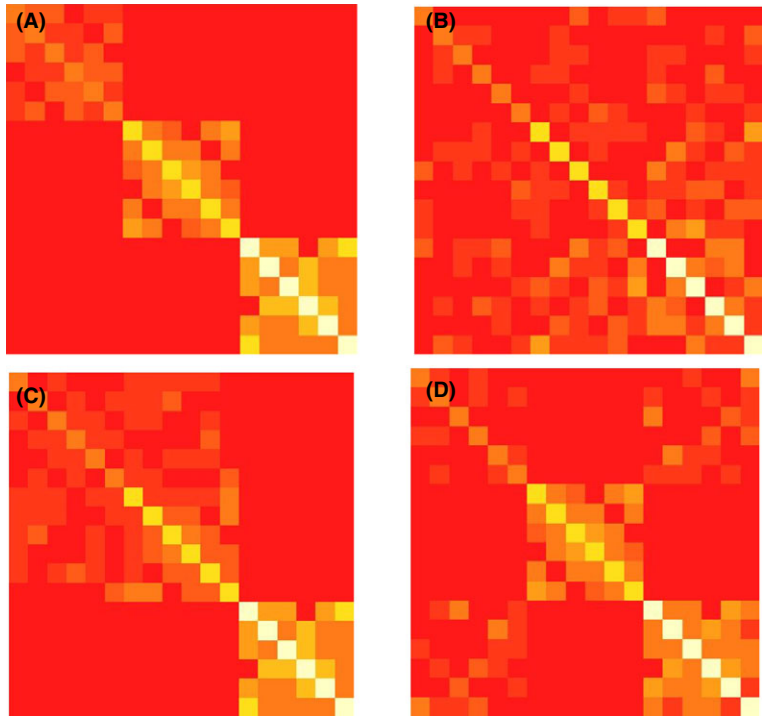

Fig. 5. Co-occurrence matrices (0 = red, 9 = white) from each condition. There were three frequency subsets in each condition (3, 6, and 9). To-be-learned pairs were manipulated in four ways to co-occur within and between each subset.

future. In the final two conditions, the 12 pairs from two frequency subsets were allowed to co-occur, and the remaining six pairs co-occurred only with themselves (i.e., within-frequency). In the *3/6 mingled* condition, the three- and six-repetition pairs co-occurred during training, and the nine-repetition pairs only appeared with other nine-repetition pairs (Fig. 5c). In the *3/9 mingled condition*, three- and nine-repetition pairs were mixed, and the six-repetition pairs could only appear with other six-repetition pairs (Fig. 5d).

4.1. Subjects

Participants were undergraduates at Indiana University who received course credit for participating. The low CD and high CD conditions had 34 and 67 participants, respectively. The 3/6 mingled condition and 3/9 mingled conditions had 66 and 40 participants, respectively.[4] None had previously participated in cross-situational experiments.

4.2. Stimuli and procedure

The 72 pseudowords and 72 objects used for Experiment 3 were the same as those used in Experiments 1 and 2, assigned to four sets of 18 word-object pairings, but several

new trial orderings were constructed to covary contextual diversity with pair frequency. Training for each condition consisted of 36 trials. Each training trial began with the appearance of three objects, which remained visible for the entire trial. After 2-s of initial silence, each of the three words was heard (randomly ordered, duration of 1-s) followed by two additional seconds of silence, for a total duration of 11 s per trial. After each training phase, participants were given an 18AFC test for knowledge of each word, randomly ordered as in Experiments 1 and 2.

4.3. Results and discussion

Fig. 6 displays the average levels of learning achieved in Experiment 3, split by condition and frequency subset. We fit a logistic mixed-effects regression model[5] to the trial-level accuracy data (N = 3,744) with CD group (18, 12, or 6 pairs) and Frequency (3, 6, or 9) represented as continuous numeric predictors (centered and scaled to have unit deviation), with Frequency, CD, and their interaction as fixed effects and including by-subjects random slopes for CD, Frequency, and their interaction. The estimated intercept was not significant (b = .22, Z = 1.50, p = .13). There was a significant positive effect of Frequency (b = 0.74, OR = 2.10, Z = 11.85, $p < 0.001$), showing that higher frequency generally

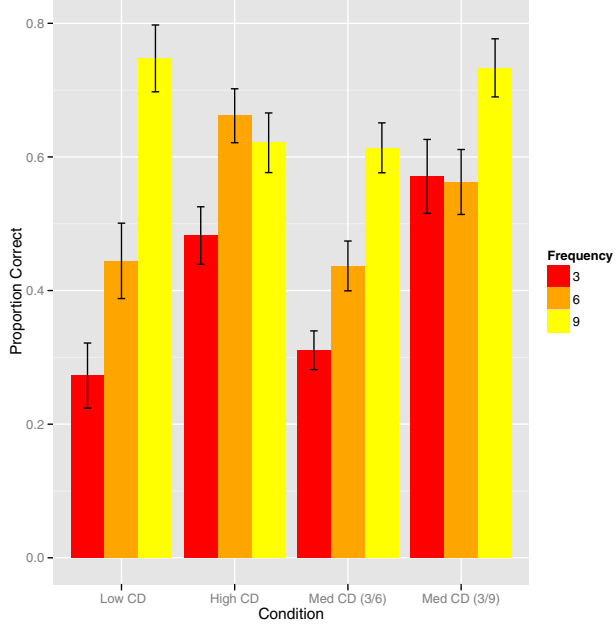

Fig. 6. Accuracy for the conditions of Experiment 3, split condition and pair subsets of differing frequency. There is a clear frequency effect in the low CD condition that disappears in the high CD condition because three- and six-frequency pair learning is bootstrapped by nine-frequency pairs. In the two mingled medium CD conditions, bootstrapping of the low-frequency pair group is evident, with learning being strongest in the 3/9 mingled condition. Error bars show ±SE.

improves learning (M_3 = .41, M_6 = .54, M_9 = .66). There was also a significant positive effect of contextual diversity (b = 0.36, OR = 1.43, Z = 3.98, $p < 0.001$), showing that increased CD benefits learning. However, there was also a significant negative interaction of Frequency and CD (b = −0.23, OR = 0.79, Z = −3.11, $p < 0.01$). This interaction is explained in detail below.

In the low CD condition, increased frequency resulted in significant increases in learning (M_3 = .26, M_6 = .45, M_9 = .75; freq 6 > 3 paired $t(33) = 3.79$, $p < .001$; freq 9 > 6 paired $t(33) = 5.6$, $p < .001$). Taken together with the results from Experiment 2, either higher frequency or higher contextual diversity can lead to better learning. However, in the high CD condition, in which all pairs were allowed to co-occur, significantly more three- and six-repetition pairs were learned than in the low CD condition (M_3 = .49, Welch $t(81.2) = 3.51$, $p < .001$; M_6 = .66, Welch $t(68.1) = 3.24$, $p < .01$), although a marginally significant fewer number of nine-repetition pairs were learned (M_9 = .63, Welch $t(82.6) = 1.84$, $p = .07$). Overall, learning was greater in the high CD condition than in the low CD condition (high CD M = .59, low CD M = .49, Welch $t(149.8) = 2.04$, $p < .05$). Thus, mixing pairs of different frequency with a higher degree of contextual diversity increases learning of the lower frequency pairs, and allows more total pairs to be learned. This is further demonstrated in the two mingled conditions which mixed two of the three frequency subsets (Fig. 5c, d).

In the 3/9 mingled condition, three-repetition pairs were learned better than in the 3/6 mingled condition (3/9 mingled M = .57, 3/6 mingled M = .31, Welch $t(60.8) = 4.21$, $p < .001$). In the two mingled conditions, learning of each nine-repetition subset remained at the same level as in the low CD condition (3/6 mingled: paired $t(35) = 1.37$, $p > .05$; 3/9 mingled: Welch $t(63.2) = .08$, $p > .05$). Thus, increasing CD helped learning, on average, by boosting acquisition of lower frequency pairs—not the higher frequency pairs. This observation also holds for the six-repetition group in the low versus high CD conditions: The high CD condition shows greater learning (low CD M_6 = .45; high CD M_6 = .66; Welch $t(68.1) = 3.24$, $p < .01$), which could be explained by the mixture of medium- and high-frequency pairs. However, in the 3/6 mingled condition, not significantly more three-repetition pairs were learned than in the low CD condition (3/6 mingled M_3 = .31 vs. low CD M_3 = .26, Welch $t(57.9) = 0.83$, $p = .41$). It seems that mingling the six-repetition pairs does not allow significant bootstrapping of the low-frequency pairs, perhaps because the six-repetition are only well-learned toward the end of training, and thus have little opportunity to be used as prior knowledge for bootstrapping. These various pairwise comparisons merely serve to bolster our intuitions for how frequency and CD, although individually beneficial, negatively interact, producing (a) much higher than expected performance for low-frequency pairs when they occur in high CD contexts (with higher frequency items), and (b) limited or no benefit for high-frequency items in high CD contexts—since these are the items that serve as the platform for bootstrapping the meaning of low-frequency items.

Frequency and CD paint only part of the picture: Environmental factors other than CD are likely affecting performance. Table 2 summarizes a few environmental statistics broken down by condition and frequency group. Although CD for the mixed 3/6 condition

Table 2
Environmental statistics and accuracy by condition and frequency in Experiment 3

Condition	Frequency	Avg. CD	Avg. Context Familiarity (CF)	Avg. Freq. of Other Co-oc. (non-zero)	Avg. Age of Exposure (AE)	Proportion Correct
Low CD	3	4	2	1.5	7	0.26
	6	4	3.5	3	4.5	0.45
	9	4.7	5	3.9	1.5	0.75
High CD	3	5	3.7	1.2	4.8	0.49
	6	8.5	4.5	1.4	4.2	0.66
	9	9.8	3.8	1.8	4.5	0.63
Mixed 3/6	3	5.5	2.9	1.1	6.5	0.31
	6	7.5	3	1.7	4.7	0.44
	9	4.7	5	3.9	1.5	0.61
Mixed 3/9	3	4.5	4.1	1.4	8	0.57
	6	4	3.5	3	4.5	0.56
	9	7.5	4.3	2.4	1.5	0.73

six-repetition pairs is higher than for the nine-repetition pairs, context familiarity—the mean frequency so far of the other stimuli appearing with a given pair—is lower and may explain the decreased performance. However, the best explanation of human word learning we present here will not be based on these summary statistics, but rather built by comparing cognitive models that are built to implement specific theories.

5. Models

The results from the three experiments showed various effects of frequency and contextual diversity in cross-situational learning. To better understand the learning mechanisms that underlie the observed behavioral effects, we use a computational modeling approach to investigate how process models accumulating statistical information trial by trial might obtain results similar to human learners. If successful, the mechanisms that the model incorporates would shed light on the human learning system. After giving a brief overview of cross-situational learning models, we compare three recent models to see which provides the best account of the bootstrapping behavior seen in Experiment 3.

Various models have been proposed for cross-situational learning, often with different goals and intuitions in mind. Models from a machine learning perspective have tried to maximize learning, without necessarily implementing psychological constraints or attempting to match human performance. For example, Yu and colleagues (Yu, 2008; Yu, Ballard, & Aslin, 2003, 2005) developed a probabilistic batch learning algorithm based on machine translation that will not show any effect of training order. The Bayesian model of Frank, Goodman, and Tenenbaum (2008) iterates multiple times over the entire training corpus to converge on a lexicon, and is thus cognitively implausible[6] for it does not produce trial-by-trial order effects, which we know to be present in human learners (e.g., Kachergis et al., 2009b).

Other models assume developmental constraints are of primary importance and follow simple rules as they hypothesize and test word meanings. Siskind's (1996) model was the first to be applied in a cross-situational learning scenario, but due to its inference based on strict constraints (e.g., mutually exclusive pairings), it is overly sensitive to noise and missing data (for an overview, see Fazly et al., 2010a). Recent hypothesis-testing models propose that learners cannot track more than one proposed meaning for a word at once, and thus only store and test a single hypothesis for each word as training proceeds (Medina, Snedeker, Trueswell, & Gleitman, 2011; Trueswell et al., 2013). Below, we describe and test the most recent of these models, the propose-but-verify model (Trueswell et al., 2013).

In another view, language learners build associations between multiple words and referents, learning an entire network of meanings with varying strength (e.g., Smith, 2000). Regier (2005) introduced an associative exemplar model of developmental word learning, but it has only been applied to simple artificial data, not experimental data. More recently, Fazly et al. (2010a) introduced a cognitively plausible incremental probabilistic model of cross-situational word learning, which has a bias to strengthen pairings that have been experienced before (i.e., a prior knowledge bias). We compare this model to a recent model introduced by Kachergis et al. (2012a), which has limited attention combined with competing biases for attending to uncertain stimuli and for pairings with prior knowledge that distinguish it from the Fazly et al. model. All three models are described below before they are applied to the data of Experiment 3, which records complicated effects of both frequency and CD within- and between-conditions, offering a challenging opportunity for modeling. Using such detailed empirical data to test models will advance our understanding of human learning mechanisms as well as other empirical phenomena (Yu & Smith, 2012).

5.1. Familiarity- and uncertainty-biased model

The model proposed by Kachergis et al. (2012a) assumes that learners do not attend equally to all possible word-object pairings and store all co-occurrences. Rather, selective storage is guided by several factors: Attention is given to pairings on the current trial, and particularly those that are familiar from previous co-occurrence. However, this factor is in competition with selective attention directed toward stimuli not already known. The latter process is based on the learner's current state of knowledge, captured by an entropy-based measurement of the uncertainty of current word-referent pairings.

Formally, given n words and n objects to be learned over a series of trials, let M be an n word $\times$ n object association matrix that is incrementally built during training. Cell $M_{w,o}$ will be the strength of association between word w and object o. Strengths are subject to general decay or forgetting but are augmented by viewing of particular pairings. Before the first trial, M is empty. On each training trial t, a set of objects O and a set of words W are presented. If there are any new words or objects are observed, new rows or columns are first added. The initial values for these new rows and columns are k, a small constant (here, 0.01).

Association strengths are generally allowed to decay, and on each new trial a fixed amount of associative weight, χ, is distributed among the associations between words and objects, and added to the (decayed) strengths. The rule used to distribute χ (i.e., attention) balances a preference for attending to unknown stimuli with a preference for strengthening already-strong associations. Consider the first time a word and referent are repeated, extra attention (i.e., χ) might be given to this pair—a bias for prior knowledge. However, as learning proceeds, novel pairings might start to stand out on trials, whereas pairings between novel objects and known words, or vice-versa, are not considered. To capture these ideas, we allocate strength using entropy (H), a measure of uncertainty that is 0 when the outcome of a variable is certain (e.g., $p(w_x|o_y) = 1$, and for all other o_z, $p(w_x|o_z) = 0$), and maximal ($log_2 n$) when every possible outcome is equally likely. In the model, on each trial the entropy of each word (and object) is calculated from the normalized row (column) vector of associations for that stimulus (i.e., $p(o|w) = M_{w,o}/\Sigma\, M_{.,o}$) like so:

$$H(w) = -\sum_{i=1}^{n} p(M_{w,i}) \cdot \log(p(M_{w,i}))$$

The update rule for adjusting and allocating strengths for the stimuli presented on a trial is:

$$M_{w,o} = \alpha M_{w,o} + \frac{\chi \cdot e^{\lambda(H(w)+H(o))} \cdot M_{w,o}}{\sum_{w \in W} \sum_{o \in O} e^{\lambda(H(w)+H(o))} \cdot M_{w,o}}.$$

In this equation, λ is a scaling parameter governing differential weighting of uncertainty and prior knowledge, α is a parameter governing forgetting, and χ is the weight being distributed. For stimuli not presented on a trial, only forgetting operates. After training, a simulated participant tested with a word w and asked to choose its associated referent from m alternatives does so in proportion to the strengths of each available referent (e.g., o) to that word ($M_{w,o}$).

5.2. Incremental probabilistic model

The Fazly et al. (2010a) model of word learning represents the meaning of each word w as a probability distribution $p(.|w)$ over the objects appearing in the corpus of trials (i.e., scene-utterance pairs). These distributions are learned incrementally as trials are experienced, much like the familiarity- and uncertainty-biased model. On a given trial presenting a set of words W and objects O, the Fazly et al. model updates the association strength of each presented word w to each presented object o in a way that more strongly associates w and o if $p(o|w)$ is high (i.e., a familiarity bias), unless some other presented word w' is already associated with o. The update rule for association scores is given by the following equation:

$$\text{assoc}(w,o) = \text{assoc}(w,o) + \frac{p(o|w)}{\sum_{w' \in W} p(o|w')}$$

where assoc(w, o) = 0 if w and o have not co-occurred. The normalizing denominator makes associations competitive, decreasing a word's alignment probability with an object if another word is already strongly associated with that object. Association scores are thus a weighted co-occurrence count, adjusted by the confidence that o is referred to by w. This is much like the familiarity bias in the Kachergis et al. model, except the update rule in that model is based on the raw association score rather than the conditional probability of o given w. Unlike the Kachergis et al. model, the Fazly et al. model does not have an uncertainty bias that encourages mapping unknown words to unknown objects; this is only accomplished by competition via the smoothed normalizing denominator. Learning performance in the Fazly et al. model is based on the association scores:

$$p(o|w) = \frac{\text{assoc}(w,o) + \lambda}{\sum_{o_j \in M} \text{assoc}(w,o) + \lambda \cdot \beta}$$

where M is the set of all objects that have been seen thus far, λ is a small smoothing constant, and β is an upper bound on the number of expected symbol types. In Fazly et al. (2010a), β was set to 8,500—the total number of words that might be expected to be learned in a developmental corpus—and $\lambda = 10^{-5}$: less than $1/\beta$ since it represents the probability of a new object going with a familiar word. Fazly et al. also thresholds the comprehension score ($p(o|w)$) at which a word w is considered to be known for object o (e.g., Fazly et al. uses $\theta = .7$). However, for our simulations this assumption is unnecessary: it does not add any flexibility in capturing the final probabilistic choice, nor does it affect the learning trajectory since comprehension scores for above-threshold words are still subject to updating.

5.3. Associative models' results

The models were fit to response-level data for all subjects and conditions simultaneously using log-likelihood as a measure of quantitative fit. Fig. 7 shows the best fit of the Fazly et al. (2010a) probabilistic incremental model ($\lambda = .017$, $\beta = 135.2$). The model shows a clear frequency effect in the Low CD condition, with higher frequency aiding learning, much like humans. However, the model shows nearly the same frequency effect in all of the other conditions, whereas people show a benefit for lower frequency pairs when they are mixed with more frequent pairs. Thus, the best fit of the Fazly et al. model does not capture the bootstrapping behavior that people show.[7] Will the uncertainty bias in the associative model enable it to match human learning?

Fig. 8 shows the best fit of the familiarity- and uncertainty-biased model to means from Experiment 3, showing that the model captures the important between- and within-condition qualitative results. Specifically, while it still captures the pure frequency effect in the

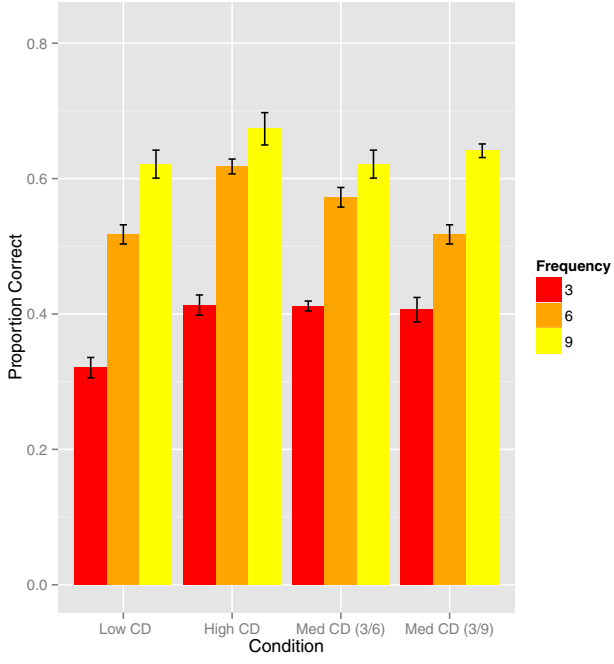

Fig. 7. Best-fitting accuracy for the Fazly et al. incremental probabilistic model to Experiment 3. The Fazly et al. model shows a clear frequency effect in the Low CD condition, much like humans. However, unlike people, the model's performance in the other conditions is much the same and is not much affected by the CD manipulations. Error bars show ±SE of item-level model performance.

Low CD condition, in the High CD condition the associative model also shows an increase in the learning of three- and six-frequency pairs, with nine-frequency pair learning remaining strong. This pattern closely matches human learning, down to the slight decrease in performance for the nine-frequency pairs, explained by the fact that more attention is going to the higher uncertainty, lower frequency pairs late in training, rather than reinforcing existing knowledge of high-frequency pairs. The associative model also qualitatively matches performance in the two mingled conditions, with the learning of low-frequency pairs being boosted when they occur in contexts with higher frequency pairs. There are some slight differences: The model shows a larger boost for three-frequency pairs in 3/6 Mingled than people show, and not as high performance for nine-frequency pairs in 3/9 Mingled. However, these differences may be in part because the same parameters ($\chi = 0.31$, $\lambda = 29.9$, $\alpha = 1.0$) were used for all participants and conditions.

Overall, the BIC of the probabilistic incremental model's best fit[8] is 4,960.3, which is worse than the BIC achieved by the associative model, 4,911.4. Thus, as well as providing a better qualitative fit, the associative model provides a better quantitative account for the data, despite having one more free parameter. The uncertainty bias seems to help explain the interaction of frequency and contextual diversity, and we can test our intuitions about how the model learns by examining the trial-to-trial learning in the model.

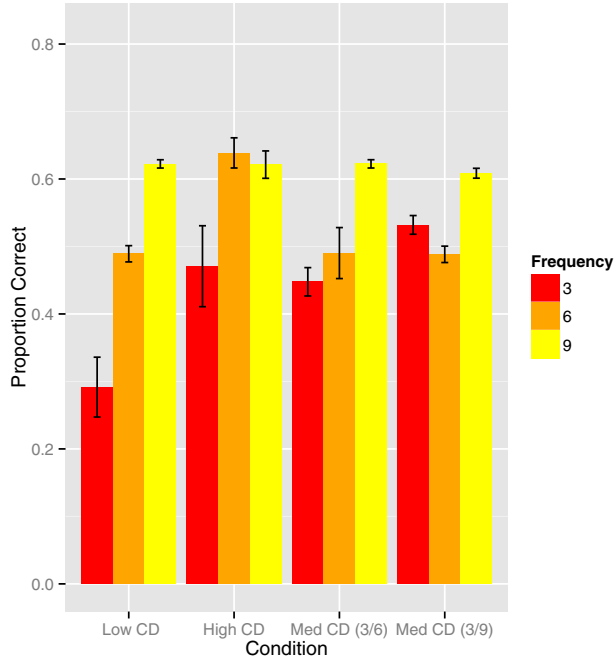

Fig. 8. The strength- and uncertainty-biased model fits qualitatively well to Experiment 3, showing both a pure frequency effect in the Low CD condition, as well as the bootstrapping of low-frequency pairs when they appear in contexts with higher frequency pairs. Error bars show ±SE of item-level model performance.

Fig. 9 shows how the model's knowledge develops over time in each condition and for each frequency group. In the Low CD condition, learners gradually—and simultaneously—learn all three frequency groups; no interaction is possible because pairs of different frequency do not co-occur. But in the High CD condition, the model first learns the high-frequency pairs, and at the end quickly learns the low-frequency pairs. Because these pairs have higher uncertainty at the end than their co-occurring high-frequency brethren, they are given more attention. That is, leveraging the uncertainty bias of the associative model, the prior knowledge of the high-frequency pairs allows the late bootstrapping of low-frequency pairs. The Appendix shows trial-by-trial learning in the Fazly et al. model for the best-fitting parameters, demonstrating that it does not capture interactions when mixing pairs of differing frequency.

Finally, we consider a recent model that is based on assumptions that are quite different than the two models evaluated above. Whereas the above models assume that multiple word-referent associations are retrieved, adjusted, and stored each time a word appears, assumptions of extremely limited memory have led other researchers to propose models that store only a single hypothesized referent for each word, and will replace this hypothesis only if it is disconfirmed (Medina et al., 2011; Trueswell et al., 2013). Elsewhere we have shown that a model implementing Medina et al.'s assumptions cannot account for the range of learning trajectories shown by individual cross-situational learners (Kachergis, Yu, &

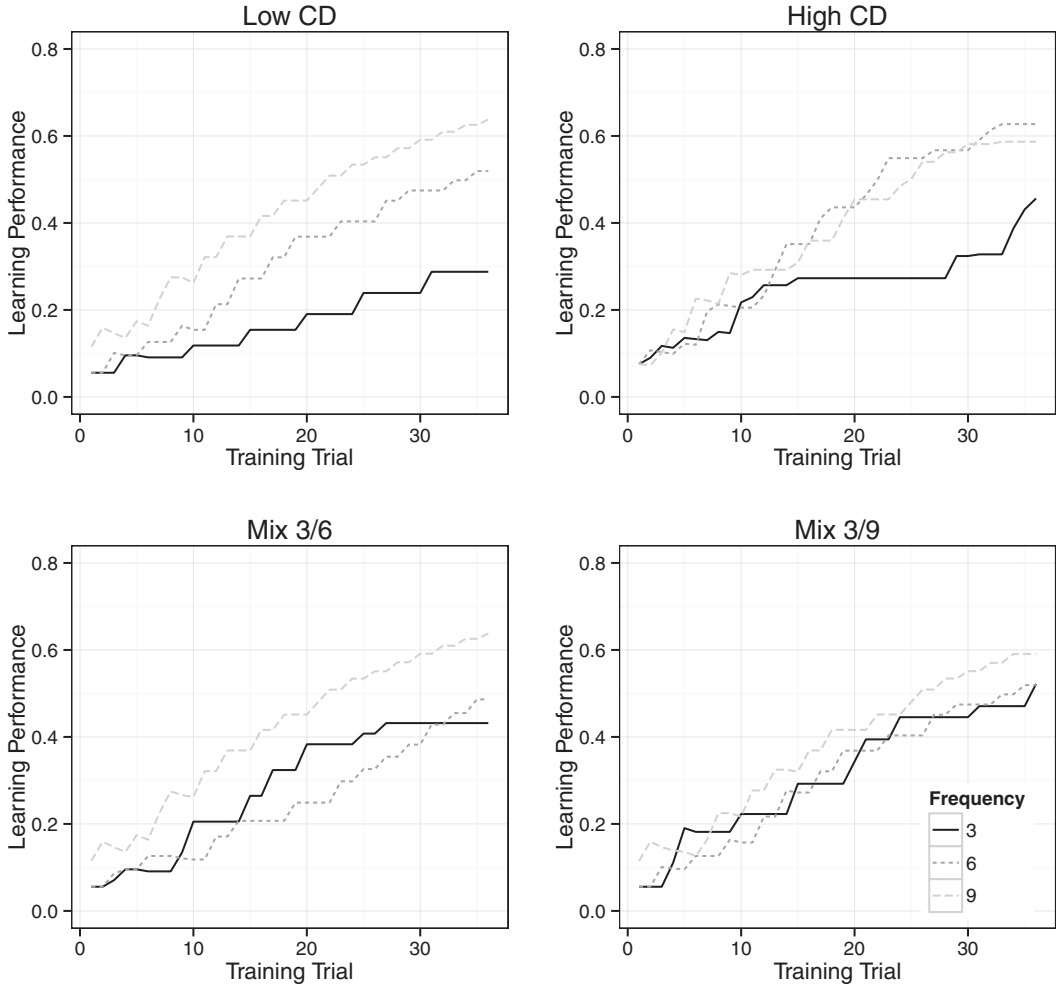

Fig. 9. The strength- and uncertainty-biased model's knowledge development by frequency in the conditions of Experiment 3 for the best-fitting parameters. In the High CD condition, notice how the model first learns the high-frequency pairs, and only later quickly learns the low-frequency pairs, which have higher uncertainty. Rather than requiring us to interrupt human learners at various points of training to view the learning process, the model allows us to predict behavior—bootstrapping, in this case.

Shiffrin, 2012b). We now investigate whether the latest hypothesis-testing model (Trueswell et al., 2013) can account for the effects of varied frequency and contextual diversity.

5.4. Propose-but-verify model

The assumptions of the hypothesis-testing approach are outlined in Medina et al. (2011):

(i) learners hypothesize a single meaning based on their first encounter with a word; (ii) learners neither weight nor even store backup alternative meanings; and (iii) on

later encounters, learners attempt to retrieve this hypothesis from memory and test it against a new context, updating it only if it is disconfirmed. Thus, they do not accrue a "best" final hypothesis by comparing multiple episodic memories of prior contexts or multiple semantic hypotheses. (p. 3)

The propose-but-verify model introduced by Trueswell et al. (2013) follows similar assumptions, positing that learners store a list of word-object pairs with only up to a single object stored for each word. The model begins with an empty list, and on each training trial the presentation of each word w causes the attempted retrieval of the stored hypothesis o_h, successful with probability α_o. If retrieval of o_h fails (probability $1-\alpha_o$), the hypothesis w-o_h is forgotten (i.e., erased from the list). If o_h is retrieved and is present on the current trial, the recall probability α is strengthened by α_r. If o_h is retrieved but is not present on the current trial, the hypothesis w-o_h is removed. For any words remaining on the trial without a hypothesis, new hypothesized objects are chosen randomly without replacement from the set of objects that are not already part of a hypothesis. This manner of selection effects a local mutual exclusivity constraint which could quickly bootstrap the meaning of a novel word-object pair, though by a different means than the associative model. Testing in the propose-but-verify model is straightforward:

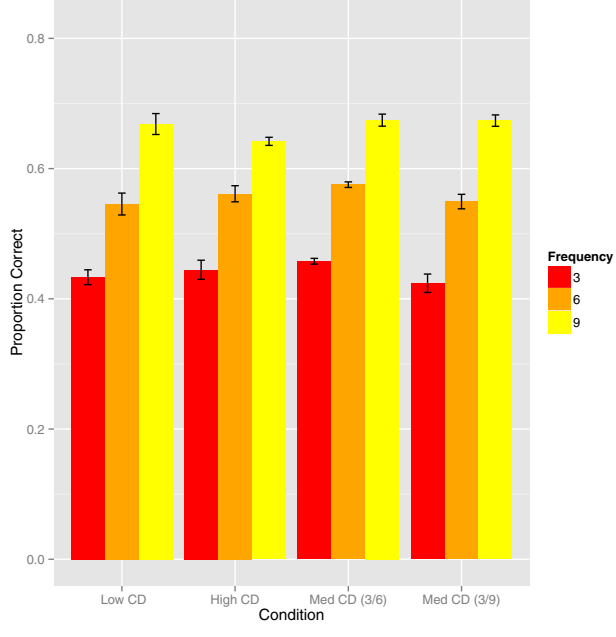

Fig. 10. Best-fitting accuracy for the propose-but-verify (Trueswell et al., 2013) model to Experiment 3. The hypothesis-testing model does not capture much of a frequency effect in the Low CD condition, nor the boost in performance for three- and six-frequency pairs in the High CD condition, relative to the Low CD condition. In fact, the High CD condition shows improvement only for the nine-frequency pairs, unlike the human data. Error bars show ±SE of item-level model performance.

For each word, the model chooses the hypothesized object for each word. For words with no hypothesis, objects are randomly chosen from the set of objects that are not part of a hypothesis.

The propose-but-verify model's two parameters—initial recall probability (α_o) and recall reinforcement (α_r)—were fit to the means of each condition and frequency grouping in Experiment 3. The best-fitting parameter values were $\alpha_o = 0.06$ and $\alpha_r = 0.31$, achieving a BIC of 4,963.3, worse than both the Fazly et al. probabilistic incremental model (BIC = 4,960.3) and the associative model (BIC = 4,911.4). To predict how well these models will generalize to future data, we examined cross-validation fits to split halves of the data, finding that the Kachergis et al. associative model still has the best performance and generalizes as well as the other models (see Appendix for details). This investigation also suggested that none of the models are greatly overfitting, and it provided a range of reasonable parameter values. In the Appendix, we also explore correlations between human and the models' item-level performance, as well as with several statistics of the training input, including frequency, CD, and two measures proposed by Fazly et al. (2010b): context familiarity and age of exposure. We found that the Kachergis et al. model best accounts for item-level human performance and is most correlated with context familiarity (i.e., the mean number of appearances of the other stimuli in a given pair's context). On the other hand, the Trueswell et al. and Fazly et al. models, are highly correlated with each other, with performance differences strongly determined by the environmental factor of pair frequency. Please see the Appendix for additional details.

6. General discussion

Our experimental manipulations suggest that three factors that significantly determine the success of cross-situational statistical learning are word-referent frequency, contextual diversity, and the degree of within-trial ambiguity. These three factors are related: Picking values for two of the factors somewhat constrains the value of the third. For example, consider a pair that is to appear six times during a three pairs/trial training set. On each of the six trials it occurs on, two other pairs must also appear. These 12 pairs may each be distinct, or some particular pairs may appear more often than once. If two pairs always co-occur, the "correct" word-referent pairs cannot be disambiguated. However, if one of the two pairs is learned prior to the appearance of the other (as may be the case for a high-frequency pair), then the other may be learned more easily, since the prior knowledge of the frequent pair reduces the uncertainty about that word and object, thus directing attention more to the uncertain objects. Thus, it is not only the diversity of the contexts in which a pair appears, but also the familiarity of the stimuli appearing with a pair that determines the likelihood of learning that pair (Fazly et al., 2010b). Context familiarity—though not explicitly manipulated—turned out to be the environmental factor most correlated with item-level human performance; stronger than frequency or contextual diversity (see Appendix).

Experiment 1 demonstrated that although varied frequency can result in increased performance for more frequent pairs, it is also possible for varied frequency to perplexingly

yield equal performance, perhaps as a result of contextual diversity and familiarity effectively reducing within-trial ambiguity. Experiment 2 showed that increased contextual diversity improves learning for equal-frequency pairs. Moreover, pairs with greater within-trial ambiguity were learned less well—despite greater contextual diversity. Experiment 3 confirmed that more frequent pairs are learned more often, even when contextual diversity is controlled. In addition, increasing the contextual diversity of two groups of different frequency by allowing these groups to co-occur augmented learning of the less frequent of these groups. Indeed, the highest learning performance observed was in conditions with varied frequency and high contextual diversity. Intriguingly, these two characteristics that yield high performance are also embedded in real-world learning environments: Words in any natural language have a skewed frequency distribution (Zipf, 1949), and naturalistic learning situations are highly complex, with many co-occurring words, events and objects (e.g., Medina et al., 2011; although perhaps not as complex from a child's perspective: Yurovsky, Smith, & Yu, 2013).

Varied frequency and contextual diversity seem to make situations more complex, but our results suggest that they facilitate statistical learning. Much structure is present in our world: words, referents, and their contexts vary in frequency, diversity, and the composition of the situation. The above experiments demonstrate that human learners are sensitive to these different kinds of regularities. Varied frequency may seem to be a natural candidate as the most important factor in cross-situational word learning, as more appearances yields more opportunities to acquire the appropriate association. However, if that pair always appears with only a few other pairs, or simultaneously appears with many other pairs, each learning opportunity is worth very little: Context is critical. Disambiguating the proper pairings via high contextual diversity or context familiarity and a reasonably small degree of within-trial ambiguity enables learning to proceed with ease. Indeed, simulation studies confirm that cross-situational learning under non-uniform frequency distributions is robust when contextual uncertainty is kept low (e.g., for Zipfian distributions: Vogt, 2012), and more important if learners apply mutual exclusivity (Reisenauer, Smith, & Blythe, 2013). Because the presence of known high-frequency pairs reduces within-trial ambiguity, highly ambiguous situations containing some familiar referents become feasible learning opportunities. This process of bootstrapping via familiar contexts may account for the rapid acquisition of vocabulary in infants, who are known to learn frequent nouns earlier than less-common nouns (Goodman, Dale, & Li, 2008). Once known, the ubiquitous nouns make possible the rapid acquisition of infrequent nouns. Indeed, this is the account provided by the uncertainty- and familiarity-biased associative model (Kachergis et al., 2012a). This model uses uncertainty to quickly learn low-frequency pairs at the end of training, as they appear with well-known high-frequency pairs. The Fazly et al. model, lacking an uncertainty bias, does not show this bootstrapping behavior while matching human performance levels. The Trueswell et al. hypothesis-testing model is able to capture pure frequency effects by having a very low initial probability of forming a hypothesis ($\alpha_o = 0.06$), accompanied by a fairly large reinforcement probability when it is encountered again ($\alpha_r = 0.31$), but does not show bootstrapping of rare pairs from high-frequency ones. Indeed, the Fazly et al. and Trueswell et al. models, despite seemingly different

mechanisms, show strongly correlated item-level predictions, being largely driven by co-occurrence frequency but not context effects (see Appendix).

Rather, the human behavioral results suggest a learning system that does not learn independent associations between individual words and referents based on mere frequency, but one that rather learns a system of associations where context is critical (see Smith, 2000; Yu, 2008). In such a system, a single word-referent pairing is correlated with all the other pairings that share the same word and all the other pairings that share the same referent, which are in turn correlated with more word-referent pairs—an entire system of them. We contend that the improvement in statistical word learning is in part due to the recruitment of accumulated latent lexical knowledge, which is used to learn subsequently appearing pairs. Indeed, the strongest environmental factor predicting item-level human performance was an item's average context familiarity, showing that the system of other associations predicts which pairs will be acquired. The associative model, which learns associations between all co-occurring words and objects incrementally, leverages competing prior knowledge and uncertainty biases to show an even stronger correlation with item-level human performance (see Appendix). Such a learning system is consistent with the finding that children who are slow to learn language have a less well-connected semantic network than normally developing children (Beckage, Smith, & Hills, 2011): Late talkers cannot bootstrap the meaning of rare words if they do not have the high frequency contextually diverse words mastered. Finally, although these experiments investigate word-object associations, we suggest that our findings may well generalize to other domains, for our model assumes domain-general mechanisms that can reproduce many associative learning behaviors (see Kachergis, 2012).

Acknowledgments

This article is an extended and updated version of a paper that appeared in the *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society* (Kachergis, Yu, & Shiffrin, 2009a). Special thanks to Gregory E. Cox and Anselm Rothe for useful discussions.

Notes

1. Data from two subjects were excluded after it was found that their average performance in every condition was below chance (chance in an 18AFC test is .056). This did not change the outcome of any statistical tests.
2. The estimated coefficients (b) are interpretable as log-odds, but they can also be transformed to an odds ratio ($OR = e^b$).
3. The number of subjects varied between conditions because several groups of participants were collected on different overlapping subsets of the conditions. Specifically, 36 participants performed the low/medium CD condition, and 26 did both the low/medium CD and four pairs/trial conditions. An additional 51 subjects in

the four pairs/trial condition were included from a study where it was used as a control condition.

4. The conditions had unequal numbers of participants because initially only Low CD, High CD, and the mingled 3/6 conditions were created and run (27 participants). The 3/9 condition was created and added for seven participants, before the experiment was deemed too long. The Low CD was dropped as it was being included in a learning trajectory study (Kachergis, Yu, & Shiffrin, 2014), so the High CD and two mingled conditions were run on 33 additional subjects, in order to have sufficient points for within-subject comparisons across conditions.
5. Model syntax: Correct ~ Freq × CD + (Freq × CD|Subject). This maximal model was chosen based on the experiment's design (see Barr, Levy, Scheepers, & Tily, 2013). More complex models, such as those treating Frequency and CD as factors, failed to converge.
6. Note that Frank et al. (2008) calls it a computational-level, rather than algorithmic-level model, and thus it may not be expected to produce human-level performance and effects of factors such as frequency and CD. In fact, the Frank et al. (2008) model usually converges on perfect performance for the trial orderings in this study.
7. Fazly, Ahmadi-Fakhr, Alishahi, and Stevenson (2010b) presented model fits that qualitatively match those of Experiment 3—at least for Low and High CD conditions. However, the results shown in that paper have much higher overall performance than the human data, and do not optimize the model's overall quantitative fit to the human data, as do the results of Fig. 7 in the present paper, which show their model's optimal quantitative fit. Model comparison is more fittingly based on quantitative fits to human performance data than more subjective qualitative fits, although of course a qualitative—and explanatory—match is also desirable. Finally, it is important to note that although the trial orderings used in the Fazly et al. (2010b) simulations were generated according to the design of Experiment 3, they were not exactly the trial orderings used here.
8. A grid search using Fazly et al.'s model implementation confirmed our best fit.

References

Adelman, J. S., & Brown, G. D. A. (2008). Modeling lexical decision: The form of frequency and diversity effects. *Psychological Review*, *115*(1), 214–227.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

Bates, D., Maechler, M., Bolker, B., & Walker, S (2015). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-8. Available at http://CRAN.R-project.org/package=lme4. Accessed August 1, 2015.

Beckage, N., Smith, L., & Hills, T. (2011). Small worlds and semantic network growth in typical and late talkers. *PLoS ONE*, *6*(5), e19348. doi:10.1371/journal.pone.0019348

Fazly, A., Ahmadi-Fakhr, F., Alishahi, A., & Stevenson, S. (2010b). Cross-situational learning of low frequency words: The role of context familiarity and age of exposure. In S. Ohlsson & R. Catrambone

(Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2615–2620). Austin, TX: Cognitive Science Society.

Fazly, A., Alishahi, A., & Stevenson, S. (2010a). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, *34*, 1017–1063.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2008). A Bayesian framework for cross-situational word-learning. In J. C. Platt et al. (Eds.), *Advances in neural information processing systems 20* (pp. 457–464). Cambridge, MA: MIT Press.

Gillette, J., Gleitman, L. R., Gleitman, H., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*, 135–176.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*(1), 3–55.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*(03), 515–531.

Hills, T., Maouene, J., Riordan, B., & Smith, L. (2010). The associative structure of language and contextual diversity in early language acquisition. *Journal of Memory and Language*, *63*, 259–273.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

Kachergis, G. (2012). Learning nouns with domain-general associative learning mechanisms. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 533–538). Austin, TX: Cognitive Science Society.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2009a). Frequency and contextual diversity effects in cross-situational word learning. In N. Taatgen, H. van Rijn, J. Nerbonne, & L. Schomaker (Eds.), *Proceedings of 31st Annual Meeting of the Cognitive Science Society* (pp. 755–760). Austin, TX: Cognitive Science Society.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2009b). Temporal contiguity in cross-situational statistical learning. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1704–1709). Austin, TX: Cognitive Science Society.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2010). Cross-situational statistical learning: Implicit or intentional? In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2362–2367). Austin, TX: Cognitive Science Society.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2012a). Cross-situational word learning is better modeled by associations than hypotheses. In *IEEE conference on development and learning/EpiRob 2012*. San Diego, CA: IEEE.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2012b). An associative model of adaptive inference for learning word-referent mappings. *Psychonomic Bulletin & Review*, *19*(2), 317–324.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2013). Actively learning object names across ambiguous situations. *Topics in Cognitive Science*, *5*(1), 200–213.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2014). Developing semantic knowledge through cross-situational learning. In P. Bello et al. (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 2568–2573). Austin, TX: Cognitive Science Society.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*, 121–157.

Medina, T., Snedeker, J., Trueswell, J., & Gleitman, L. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of the Sciences*, *108*, 1–6. doi:10.1073/pnas.1105040108

Merriman, W. E., & Bowman, L. L. (1989). The mutual exclusivity bias in children's word learning. *Monographs of the Society for Research in Child Development*, *54*(3/4), 1–129.

R Development Core Team (2010). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. Available at: http://www.R-project.org/. Accessed December 1, 2013.

Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive Science*, *29*, 819–865.

Reisenauer, R., Smith, K., & Blythe, R. A. (2013). Statistical mechanics of lexicon learning in an uncertain and nonuniform world. *Physical Review Letters*, *110*, 258701.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*, 39–91.

Smith, L. B. (2000). How to learn words: An associative crane. In R. Golinkoff & K. Hirsh-Pasek (Eds.), *Breaking the word learning barrier* (pp. 51–80). Oxford, UK: Oxford University Press.

Smith, L. B. (2001) How domain-general processes may create domain-specific biases. In M. Bowerman & S. Levinson (Eds.) *Language Acquisition and Conceptual Development*. Cambridge University Press.

Smith, K., Smith, A. D. M., & Blythe, R. A. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, *35*(3), 480–498.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*, 1558–1568.

Suanda, S. H., & Namy, L. L. (2012). Detailed behavioral analysis as a window into cross-situational word learning. *Cognitive Science*, *36*(3), 545–559.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*, 126–156. doi:10.1016/j.cogpsych.2012.10.001

Vogt, P. (2012). Exploring the robustness of cross-situational modelling under Zipfian distributions. *Cognitive Science*, *36*(1), 726–739.

Yu, C. (2008). A statistical associative account of vocabulary growth in early word learning. *Language Learning and Acquisition*, *4*(1), 32–62.

Yu, C., Ballard, D. H., & Aslin, R. N. (2003). The role of embodied intention in early lexical acquisition. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th Annual Meeting of Cognitive Science Society* (pp. 1293–1298).

Yu, C., Ballard, D. H., & Aslin, R. N. (2005). The role of embodied intention in early lexical acquisition. *Cognitive Science*, *29* (6), 961–1005.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*, 414–420.

Yu, C., & Smith, L. B. (2012). Modeling cross-situational word-Referent learning: Prior questions. *Psychological Review*, *119*(1), 21–39.

Yurovsky, D., Smith, L. B., & Yu, C. (2013). Statistical word learning at scale: The baby's view is better. *Developmental Science*, *16*, 959–966.

Yurovsky, D., & Yu, C. (2008). Mutual exclusivity in cross-situational statistical learning. In B. C. Love, K. McRae & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 715–720). Austin, TX: Cognitive Science Society.

Yurovsky, D., Yu, C., & Smith, L. B. (2013). Competitive processes in cross-situational word learning. *Cognitive Science*, *37*, 891–921.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley.

Appendix A:

A.1. Correlation of model performance and environmental factors with behavior

To examine which environmental factors (e.g., frequency and CD) the models are sensitive to in comparison to those that best explain human performance, we looked at the correlations for the 72 word–object pairs in Experiment 3 between several item-level factors and each model's item-level performance using the best-fitting group parameters. In addition to pair frequency (3, 6, or 9) and contextual diversity (CD; range: 3–11), we used two statistics

proposed by Fazly et al. (2010b). Age of Exposure (AE) is operationalized as the trial index where a pair first appears (range: 1–15). Of course, higher frequency pairs are more likely to appear earlier, meaning AE will likely be negatively correlated with frequency—but the two may have unique impacts on learning. Another statistical measure of the training input,

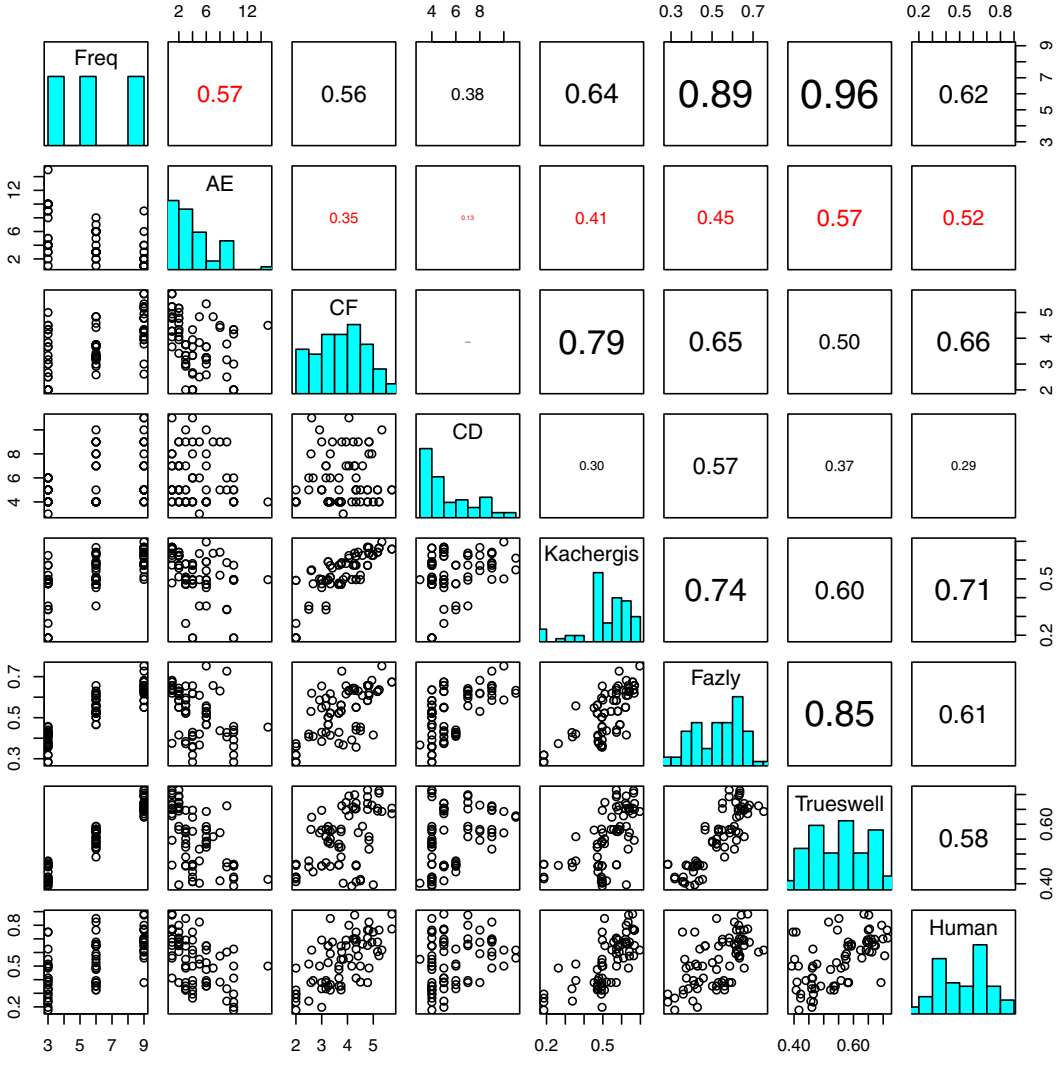

Fig. A1. A scatterplot matrix of item-level statistics—including frequency (Freq.), age of exposure (AE), context familiarity (CF), and contextual diversity (CD)—compared to performance of the three models (Kachergis, Fazly, and Trueswell) and of humans for the 72 word–object pairs of Experiment 3. Correlation coefficients are shown in the upper right half (*red* values are negative). Human performance is most highly correlated with the Kachergis et al. model, followed by CF and frequency. The Trueswell et al. and Fazly et al. models are highly correlated with frequency, and with each other. The Kachergis et al. model's performance is most correlated with CF, followed by frequency.

a word's Context Familiarity (CF) is defined as the average familiarity (i.e., co-occurrences) of the pairs appearing with a given word, across all of its occurrences (range: 2.0–5.7). Items with higher CF should be more likely to be acquired (Fazly et al., 2010b).

The distribution and correlation between these item-level measures (Freq., AE, CF, and CD) and the performance of each model (Kachergis, Fazly, and Trueswell) and humans (Human) on the corresponding items are shown in Fig. A1. Of the three models, the Kachergis et al. model has the strongest item-level correlation with human performance (r = .72 compared to Fazly et al.'s r = .61 and Trueswell et al.'s r = .58). The Trueswell et al. and Fazly et al. models are in fact highly-correlated with each other

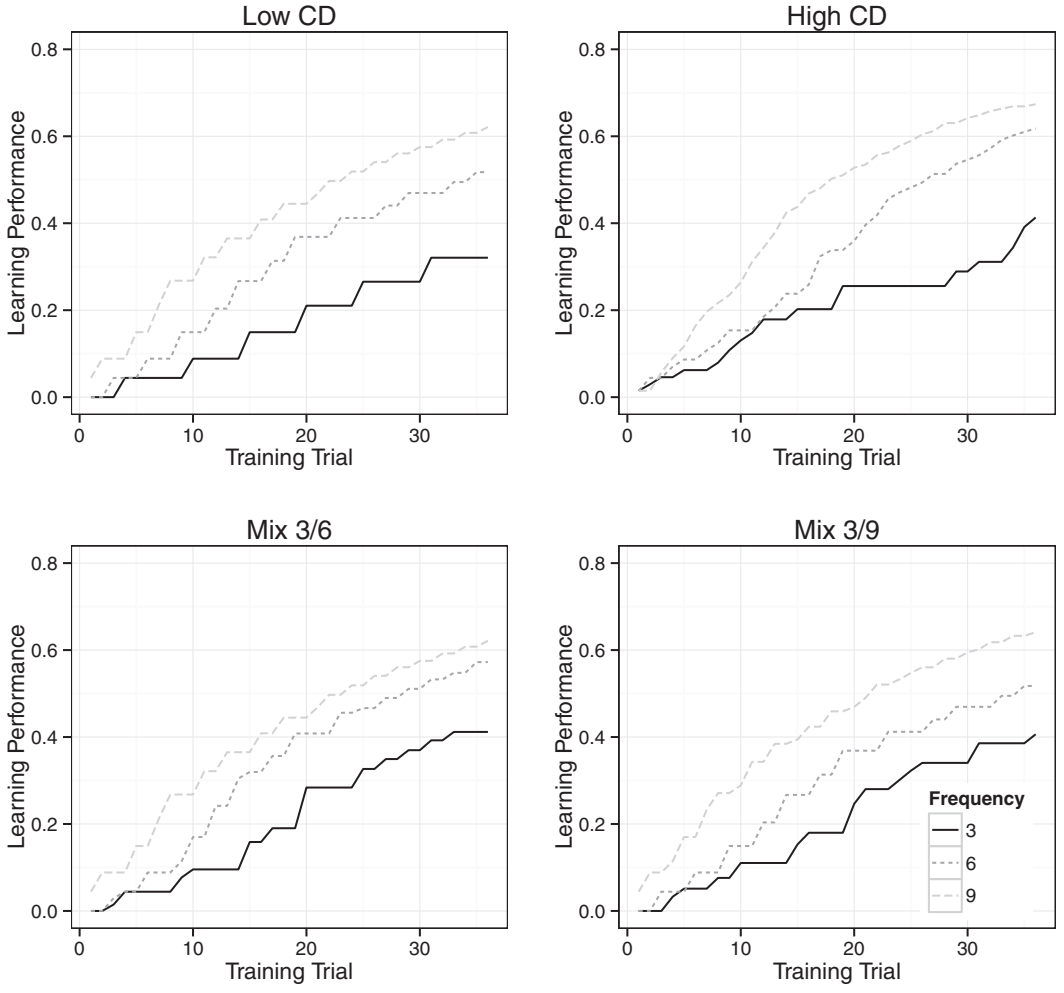

Fig. A2. The Fazly et al. probabilistic incremental model's knowledge development by frequency in the conditions of Experiment 3 for the best-fitting parameters. Regardless of condition, the model learns the different frequency groups nearly in tandem, and does not show late-stage bootstrapping of low from higher frequency pairs.

(r = .85), with both of their performance being strongly correlated with frequency (Fazly et al.'s r = .89 and Trueswell et al.'s r = .96; compare to Kachergis et al.'s r = .64). Aside from the Kachergis et al. model, human performance is most correlated with CF (r = .66), followed by frequency (r = .62). CF, which measures how familiar the contexts that a word appears in, is more strongly correlated with human performance than CD (r = .29), which measures the dispersion of an item's appearance across contexts, but not their familiarity. Indeed, CF is the factor that best captures the behavior of the Kachergis

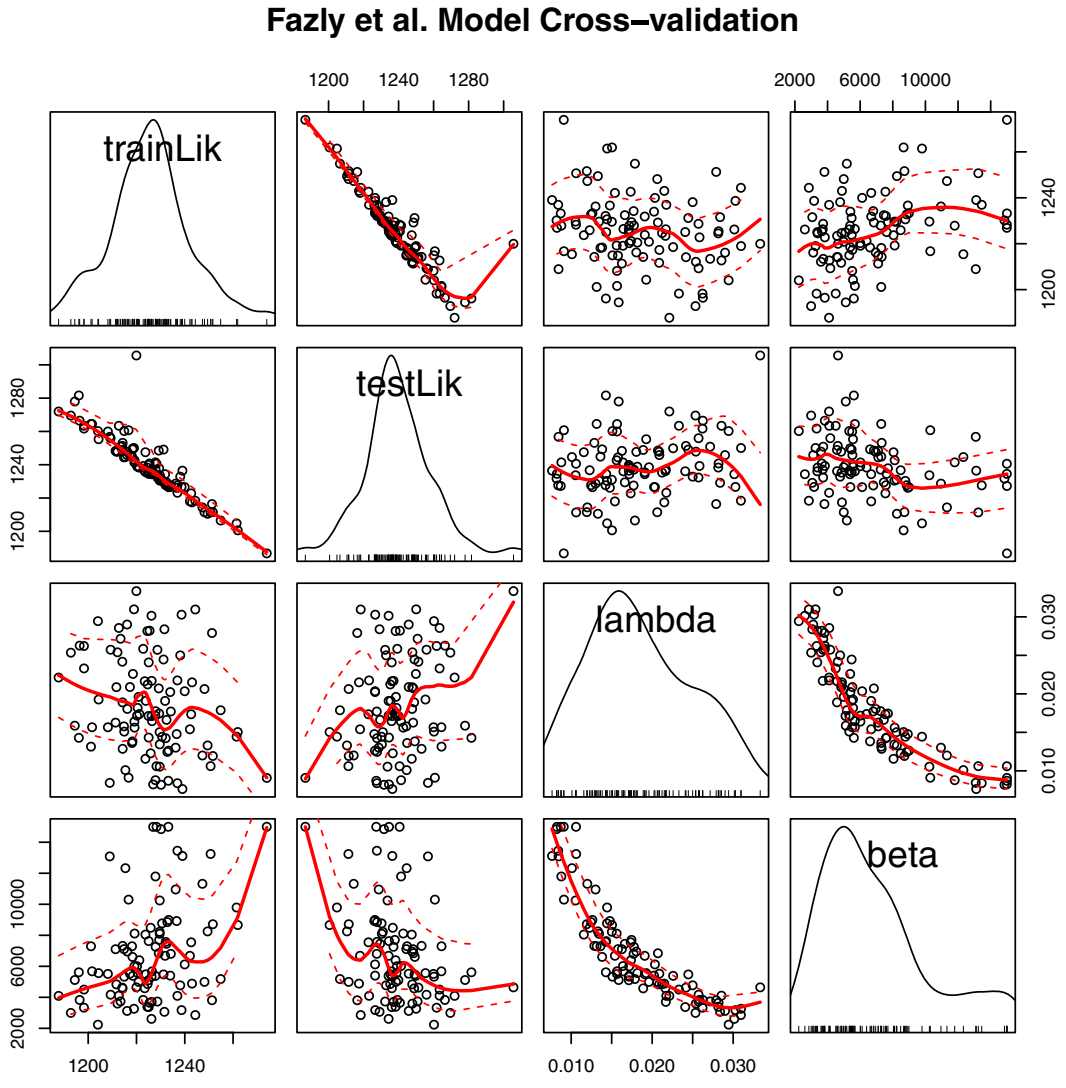

Fig. A3. Scatterplot of the Fazly et al. probabilistic incremental model's best-fitting parameters versus training data fit (trainLik) and the held-out testing data fit (testLik) for the 100 random cross-validation subsets. The lambda and beta parameters seem to trade off, suggesting that the model could be reparameterized.

et al. model (r = .79). Although many of the statistical measures are related to each other (e.g., AE and frequency's negative correlation), seeing which factors correlate most with each model—and with human performance—gives us a sense of what effects are produced by the mechanisms when applied to structured statistical input. In summary, the Kachergis et al. model is the best explanation of the human data, seemingly by capturing context familiarity and frequency effects. The Fazly et al. and Trueswell et al. models fare less well, capturing essentially the frequency effects.

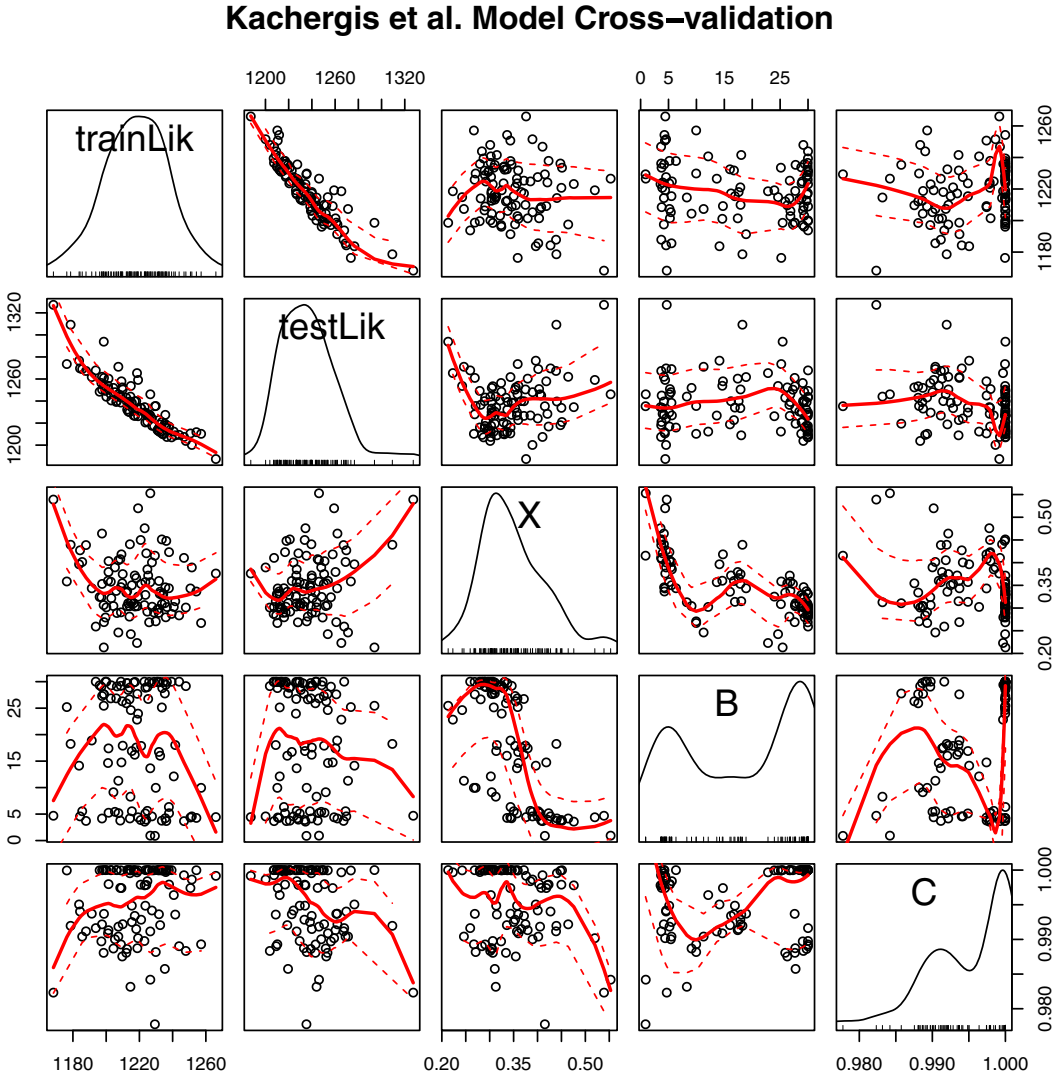

Fig. A4. Scatterplot of the Kachergis et al. associative model's best-fitting parameters versus fit to the training data (trainLik) and to the held-out testing data (testLik) for the 100 random subsets of half the data. (*N.b.*: Parameter α is C in the figure.)

A.2. Knowledge development in the Fazly et al. model

Fig. A2 shows the Fazly et al. probabilistic incremental model's knowledge development in the conditions of Experiment 3, using the best-fitting parameters (λ = .017, β = 135.2). Unlike the Kachergis et al. associative model, this model does not seem to bootstrap the meaning of low-frequency words from co-occurrence with higher frequency items. Instead, low-frequency pairs are learned nearly in tandem with higher frequency pairs in every condition, and performance for them levels off during the first half of training.

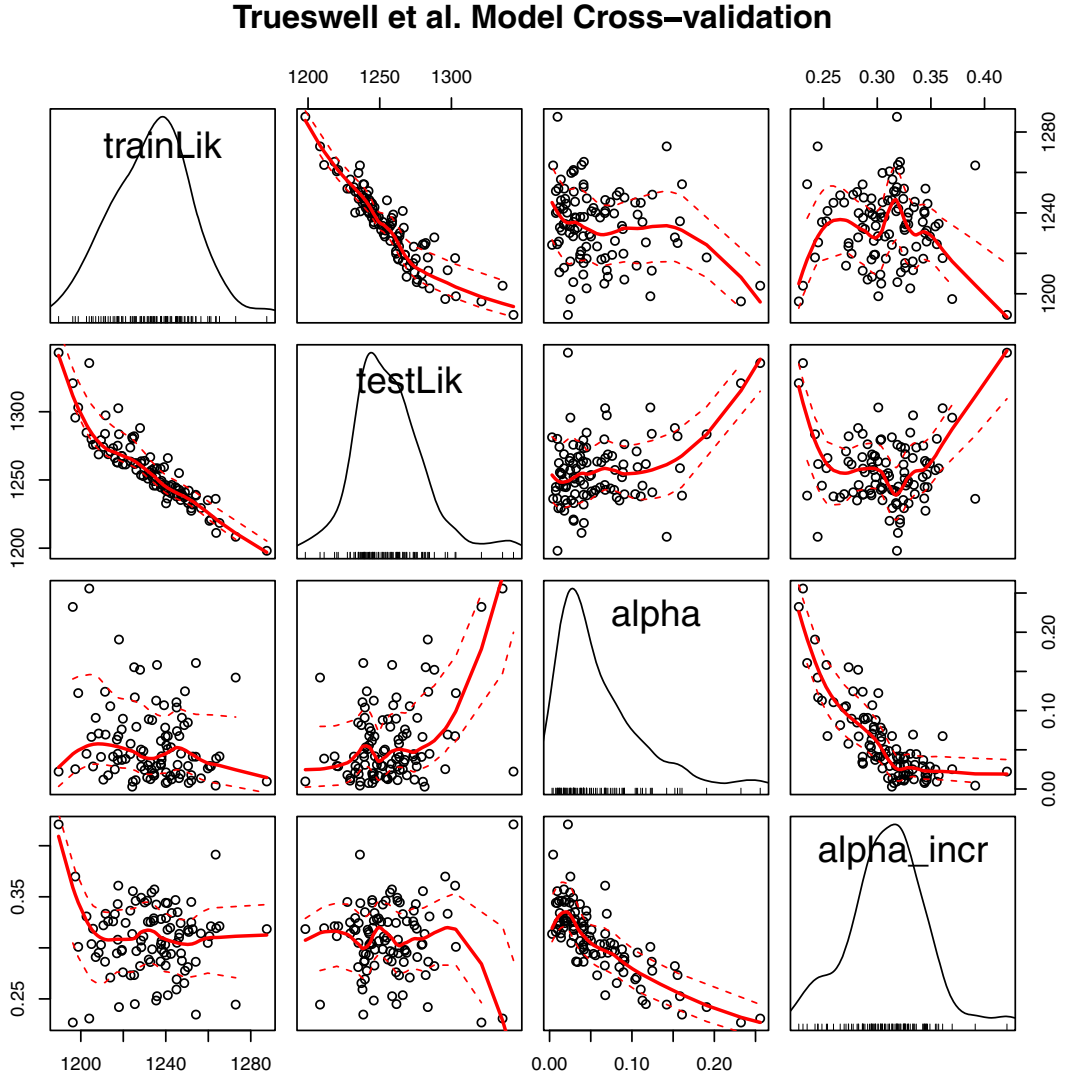

Fig. A5. Scatterplot of the Trueswell et al. propose-but-verify model's best-fitting parameters versus fit to the split halves of training data (trainLik) and to test data (testLik) for 100 random subsets.

A.3. Cross-validation of parameters

To see how well the models can be expected to generalize, we tested them via cross-validation. For each of 100 iterations, we fit each model's parameters to a randomly selected half of Experiment 3's data, and used the remaining half of the data to test how well the model and parameters generalize. We sought to minimize negative log-likelihood of the training data, and we hoped to see a similarly low log-likelihood on the held-out test data; a large increase would indicate that the model is overfitting the training data, and failing to generalize to the held-out data—and thus future data. The distribution of parameters values and the models' fit to the training and validation testing data are shown in Fig. A3 (Fazly et al. model), Fig. A4 (Kachergis et al. model), and Fig. A5 (Trueswell et al. model).

The mean (and median) values for the parameters of the Fazly et al. model are $\lambda = .018$ (.017) and $\beta = 6{,}773$ (5,675), with mean negative log-likelihood fits (lower is better) of 1,225 to the training data and 1,239 for the held-out test data. The mean (median) values for the Kachergis et al. model's parameters are $\chi = .347$ (.336), $\beta = 18.31$ (18.6), and $\alpha = .995$ (.998), with mean negative log-likelihoods of 1,218 for the training data and 1,237 for the test data. The mean (median) values for the parameters of the Trueswell et al. model are $\alpha = .066$ (.056) and $\alpha_r = .293$ (.296), and the mean negative log-likelihoods were 1,236 for the training data and 1,251 for the validation data. Overall, the Kachergis et al. model had the best training and validation testing likelihoods, followed by the Trueswell et al. model and then the Fazly et al. model. The prediction error of all three models was in roughly the same range, with means of 18.6, 15.2, and 13.9 for the Kachergis, Trueswell, and Fazly et al. models, respectively. These differences were not significantly different. Finally, the mean best-fitting parameters for all three models found for the cross-validation subsets were similar to those found when fitting the entire dataset. We hope other researchers will find these parameter distributions informative when evaluating future studies.