Feature overwriting as a finite mixture process: Evidence from comprehension data

Shravan Vasishth (vasishth@uni-potsdam.de)
Department of Linguistics, University of Potsdam, Germany

Lena Jäger (lena.jaeger@uni-potsdam.de)
Department of Linguistics, University of Potsdam, Germany.

Bruno Nicenboim (bruno.nicenboim@uni-potsdam.de)
Department of Linguistics, University of Potsdam, Germany.

Abstract

The ungrammatical sentence *The key to the cabinets are on the table* is known to lead to an illusion of grammaticality. As discussed in the meta-analysis by Jäger et al., 2017, faster reading times are observed at the verb *are* in the agreement-attraction sentence above compared to the equally ungrammatical sentence *The key to the cabinet are on the table*. One explanation for this facilitation effect is the feature percolation account: the plural feature on *cabinets* percolates up to the head noun *key*, leading to the illusion. An alternative account is in terms of cue-based retrieval account (Lewis & Vasishth, 2005), which assumes that the non-subject noun *cabinets* is misretrieved due to a partial feature-match when a dependency completion process at the auxiliary initiates a memory access for a subject with plural marking. We present evidence for yet another explanation for the observed facilitation. Because the second sentence has two nouns with identical number, it is possible that these are, in some proportion of trials, more difficult to keep distinct, leading to slower reading times at the verb in the first sentence above; this is the feature overwriting account of Nairne, 1990. We show that the feature overwriting proposal can be implemented as a finite mixture process. We reanalysed ten published data-sets, fitting hierarchical Bayesian mixture models to these data assuming a two-mixture distribution. We show that in nine out of the ten studies, a mixture distribution corresponding to feature overwriting furnishes a superior fit over both the feature percolation and the cue-based retrieval accounts.

Keywords: Feature overwriting; feature percolation; cue-based retrieval; sentence processing; interference; reading; Bayesian hierarchical mixture models

Introduction

It is well-known that sentences such as (1a) can lead to an illusion of grammaticality. The sentence is ungrammatical because of the lack of number agreement between the subject *key* and the auxiliary *are*. Note that the second noun, *cabinets*, and the auxiliary *are* agree in number, but no syntactic agreement is possible between these two elements.

(1) a. The key to the cabinets are on the table.
b. The key to the cabinet are on the table.

Many sentence comprehension studies have shown that the illusion has the effect that the auxiliary *are* is read faster in (1a) compared to the equally ungrammatical sentence (1b) (see Jäger, Engelmann, & Vasishth, 2017 for a review). In contrast to (1a), in (1b) the second noun (*cabinet*) is singular and does not agree with the auxilary in number.

Several explanations have been proposed for the illusion of grammaticality in (1a) vs. (1b). We discuss two of these here. The feature percolation account proposes that in (1a) the plural feature on *cabinets* can, in some proportion of trials, move or percolate up to the head noun *key* (see Patson & Husband, 2016 for recent evidence for this model). The head noun now has the plural feature, leading to an illusion of grammaticality compared to (1b), where no such feature percolation occurs. Another prominent explanation, due to Wagers, Lau, and Phillips (2009), is the retrieval interference account. Here, in ungrammatical sentences like (1a), a singular verb would be predicted; but when the plural verb *are* is encountered, a cue-based retrieval process (Lewis & Vasishth, 2005) is triggered: The verb triggers an access (called a retrieval) for a noun that is plural marked and is a subject. A parallel cue-based associative memory access leads to the retrieval of a partially matching noun in memory (*cabinets*) that agrees in number but is not the subject. This partial match leads to a successful retrieval and an illusion of grammaticality.[1]

As we show next, there is evidence for both these accounts: a facilitatory effect is generally present in the published data.

The facilitatory effect in reading time in the "illusion of grammaticality" data-sets

We first establish that a facilitatory effect is found in studies comparing sentences like (1a) and (1b). In connection with the meta-analysis relating to studies on cue-based retrieval reported in Jäger et al. (2017), we had obtained the raw data from 10 studies on sentences like (1a) and (1b). These were reading-time studies reported in Dillon, Mishler, Sloggett, and Phillips (2011), Lago, Shalom, Sigman, Lau, and Phillips (2015), and Wagers et al. (2009). Except for the eyetracking experiment by Dillon and colleagues, all the other studies were self-paced reading experiments. In these data-sets, the dependent measure was reading time in milliseconds at the auxiliary or the region following it. Most of the 10 studies

[1] The cue-based retrieval account may a priori be implausible because it predicts that an incorrect dependency is built between *cabinets* and *are*; building such a dependency would imply that the sentence has the implausible meaning that the cabinets are on the table. The reader should detect such an implausible meaning and this should lead to a slowdown rather than facilitation.

found statistically significant effects in this post-critical region. What is noteworthy here is the consistently negative sign of the effect of interest; this consistency is much more informative than the statistical significance of individual studies.

We first reanalyzed these 10 data-sets in order to confirm the facilitatory effect reported.[2] We fit Bayesian hierarchical models to each data-set using Stan (Stan Development Team, 2016). We fit Bayesian models because of the ease with which statistical models can be defined flexibly to reflect the cognitive process of interest.

The model specification was as follows. Assume that (i) i indexes participants, $i = 1,\ldots,I$ and j indexes items, $j = 1,\ldots,J$; (ii) y_{ij} is the reading time in milliseconds for the i-th participant reading the j-th item; and (iii) the predictor x, which represents the experimental manipulation, is sum-coded (± 1). In our case, the condition (1a) is coded $+1$ and the condition (1b) is coded -1.

Then, the data y_{ij} (reading times in milliseconds) are defined to be generated by the following process:

$$y_{ij} \sim LogNormal(\beta_1 + \beta_2 x_{ij} + u_i + w_j, \sigma^2_e) \qquad (1)$$

where $u_i \sim Normal(0, \sigma^2_u)$, $w_j \sim Normal(0, \sigma^2_w)$ and σ^2_e is the error variance. The terms u_i and w_j are called varying intercepts for participants and items respectively; they represent by-subject and by-item adjustments to the fixed-effect intercept β_1. The variances σ^2_u and σ^2_w represent between-participant (respectively item) variance.[3] The facilitation effect is the estimate of β_2 (on the log scale).

As priors, we chose the Cauchy(0,2.5) distribution for all coefficients, and a half-Cauchy (with only positive values) for the standard deviations. This are mildly informative priors (Gelman et al., 2014) which express the belief that that the most likely value of the parameter is near 0, but allows for a wide range of non-zero values because of the fat tails of the Cauchy.

As shown in Figure 1, the effects in each study consistently show negative estimates of β_2, which indicates a facilitation in reading time at the auxiliary or a subsequent region. This is consistent with both the feature percolation and retrieval interference accounts. There is a third explanation for the observed facilitation effect in these studies, which we turn to next.

An alternative explanation for the facilitatory effect Consider the ungrammatical example sentences again. These are repeated below for convenience:

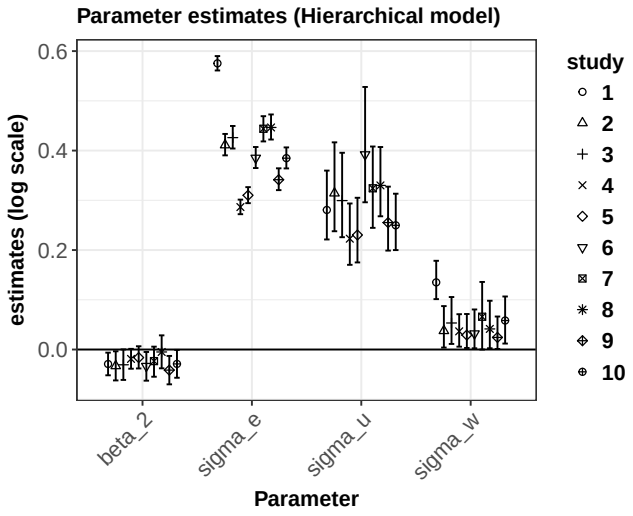

Figure 1: The parameter estimates of the hierarchical model fitted to the 10 data-sets. The condition representing (1a) is coded $+1$ and the condition representing (1b) is coded -1, so that parameter `beta_2` shows a facilitation effect if its value is negative. Shown are the estimates of the facilitatory effect (`beta_2`), and the standard deviations of (i) the error (`sigma_e`), (ii) the by-subjects varying intercepts (`sigma_u`), and (iii) the by-items varying intercepts (`sigma_w`).

(2) a. The key to the cabinets are on the table.
b. The key to the cabinet are on the table.

In example (2b), both the nouns are marked singular, whereas in example (2a) the nouns have different number marking. As discussed in Villata and Franck (2016), the similarity in number of the two nouns in (2b) could be the underlying cause for increased processing difficulty, compared to (2a). The identical number marking in (2b) could lead to increased confusability between the two nouns, leading to longer reading times at the moment when a subject noun is to be accessed at the auxiliary verb. The feature overwriting model of Nairne (1990) formalizes this idea. To quote (p. 252): *An individual feature of a primary memory trace is assumed to be overwritten, with probability F, if that feature is matched in a subsequently occurring event. Interference occurs on a feature-by-feature basis, so that, if feature b matches feature a, the latter will be lost with probability F.*

The Nairne proposal has a natural interpretation as a finite mixture process. Specifically, feature overwriting could occur with a higher probability in example (2b) compared to (2a). This assumption implies that the reading times in both (2b) and (2a) are generated from a mixture of two distributions. In a particular trial, if no feature overwriting occurs, the reading time would come from a Lognormal distribution with some location and scale parameters; this situation would result in minimal processing difficulty in carrying out a retrieval and detecting the ungrammaticality. In other trials, when feature overwriting does occur, the reading time would have a larger

[2] The published studies had other experimental conditions that we do not discuss here. The published studies also used a trimming procedure to analyze the data, and their analysis was done on the raw millisecond scale. Thus, our analysis has some differences from the original analyses, but the conclusions are substantially unchanged.

[3] This so-called crossed participants and items varying intercepts linear mixed model can be made more complex by adding varying slopes for the factor X by participant and by item, but for space reasons we do not consider these more complex models here.

location parameter, and possibly also a larger scale parameter; this would represent the cases where additional difficulty occurred due to feature overwriting.[4]

An explicit assumption here is that feature overwriting could occur in both (2b) and (2a), but the proportion would be higher in (2b). It is also possible to assume that feature overwriting only occurs in (2b), but due to space reasons we do not consider this and other alternative models here.

Thus, in the mixture model implementation of the Nairne proposal, one distribution will have a larger location parameter (and perhaps also the scale parameter). In the modelling presented below, one goal is to estimate the mixing proportions of these distributions. In the results section, we will refer to the proportion of the slow reading time distributions in (2b) as `prob_hi`, and in (2a) `prob_lo`. The suffixes `hi` and `lo` here refer to whether we expect confusability to be high or low.

To summarize, the feature percolation, cue-based retrieval, and feature overwriting models all predict facilitation in the ungrammatical sentences (2a) compared to (2b), but the underlying generative process assumed in each model is different. Feature percolation and feature overwriting can be seen as finite mixture models of different types, and cue-based retrieval can be seen as implemented by the standard hierarchical model. Our goal here is to implement all the three proposals as statistical models and then compare their relative fit to the data in order to adjudicate between them. Before we do this, we introduce finite mixture models.

Finite mixture models

A finite mixture model assumes that the independently distributed outcome $y_i, i = 1,\ldots,N$ is drawn from one of several distributions. Each distribution's identity is controlled by a Categorical distribution. For example, assume that we have K distributions with location parameter (the mean) $\mu_k \in \mathbb{R}$ and scales (standard deviation) $\sigma_k \in (0,\infty)$, where $k = 1,\ldots,K$. Assume also that we have a vector of probabilities $<\lambda_1,\ldots,\lambda_K>=\Lambda$ that represent the mixing proportions. The parameters λ_k are non-negative values and they sum to 1.

Thus, if the K distributions are mixed in proportion Λ, where $\lambda_k \geq 0$ and $\sum_{k=1}^{K}\lambda_k = 1$, for each outcome y_i there is a latent variable $z_i \in \{1,\ldots,K\}$ with a Categorical distribution[5] parameterized by $\lambda : z_i \sim \texttt{Categorical}(\lambda)$. The variable y_i is then distributed as follows:

$$y_i \sim Normal(\mu_{z_i}, \sigma^2_{z_i}) \tag{2}$$

Assuming that each of the K mixture distributions $f(\cdot)$ has a vector of parameters θ_k associated with it, the mixture density can be written in the following manner:

$$p(y_i \mid \theta, \Lambda) = \lambda_1 \cdot f(y_i \mid \theta_1) + \cdots + \lambda_K \cdot f(y_i \mid \theta_K) \tag{3}$$

A random variable Y with the above density can then be written in abbreviated form as follows.

$$Y \sim \lambda_1 f(y \mid \theta_1) + \cdots + \lambda_K f(y \mid \theta_K) \tag{4}$$

In this paper, we consider a mixture of LogNormals with $K = 2$; this is because the feature overwriting model assumes a mixture of two distributions. We choose LogNormals to model reading times because reading times must be greater than 0 and follow a LogNormal distribution. We will write the models as follows:

$$\begin{aligned} Y \sim & \lambda_1 \cdot LogNormal(\mu_1 + \delta, \sigma^2_1) + (1-\lambda_1) \cdot LogNormal(\mu_1, \sigma^2_2) \\ & \text{where } \sigma^2_1 = \sigma^2_2 \text{ or } \sigma^2_1 \neq \sigma^2_2 \end{aligned} \tag{5}$$

The parameter δ marks the shift in the mean in the first mixture distribution relative to the second mixture distribution. Note that the scale parameters (σ_1, σ_2) can be either identical (homogeneous variances) in both distributions, or different (heterogeneous variances). We will consider both types of models here.

The above models assume that the data are independent. When we have repeated measures data, the independence assumption is no longer valid. In order to address this issue, finite mixture models can be made hierarchical by adding varying intercepts for subjects (indexed by i) and items (indexed by j):

$$\begin{aligned} y_{ij} \sim & \lambda_1 \cdot LogNormal(\mu_1 + \delta + u_i + w_j, \sigma^2_1) + \\ & (1-\lambda_1) \cdot LogNormal(\mu_1 + u_i + w_j, \sigma^2_2) \end{aligned} \tag{6}$$

where $u_i \sim Normal(0, \sigma^2_u)$ and $w_j \sim Normal(0, \sigma^2_w)$. Thus, the mixture model with $K = 2$ will have the following parameters: four variance components, $\sigma^2_1, \sigma^2_2, \sigma^2_u$, and σ^2_w; two coefficients μ_1 and δ; and two probabilities λ_1 and $\lambda_2 = (1-\lambda_1)$.

An evaluation of the Nairne feature overwriting proposal

Method

Implementing the Nairne proposal We fit the homogeneous and heterogeneous variance hierarchical mixture models to the 10 reading time data-sets that compared reading times at the auxiliary or the following region for sentences like (2a) and (2b).

The data were assumed to be generated from a two-mixture Lognormal distribution with either a homogeneous variance in both mixture distributions, or heterogeneous variances.

[4] In grammatical sentences like *The key to the cabinet/s is...*, both feature overwriting and cue-based retrieval predict a slowdown when the nouns have the same number. The literature largely shows no difference in reading time. But the two models' relative performance can still be investigated; we plan to do this in future work.

[5] The Categorical distribution can be seen here as the Bernoulli distribution in the case where K=2. In this paper, we focus only on the K=2 case.

Thus, for the high confusability condition (2b), we considered two models:

$$\begin{aligned}&\underline{\text{Homogeneous variance feature overwriting model}}\\ y_{ij} \sim &\text{prob\_hi} \cdot \mathit{LogNormal}(\beta+\delta+u_i+w_j,\sigma_e^2)+\\ &(1-\text{prob\_hi}) \cdot \mathit{LogNormal}(\beta+u_i+w_j,\sigma_e^2)\\ &\text{where:}\\ &u_i \sim \mathit{Normal}(0,\sigma_u^2), w_k \sim \mathit{Normal}(0,\sigma_w^2)\end{aligned} \tag{7}$$

$$\begin{aligned}&\underline{\text{Heterogeneous variance feature overwriting model}}\\ y_{ij} \sim &\text{prob\_hi} \cdot \mathit{LogNormal}(\beta+\delta+u_i+w_j,\sigma_{e'}^2)+\\ &(1-\text{prob\_hi}) \cdot \mathit{LogNormal}(\beta+u_i+w_j,\sigma_e^2)\\ &\text{where:}\\ &u_i \sim \mathit{Normal}(0,\sigma_u^2), w_k \sim \mathit{Normal}(0,\sigma_w^2)\end{aligned} \tag{8}$$

In both models, y_{ij} is the reading time in milliseconds from subject i and item j. The probability `prob_hi` represents the mixing probability of the distribution that generates the slow reading times corresponding to trials where feature overwriting occurred (2b). Although not shown, another mixture distribution is defined for example (2a); here, `prob_lo` represents the mixing probability of the distribution that generates the slower reading times corresponding to the trials where feature overwriting occurred.

The homogeneous variance model assumes that both mixture distributions have the same standard deviation σ_e. The heterogeneous mixture model assumes that the mixture distribution that leads to the slower reading times is assumed to have both a different mean ($\beta+\delta$) and a different standard deviation ($\sigma_{e'}$) than the other distribution. Alternative models can be fit which relax these assumptions, but due to space constraints we consider only these two models.

We had the following priors for the parameters:

$$\begin{aligned}\text{prob\_hi} &\sim \mathit{Beta}(1,1)\\ \beta,\delta &\sim \mathit{Cauchy}(0,2.5)\\ \sigma_e,\sigma_{e'},\sigma_u,\sigma_w &\sim \mathit{Cauchy}(0,2.5)\\ &\text{constraint: } \sigma_e,\sigma_{e'},\sigma_u,\sigma_w > 0\end{aligned} \tag{9}$$

The priors for the variance components (the standard deviations σ_e, $\sigma_{e'}$, σ_u, σ_w) and the coefficients representing the means of the Lognormal distributions (β,δ) are mildly informative priors, as in the standard hierarchical model above. These Cauchy priors assume that values of the parameters near 0 are the most likely ones, but extreme values are possible. The Beta(1,1) prior for the mixing probabilities expresses a large prior uncertainty, and express the assumption that the probability is equally likely to be any value between 0 and 1.

Baseline models As baselines, we fit a model corresponding to the retrieval interference account (the standard hierarchical model shown in equation 1 and summarized in Figure 1), and the feature percolation proposal. The latter also assumes a mixture distribution, but only for the condition corresponding to example (2a). Recall that the claim is that in ungrammatical sentences, in some proportion of trials the plural feature on the distractor *cabinets* moves up to the head noun. In (2b), no such mixture process should occur because percolation never occurs; hence a standard hierarchical LogNormal distribution can be assumed here. We therefore defined the following generative process for (2a):

$$\begin{aligned}&\underline{\text{Feature percolation model}}\\ y_{ij} \sim &\text{prob\_perc} \cdot \mathit{LogNormal}(\beta+\gamma+u_i+w_j,\sigma_e^2)+\\ &(1-\text{prob\_perc}) \cdot \mathit{LogNormal}(\beta+u_i+w_j,\sigma_e^2)\\ &\text{where:}\\ &u_i \sim \mathit{Normal}(0,\sigma_u^2), w_k \sim \mathit{Normal}(0,\sigma_w^2), \gamma<0\end{aligned} \tag{10}$$

Note that in the specification above the parameter γ, which represents the change in the location parameter, is constrained in the model to be negative; this is because the assumption in the feature percolation proposal is that percolation leads to faster reading time.

For sentences like (2b), in which no percolation is assumed to occur, we simply assumed a LogNormal generative process:

$$y_{ij} \sim \mathit{LogNormal}(\beta+u_i+w_j,\sigma_e^2) \tag{11}$$

Model comparison Having fitted the homogeneous and heterogeneous variance models, as well as the baseline models (the cue-based retrieval and feature percolation models), we need a method for comparing the quality of fit of the mixture models relative to the standard hierarchical models. We use an approximation of the leave-one-out cross-validation (LOO), as discussed in Vehtari, Gelman, and Gabry (2016). We find this approach attractive because it focuses on the predictive performance of the model. LOO compares the expected predictive performance of alternative models by subsetting the data into a training set (for estimating parameters) by excluding one observation. The difference between the predicted and observed held-out value can then be used to quantify model quality by successively holding out each observation. The sum of the expected log pointwise predictive density, $\widehat{elpd}$, can be used as a measure of predictive accuracy, and the difference between the $\widehat{elpd}$'s of competing models can be computed, including the standard deviation of the sampling distribution of the difference in $\widehat{elpd}$. When comparing a model M1 with another model M2, if M2 has a higher $\widehat{elpd}$, then it has a better predictive performance compared to M1. The model comparisons are transitive; if a third model M3 has a higher $\widehat{elpd}$ than M2, then it has a better performance than M1 as well. Vehtari and colleagues have developed an efficient computation of LOO using Pareto-smoothed

importance sampling (PSIS-LOO), This is what we use here. For details of PSIS-LOO, see Vehtari et al. (2016).

Results

Table 1 shows model comparisons between the standard hierarchical model, corresponding to the retrieval interference account, and the homogeneous variance model. The table shows that apart from study 1, the homogeneous variance feature overwriting model is clearly superior to the retrieval interference model because it has higher $\widehat{elpd}$ values. Table 1 also shows that the homogeneous variance feature overwriting model furnishes a better fit than the feature percolation model. Finally, the table shows that, except for study 1, the heterogeneous variance model is superior to the homogeneous variance model.

Since the model comparisons are transitive, we can conclude that, among the models compared, the heterogeneous variance feature overwriting model characterises the data best. We therefore focus on the parameter estimates of the heterogeneous variance model below. The estimates from the models for the 10 data-sets are shown in Figure 2. In this model, two noteworthy points are the following: (i) The variance of the high confusability distribution (`sigmap_e`; this corresponds to $\sigma_{e'}$ in the models defined earlier) is relatively large compared to the other variance components; (ii) The difference in probabilities of the two mixture distributions, `diffprob`, is generally greater than 0 across all the studies; however, the uncertainty in the estimate of the probability in study 1 is very high. These two observations suggest that there is more variability in the reading time when the feature overwriting occurs, and that there some evidence that the proportion of trials with feature overwriting is higher in the condition with two singular nouns, consistent with the Nairne proposal.

In summary, overall there is good motivation to assume that in the condition with two singular nouns (example 2b), a proportion of trials comes from a distribution with a larger mean and larger standard deviation, and this proportion is higher than in the condition with one singular and one plural noun (example 2b).

Discussion

We implemented as a statistical model the proposal that nouns with similar feature marking (here, number) may be more confusable due to feature overwriting in some proportion of trials, which in turn leads to occasional increase in difficulty in accessing the correct noun when a dependency is to be completed between the subject and the verb. By fitting Bayesian hierarchical two-mixture models, we showed that 9 out of the 10 data-sets showed evidence for this increased confusability in one condition over the other. The feature overwriting account for the ungrammatical sentences (2a, 2b) appears to be superior to both the retrieval interference and feature percolation accounts.

The three accounts make the same predictions for ungrammatical sentences—a facilitation effect. The modelling presented here allows us to quantitatively compare the relative fit of these proposals for these otherwise indistinguishable accounts. An interesting future direction is to evaluate the predictions of these models for grammatical sentences such as those considered in Franck, Colonna, and Rizzi (2015); Villata and Franck (2016). We plan to address this in future work.

Acknowledgments

Our thanks to Brian Dillon, Sol Lago, Colin Phillips, and Matt Wagers for generously sharing their data. We also benefitted a great deal from discussions with Julie Franck, Whitney Tabor, Aki Vehtari, and Michael Betancourt. For partial support of this research, we thank the Volkswagen Foundation through grant 89 953 to the first author.

References

Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2011). A computational cognitive model of syntactic priming. *Journal of Memory and Language*, *69*(4), 85-103.

Franck, J., Colonna, S., & Rizzi, L. (2015). Task-dependency and structure-dependency in number interference effects in sentence comprehension. *Frontiers in psychology*, *6*, 349.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Third ed.). Chapman and Hall/CRC.

Jäger, L. A., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language*, *94*, 316-339. doi: 10.1016/j.jml.2017.01.004

Lago, S., Shalom, D. E., Sigman, M., Lau, E. F., & Phillips, C. (2015). Agreement attraction in spanish comprehension. *Journal of Memory and Language*, *82*, 133–149.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*, 1–45.

Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition*, *18*(3), 251–269.

Patson, N. D., & Husband, E. M. (2016). Misinterpretations in agreement and agreement attraction. *The Quarterly Journal of Experimental Psychology*, *69*(5), 950–971.

Stan Development Team. (2016). Stan modeling language users guide and reference manual, version 2.12 [Computer software manual]. Retrieved from `http://mc-stan.org/`

Vehtari, A., Gelman, A., & Gabry, J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*.

Villata, S., & Franck, J. (2016). *Similarity-based interference in agreement comprehension and production: Evidence from object agreement.* (Manuscript)

Wagers, M., Lau, E., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, *61*(2), 206–237.

	(a) Standard HLM vs. Homogeneous variance mixture model		(b) Percolation vs. Homogeneous variance mixture model		(c) Homogeneous variance vs. Heterogeneous variance mixture model	
Study	elpd_diff	SE	elpd_diff	SE	elpd_diff	SE
1	-0.29	1.67	29.55	6.97	0.57	1.09
2	56.98	13.57	76.34	14.26	15.20	6.07
3	97.62	16.10	112.40	17.43	57.12	11.11
4	71.29	14.08	84.78	14.12	19.66	8.77
5	112.74	18.17	120.45	18.56	63.28	18.12
6	66.84	12.59	85.97	13.88	43.58	12.18
7	72.45	13.76	80.93	14.72	80.92	14.41
8	88.50	14.60	90.22	14.77	40.17	11.87
9	78.35	14.21	108.10	16.04	26.21	7.76
10	90.08	14.14	105.23	15.02	33.59	11.95

Table 1: Comparison of the 10 sets of hierarchical models using PSIS-LOO. Shown are the differences in $\widehat{elpd}$ between (a) the standard hierarchical model and the homogeneous variance mixture model; (b) the feature percolation model and the homogeneous variance mixture model; and (c) the homogeneous vs. heterogeneous variance mixture model. Also shown are standard errors for each comparison. If the difference in $\widehat{elpd}$ is positive, this is evidence in favour of the second model. The pairwise model comparisons are transitive. These comparisons show that the heterogeneous variance mixture model has the best predictive performance.

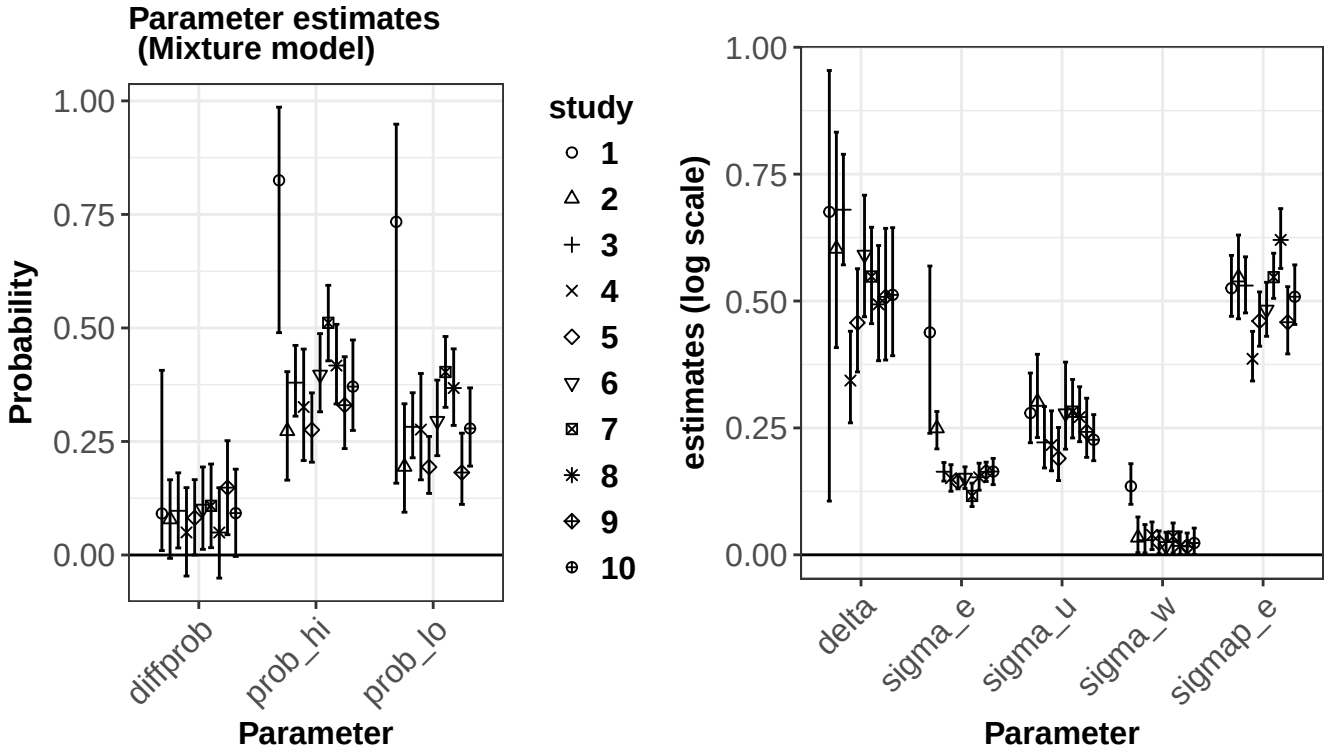

Figure 2: Parameter estimates for the heterogeneous variance hierarchical mixture models.