Appendix to *A bootstrapping model of frequency and context effects in word learning*

George Kachergis, Chen Yu, and Richard M. Shiffrin

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, Ahmadi-fakhr, Alishahi, & Stevenson (2010). 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, 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, Ahmadi-fakhr, Alishahi, & Stevenson, 2010).

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 Figure 1. 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 (r = .85), with both of their performance being strongly correlated with frequency (Fazly *et al.* r = .89 and Trueswell *et al.* r = .96; compare to Kachergis *et al.* 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 *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.

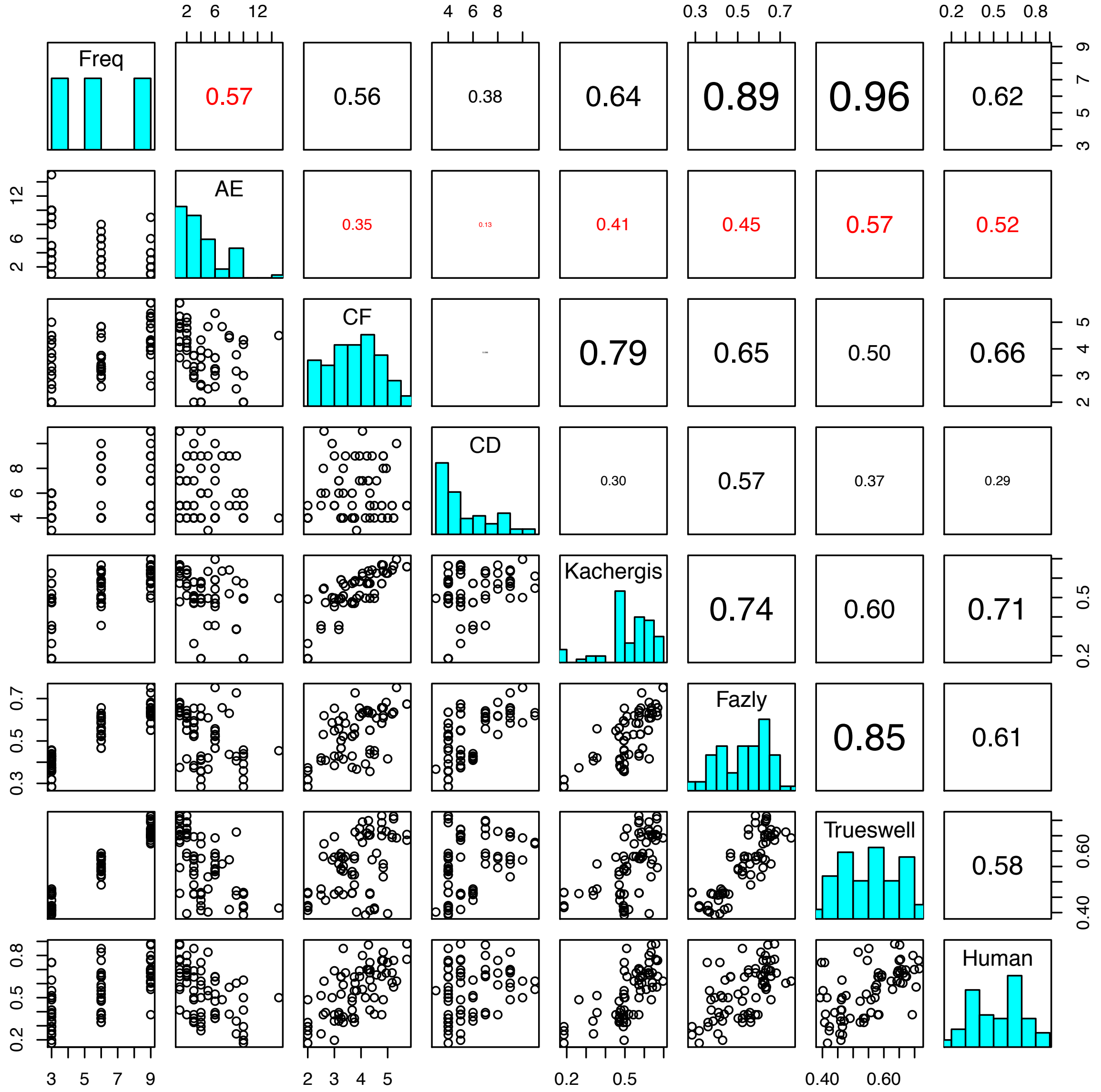

Figure 1. A scatterplot matrix of item-level statistics—including frequency (Freq.), age of exposure (AE), context familiarity (CF), and contextual diversity (CD)—compared to performance of the three models (Kachergis, Fazly, and Trueswell) and of humans for the 72 word-object pairs of Experiment 3. 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.

2. Knowledge Development in the Fazly *et al.* Model

Figure 2 shows the Fazly *et al.* probabilistic incremental model's knowledge development in the conditions of Experiment 3, using the best-fitting parameters ($\lambda = .017$, $\beta = 135.2$). 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.

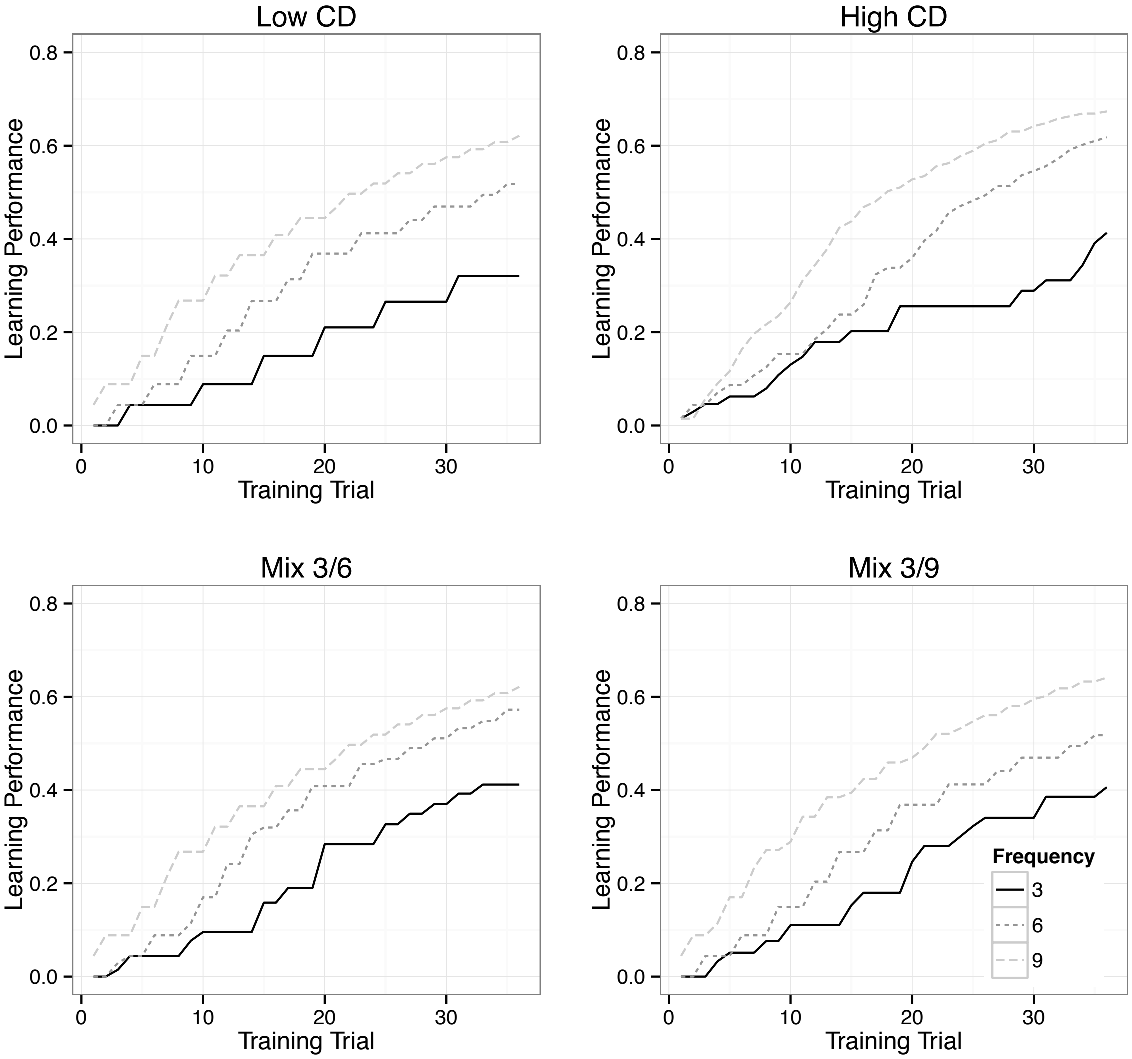

Figure 2. The Fazly *et al.* probabilistic incremental model's knowledge development by frequency in the conditions of Experiment 3 for the best-fitting parameters. Regardless of condition, the model learns the different-frequency groups nearly in tandem, and does not show late-stage bootstrapping of low from higher-frequency pairs.

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 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 Figure 3 (Fazly *et al.* model), Figure 4 (Kachergis *et al.* model), and Figure 5 (Trueswell *et al.* model).

The mean (and median) values for the parameters of the Fazly *et al.* model are λ = .018 (.017) and β = 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 χ = .347 (.336), β = 18.31 (18.6), and α = .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 α = .066 (.056) and α_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.

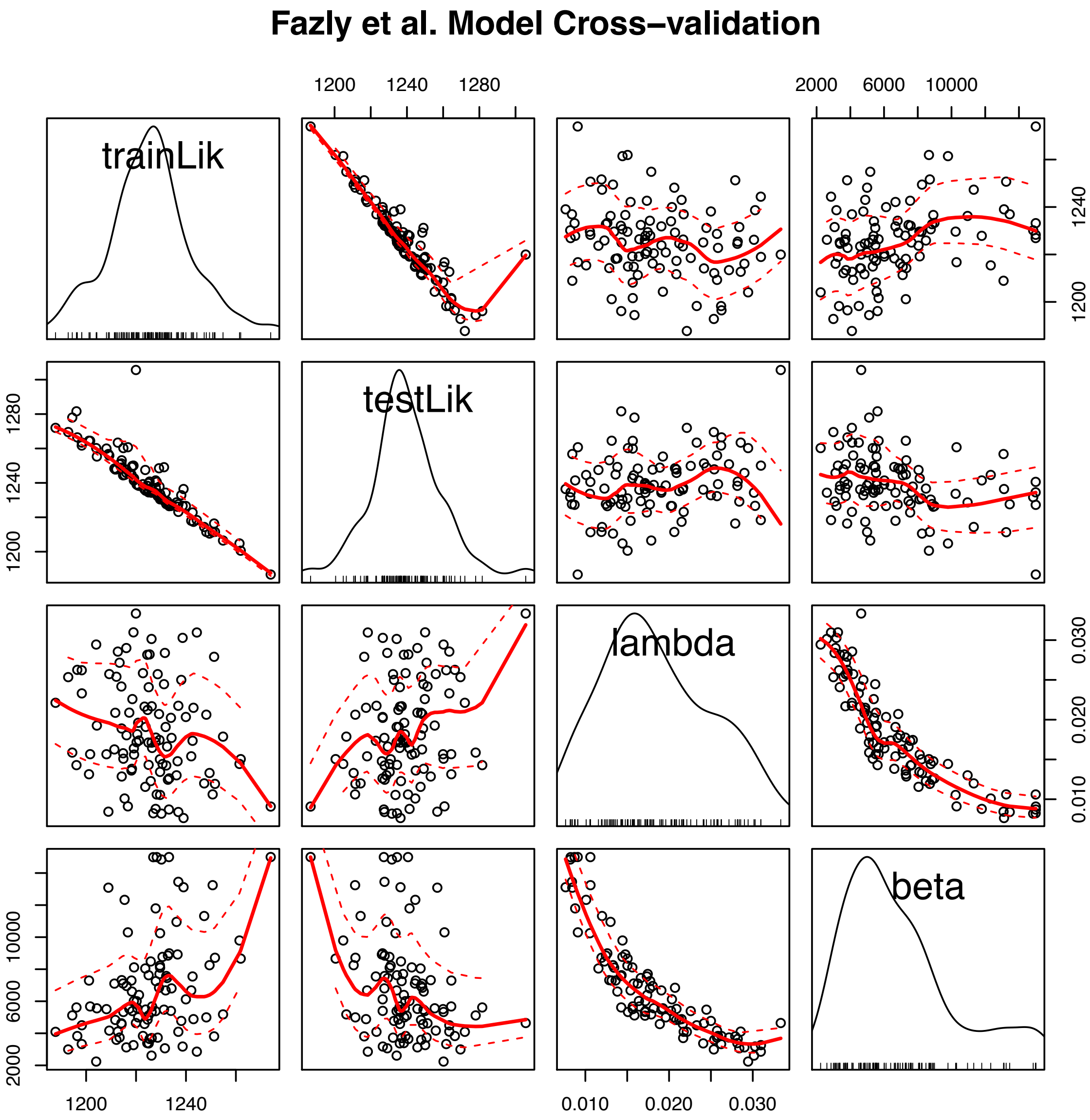

Figure 3. Scatterplot of the Fazly et al. probabilistic incremental model's best-fitting parameters versus training data fit (trainLik) and the held-out testing data fit (testLik) for the 100 random cross-validation subsets. The lambda and beta parameters seem to trade off, suggesting that the model could be reparameterized.

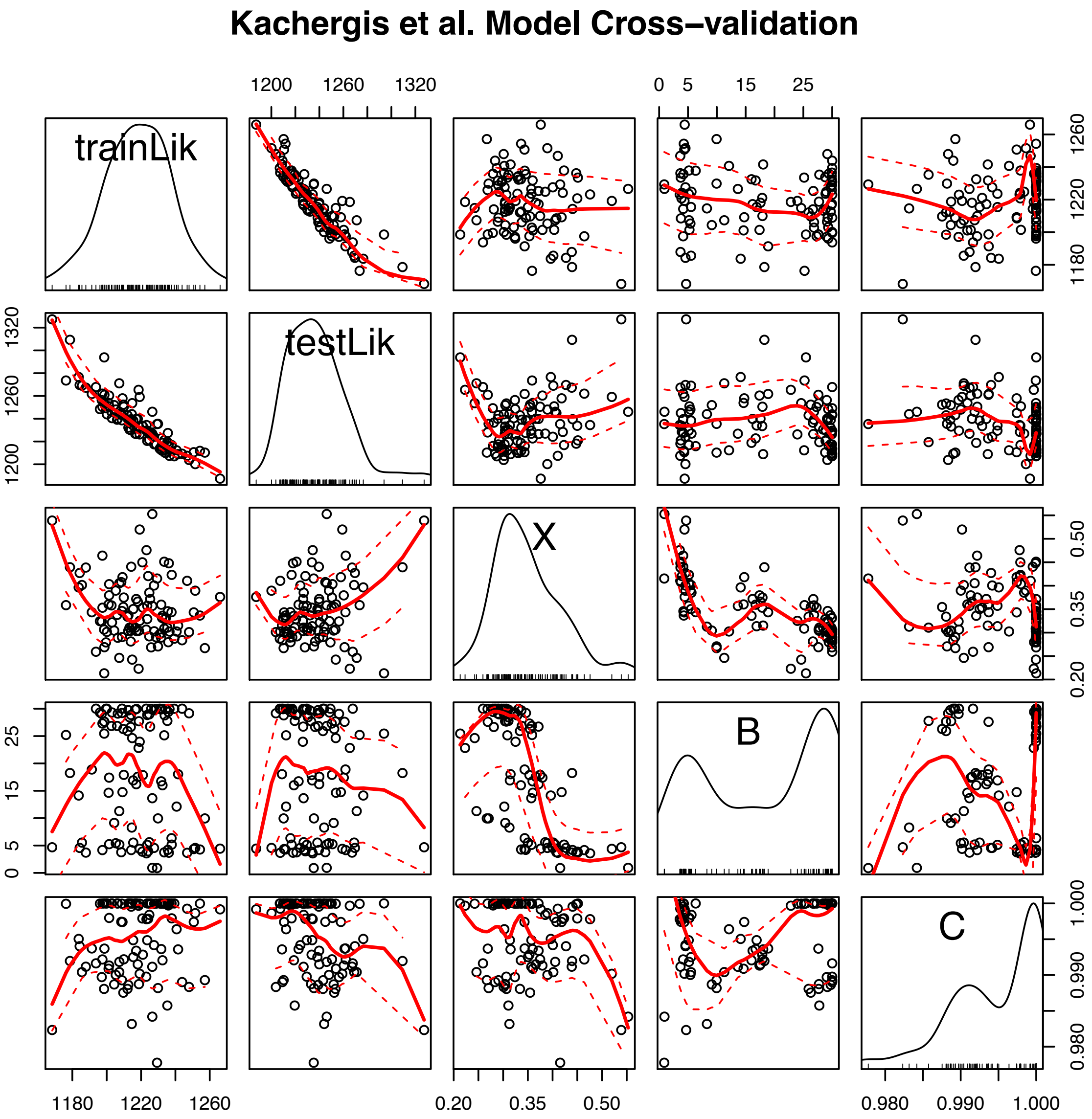

Figure 4. Scatterplot of the Kachergis et al. associative model's best-fitting parameters versus fit to the training data (trainLik) and to the held-out testing data (testLik) for the 100 random subsets of half the data. (*N.b.*: Parameter α is C in the figure.)

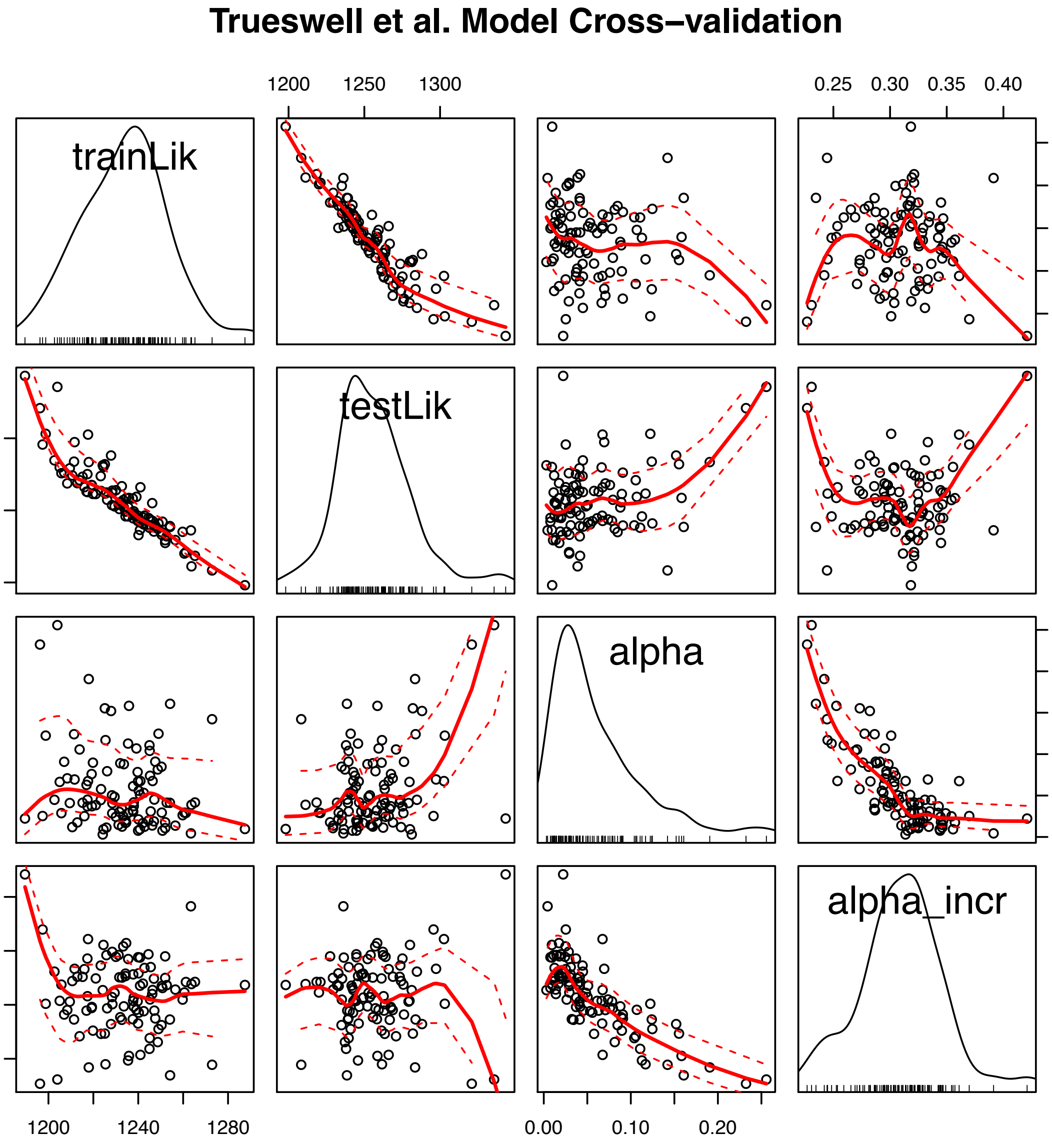

Figure 5. Scatterplot of the Trueswell *et al*. propose-but-verify model's best-fitting parameters versus fit to the split-halves of training data (trainLik) and to test data (testLik) for 100 random subsets.