Spring 2021

Posterior Predictive Critique of a Psychometric Bayesian Model for Assessing Aphasia

Ashlynn Crisp
Portland State University

Posterior Predictive Critique of a Psychometric Bayesian Model for Assessing Aphasia

by

Ashlynn Crisp

An undergraduate honors project submitted
in partial fulfillment of the requirements for the

Fariborz Maseeh Department of Mathematics and Statistics
Honors Track

Under the supervision of
Prof. Daniel Taylor-Rodriguez

Portland State University

June 12, 2021

1 Introduction

For persons with aphasia, naming tests are useful for assessing the severity of the disease and observing progress toward recovery. The Philadelphia Naming Test (PNT) is a leading naming test composed of 175 items. The items are common nouns which are one to four syllables in length and with low, medium, and high frequency (Fergadiotis et al., 2015). Since the target word is known to the administrator, the response from the patient can be classified as correct or an error. If the patient commits an error, the PNT provides procedures for classifying the type of error in the response (Fergadiotis et al., 2015). Item response theory can be applied to PNT data to provide estimates of item difficulty and subject naming ability (e.g., De Ayala (2013)).

Walker et al. (2018) developed a multinomial processing tree (MPT) model to draw more insight from the types of errors patients commit in responding to an item. The MPT model expands on existing models by considering items to be heterogeneous and estimating multiple latent parameters for patients to more precisely determine at which step of word production a patient's ability is affected. These latent parameters represent the theoretical cognitive steps taken in responding to an item, shown in Figure 1.

The purpose of this paper is to provide an assessment of the goodness-of-fit of the MPT model through posterior predictive checking. Background information for the MPT model is provided in the next section, followed by details of the statistical methods applied and results. The paper concludes with a discussion of areas for improvement for the MPT model and implications of use with the current design.

MPT Model Architecture

Figure 1 shows the structure of the MPT model where the probability of successfully performing each of the processes is denoted by the letters a-h. Of the eight processes specified by the model, Attempt, LexSem, LexPhon, LexSel, and Phon are dependent on the patient's ability and the item difficulty. There is one participant-only parameter Sem, one item-specific parameter Word-T, and one global parameter Word-L. The word outcome is dependent on which processes the subject performs correctly and will end in one of eight possible response categories: Correct, Semantic, Formal, Mixed, Unrelated, Neologism, Abstruse Neologism, and Non-naming Attempt.

The MPT model applies the Rasch model from item response theory to estimate the Attempt, LexSem, LexPhon, LexSel, and Phon parameters. Given the ability estimate θ_{ts} for subject t and item difficulty estimate δ_{ks} for item k, the probability of correctly performing process s is calculated as:

$$\psi_{stk} = \frac{e^{(\theta_{ts} - \delta_{ks})}}{1 + e^{(\theta_{ts} - \delta_{ks})}} \tag{1}$$

Each of the ψ_{stk} probabilities are assumed to be independent of each other. Letting j represent the response category, the probability of following a particular branch i to category j is

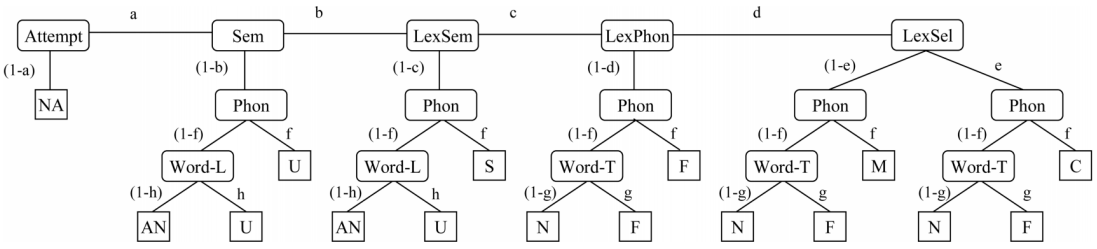


Figure 1: MPT model structure

then given by:

$$P(B_{kji}|\boldsymbol{\psi}_t) = \prod_{s}^{S} \psi_{ts}^{a_{kjis}} (1 - \psi_{ts})^{b_{kjis}} \tag{2}$$

where a_{kjis} equals 1 if process s was performed correctly and equals 0 otherwise. Conversely, b_{kjis} equals 1 if process s was not performed correctly and is 0 otherwise.

Since there are multiple branches leading to the same error type, the probability of a certain response category occurring is the sum of the probabilities for all branches leading to that error. Let I_j be the number of possible paths for response category j. For example, $I_j = 1$ for the Correct category since there is one path terminating with Correct; however, $I_j = 4$ for the Formal category since there are four paths terminating with Formal. For each person t, the probability of a particular response category j on item k is given by:

$$P(C_{kj}|\boldsymbol{\psi}_t) = \sum_{i=1}^{I_j} \prod_{s=1}^{S} \psi_{st}^{a_{skji}} (1 - \psi_{st})^{b_{skji}} \tag{3}$$

The category counts $\boldsymbol{n}_t = (n_{11t}, \ldots, n_{kJt}, \ldots, n_{K1t}, \ldots, n_{KJt})$ for each person t on item k and category j follows a product-multinomial distribution:

$$P(\boldsymbol{N}_t|\boldsymbol{\psi}_t) = \prod_{k=1}^{K} \prod_{j=1}^{J} P(C_{kj}|\psi_t)^{n_{kjt}} \tag{4}$$

The MPT model is composed of 2,923 parameters in total. There are 1,872 parameters which are subject-specific (312 subjects times 6 subject-specific parameters), 1,050 item-specific parameters (175 items times 6 item-specific parameters), and one global parameter. The posterior distributions are calculated through Gibbs sampling with standard normal priors for subject ability and item difficulty parameters and standard uniform for the Sem, Word-T, and Word-L parameters.

2 Methods

We fit the MPT model using RStan (Stan Development Team, 2020) for implementing the sampling algorithm. Four chains were run in parallel to check for model convergence, which was observed in traceplots. The chains were thinned to reduce auto-correlation, resulting in 2,000 MCMC draws which were used for inference. For each of the 2,000 draws, a sample from the posterior predictive distribution was sampled.

To test the goodness-of-fit of the MPT model, posterior predictive checks were used. If the posterior distributions of the model parameters fit the data well, samples drawn from the posterior predictive distribution should be similar to the observed data. That is, we can determine how well the model fits the observed data by the extent to which the posterior predictive data aligns with the observed data. The posterior predictive distributions were analyzed by calculating model prediction accuracy and discrepancy statistics. Additionally, Bayesian credible intervals were calculated to determine confidence in the participant naming ability point estimates.

Model Prediction Accuracy

Each draw from the posterior predictive distribution is a vector of length 8 composed of seven 0's and one 1 which indicates the response category predicted from the corresponding draw from the posterior distribution. Following the approach of Walker et al., for each person and item combination, the mode (i.e., the most frequently predicted response category from the 2,000 posterior predictive samples) was taken as the predicted response type. The MPT model's accuracy and

fit is calculated by reconstructing the data from the posterior predictive distribution. In other words, the accuracy is the percent of the original data which can be correctly recovered from the model parameters. Sensitivity and specificity (a.k.a., true positive and true negative rates, respectively), as well as precision were calculated for each response category to observe the MPT model's accuracy across response types.

T_1 and T_2 Statistics

Klauer (2010) developed the T_1 and T_2 statistics to use as posterior predictive checking metrics for multinomial processing tree models. We computed these T_1 and T_2 discrepancy statistics to assess the recovery of the mean and covariance structure of the data, respectively. For these statistics, we consider each item which a subject responds to as a subtree. This is to account for item heterogeneity, which is a core feature of the MPT model. Using this approach, there are 175 subtrees, and each participant navigates each subtree once per item they are shown. We calculated the T_1 and T_2 statistics as follows: on every draw s from the 2,000 posterior distribution samples, we calculated $T(\text{data}, \boldsymbol{\psi})$. For the same draw from the posterior distribution, a draw from the posterior predictive distribution was taken and $T(\text{data}^{\text{pred}}, \boldsymbol{\psi})$ was calculated. This creates 2,000 T_1 and T_2 statistics from the observed data and the corresponding 2,000 T_1 and T_2 statistics for the predicted data. The p-value for each T statistic is the proportion of times where $T(\text{data}^{\text{pred}}, \boldsymbol{\psi}) > T(\text{data}, \boldsymbol{\psi})$. Klauer states that the model is adequately recovering the mean or covariance if the p-value is not small. A small p-value would indicate the T statistics from the posterior predictive data are mostly smaller than the T statistics from the observed data. This would suggest the model is not adequately recovering the mean or covariance structure observed in the data.

The T_1 statistic tests if the mean category frequencies across participants in the data are recovered by the model. For each person t, the expected category counts are given by the number of trials on each item multiplied by the probability of each category for that item. Since each item is only attempted once per person, the expected category counts $\hat{n}_{kjt}$ on item k for category j and person t is $\hat{n}_{kjt} = p(C_{kj}|\boldsymbol{\psi}_{kjt})$. The average expected category counts over persons is then $\hat{n}_{kj.} = \frac{1}{T}\sum_{t=1}^{T}\hat{n}_{kjt}$. The observed category counts n_{kjt} are the frequencies observed in either the data or the samples from the posterior predictive distribution. The average category counts for the observed frequencies are similarly averaged over persons such that $n_{kj.} = \frac{1}{T}\sum_{t=1}^{T} n_{kjt}$. The T_1 statistic is then given by:

$$T_1(n, \theta) = \sum_{kj} \frac{(n_{kj.} - \hat{n}_{kj.})^2}{\hat{n}_{kj.}} \tag{5}$$

The T_2 statistic tests if the covariance structure observed in the data is adequately recovered by the model. The expected covariance for an item/category combination $\sigma_{kj,lm}$ is composed of two parts. Part (1) is the covariance over persons of the expected counts multiplied by $\frac{(T-1)}{T}$. This covariance is calculated for all item and category combinations. Letting k, l represent two items and j, m represent two category types, this covariance is given by $\frac{1}{T}\sum_{t=1}^{T}(\hat{n}_{kjt} - \bar{\hat{n}}_{kjt})(\hat{n}_{lmt} - \bar{\hat{n}}_{lmt})$ where $\bar{\hat{n}}_{kjt}$ and $\bar{\hat{n}}_{lmt}$ are the means of the expected counts over t for that particular item/category combination. Note that whenever $k = l$ and $j = m$ this is simply the sample variance of $\hat{n}_{kjt}$ over t multiplied by $\frac{(T-1)}{T}$.

Part (2) is first computed for each person t. This person-level covariance is given by the covariance formula for the product multinomial distribution shown in Equation 4. The covariance between categories from two distinct items is assumed to be 0. For distinct j, m within the same item k, this is $-p_{kjt}p_{kmt}$. When $j = m$, the variance is given by $p_{kjt}(1 - p_{kjt})$. These variances/covariances are averaged over t. Then, parts (1) and (2) are summed together, such that

$$\sigma_{kj,lm} = \frac{1}{T}\sum_{t=1}^{T}(\hat{n}_{kjt} - \bar{\hat{n}}_{kjt})(\hat{n}_{lmt} - \bar{\hat{n}}_{lmt}) + \frac{1}{T}\sum_{t=1}^{T}\text{cov}(\hat{n}_{kjt}, \hat{n}_{lmt}).$$

The observed covariance $s_{kj,lm}$ is calculated from the data or predicted data and multiplied by $\frac{(T-1)}{T}$. The T_2 statistic is given by:

$$T_2(n,\theta) = \sum_{kj}\sum_{lm} \frac{(s_{kj,lm} - \sigma_{kj,lm})^2}{\sqrt{\sigma_{kj,kj}\sigma_{lm,lm}}} \tag{6}$$

Subject Ability Credible Intervals

Walker et al. constructed 95% Bayesian credible intervals to assess the reliability of the subject ability estimates provided by the MPT model. The intervals were calculated by taking the mean of the posterior distribution plus or minus two standard deviations. However, subject abilities in the MPT model are not required or known to be normally distributed. Here, we calculate Bayesian 95% highest posterior density credible interval to assess the certainty of the ability estimates.

3 Data

Response categories to the Philadelphia Naming Test were collected from 312 subjects. Table 1 shows the description of the category responses provided by Walker et al. Of the 312 subjects, the data for 289 of them were collected from the Moss Aphasia Psycholinguistics Project Database (MAPPD) (Mirman et al., 2010).

Each subject was shown each of the 175 items in the PNT, giving 54,600 total responses. The total category counts from all subjects are shown in Table 1. The category counts as a percentage of the data are similar to those used by Walker et al. The largest difference is in the Correct category. The data for this project contained 61.6% Correct responses. The data used by Walker et al. contained 55.8% Correct responses. The proportions of the other response categories differed by less than 2%.

4 Results

Prediction Accuracy

Prediction accuracy for the MPT model varies drastically by response types. Figure 2 shows the overlap of counts between the predicted and observed data. The entries on the minor diagonal are

Response Category	Code	Description	Example Target: *cat*
Correct	C	The response matches the target.	cat
Semantic	S	The response is a word with only a semantic relation to the target. Semantically related responses are judged by the scorer as having a taxonomic or associative relation to the target.	dog
Formal	F	The response is a word with only a phonological relation to the target. Phonologically related responses share the initial or final phoneme with the target, or a single phoneme in the same word position aligned from center to center, or two phonemes in any word position.	hat
Mixed	M	The response is a word with both a semantic and phonological relation the target.	rat
Unrelated	U	The response is a word with neither a semantic nor a phonological relation to the target.	fog
Neologism	N	The response is not a word, but it has a phonological relation to the target.	cag
Abstruse Neologism	AN	The response is not a word, nor does it have a phonological relation to the target.	rog
Non-naming Attempt	NA	All other responses, including omissions, descriptions, non-nouns, picture parts, and fragments are considered non-naming attempts.	I don't know

Table 1: Response category classification

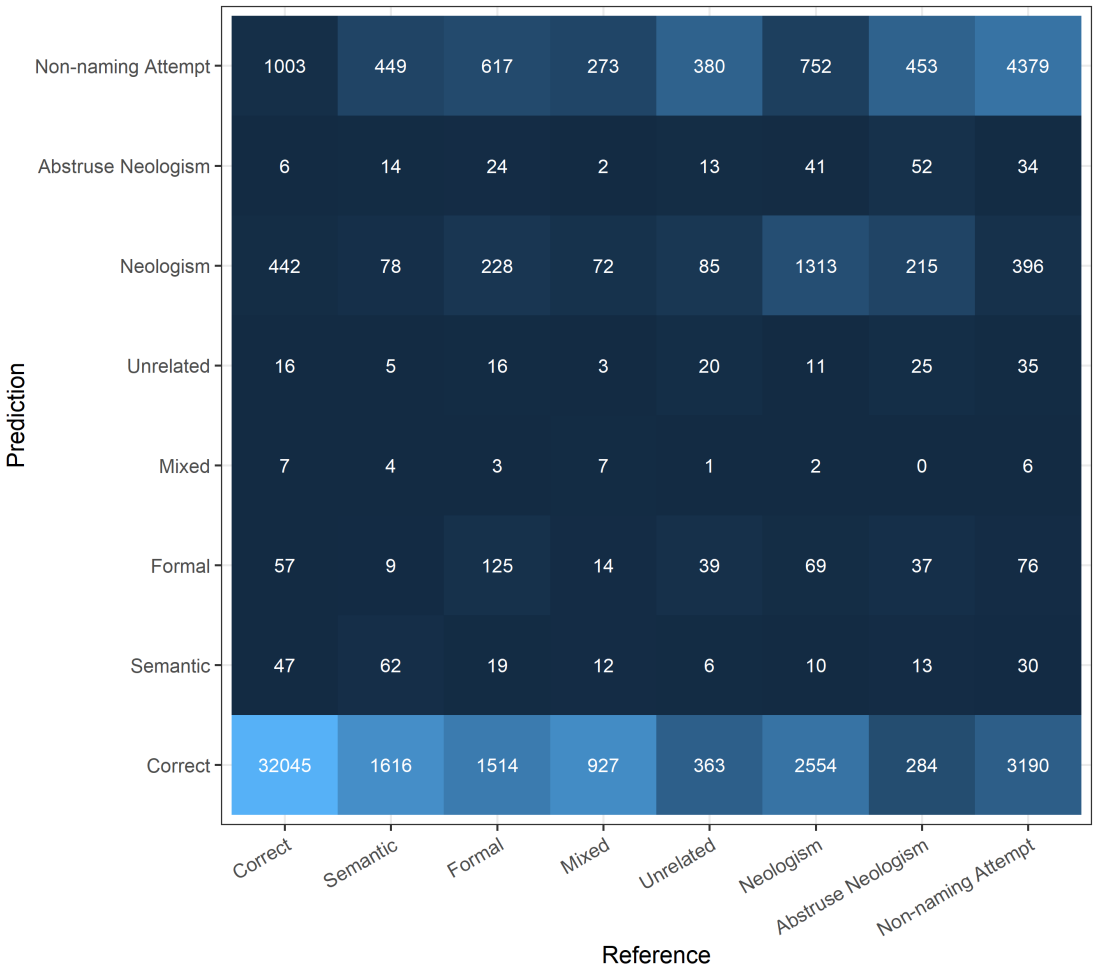


Figure 2: Confusion matrix showing distribution of error classification

the counts of correctly classified categories. Any counts off of the minor diagonal are incorrectly classified. Ideally, the minor diagonal would have the highest counts for each category. However, there is weak recovery of the less common response categories. The MPT model rarely predicts a response type other than Correct, with the second most predicted category being Non-naming Attempt.

Sensitivity, Specificity, and Precision

Table 3 shows the predicted category counts and accuracy statistics. The sensitivity is the proportion of true positives (i.e., correctly predicted ones) out of all the predicted positives. For example, the sensitivity for the Correct response category was 95%, meaning of the 33,623 Correct responses in the data, the model accurately predicted 32,045 (95%) of them. For less common response categories, the sensitivity is very low. For example, of the 2,237 Semantic responses, only 62 (2.8%) were accurately predicted. The MPT model predicts the Correct and Non-naming Attempt categories the majority of the time, which gives these two categories a relatively high sensitivity. However, the less common categories are very rarely recovered by the model giving these categories low sensitivity (27.7% and 4.9% for the two next highest sensitivity values).

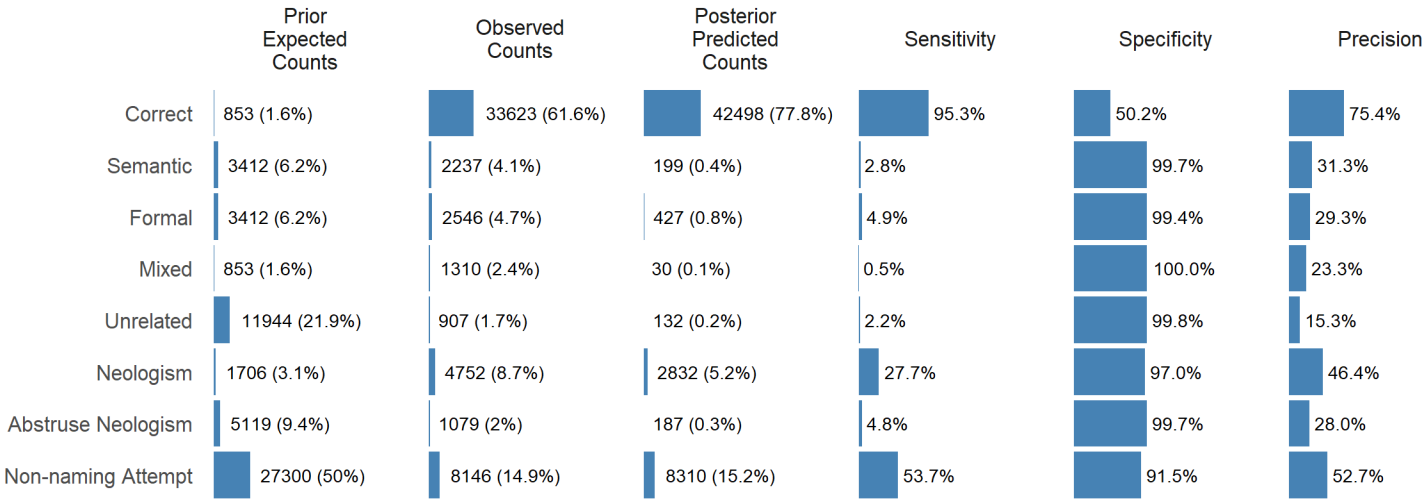


Figure 3: Category counts and model accuracy results

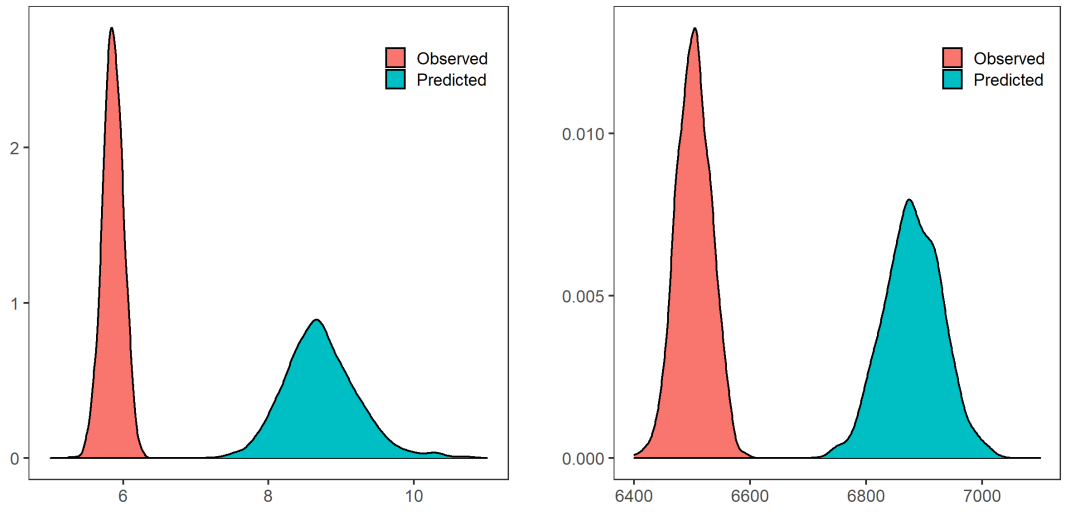
The specificity is the proportion of true negative (i.e., correctly classified zero) responses in the data. The specificity of the MPT model is low for the Correct response category and very high for all other categories. Looking at the Correct category, there are 20,977 negative entries in the data. Of these, the model accurately predicted 10,529 (50.2%) to be negative. In other words, 50% of the responses in the data which were a zero in the Correct category, were predicted by the model to be Correct. This can be seen in Figure 2 by observing the high predicted counts in the last row of the matrix. The responses incorrectly classified as Correct are the Semantic through Non-naming Attempt categories, all of which are predicted as Correct frequently. Conversely, looking at a less common response type, Mixed has 53,290 negative entries in the data. The model predicted a negative response for 53,270 (99.96%) of them.

The precision of the model is the proportion of positive predictions which are accurately classified. For example, the Correct category was accurately predicted 33,045 (75.4%) of the 42,498 total predictions in the Correct category. However, precision is low for less common response categories such as Unrelated. For the Unrelated category, only 20 (15.3%) of the 132 times Unrelated was predicted agreed with the data. Other than the Correct and Non-naming Attempt categories, the precision is less than 50%. When the model predicts a positive response for one of these categories, it is more likely than not a false positive.

T_1 and T_2 Statistics

Figure 4a shows the distribution of T_1 statistics for the observed and predicted data. The T_1 statistic measures the recovery of the mean category counts over subjects in the data. The statistic is larger in the predicted data for all but 4 of the 2,000 samples. This gives an extremely high p-value of 0.998, suggesting the discrepancy between the posterior mean and the observed data is less than that between the posterior means and the draws from the posterior predictive distribution. In other words, the average expected category counts are closer to the observed counts than the posterior predicted counts.

Figure 4b shows the distributions of T_2 statistics from the observed and predicted data. The T_2 statistic is larger for the predicted data for every draw from the posterior distribution. This gives a p-value of 1, indicating the MPT model is adequately recovering the covariance structure present in the data.



(a) T_1 Statistic

(b) T_2 Statistic

Figure 4: Distributions of T_1 and T_2 statistics from observed and posterior predictive data

Credible Intervals

Participant abilities were estimated for each of the 6 subject-dependent parameters. Credible intervals were calculated using the highest posterior density interval to observe the variability in these estimates. These 95% credible intervals are shown in Figure 5. The Sem parameter is a probability, limiting the support to be from 0 to 1 inclusive. As the mean ability estimate increases, the credible intervals for Sem become narrower and the confidence in the point estimate increases. For each processing step which is estimated by the Rasch model given in Equation 1, the variability in the participant ability estimate increases as the mean ability increases. That is, for subjects with a high mean ability estimate, there is more variability compared to a subject with a low mean estimate. This is reflected in wider credible intervals for subjects with higher ability.

5 Discussion

We replicated the MPT model on PNT data which had significant overlap with the data used by Walker et al. Our posterior predictive counts follow a similar trend to Walker et al. with the largest discrepancy being for the Correct category. Our data was composed of 61.6% Correct responses, while Walker et al. had 55.8% Correct. This difference in Correct responses contributes to our posterior predictive counts being more biased toward Correct. Since the MPT model is sensitive to data composition, users should be aware that the estimates provided by the MPT model may be extremely biased toward high category counts observed in the data.

The prediction accuracy of the MPT model shows the high bias toward the two most frequent response categories, Correct and Non-naming Attempt. These two categories together constitute 76.5% of the observed data. The sensitivity of the MPT model is between 0.5% and 28% for the other 6 categories, demonstrating poor recovery of these less common responses, which are possibly the most interesting as well. The specificity is high for infrequent response categories due to the rarity of the event. The low specificity for Correct is owed to the model over-predicting the category. The precision shows that while the model does predict each response category, it does so with limited accuracy. The lack of agreement between the observed data and the

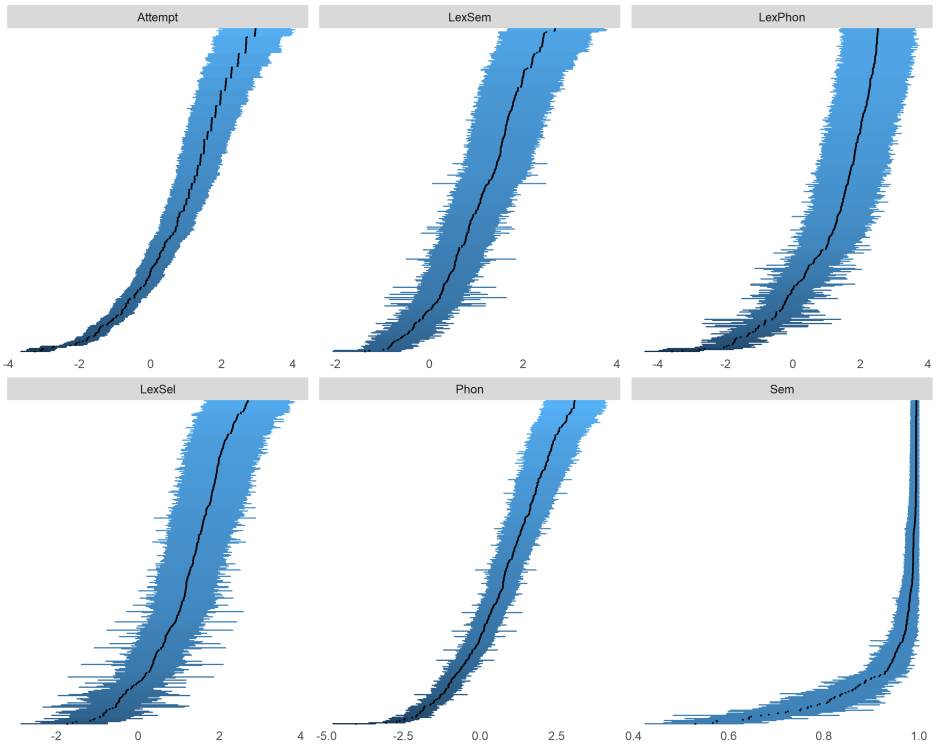


Figure 5: 95% Highest Posterior Density Credible Intervals

posterior predictive data reflects that the model is not fitting the data closely.

The T_1 and T_2 statistics are chi-squared-like tests designed to measure the discrepancy between the observed and expected mean and covariance, respectively. The p-value from this test is the proportion of times the T statistic is larger for the posterior predicted data than the observed data. The model is adequately recovering the mean or covariance structure if the p-value is not close to 0. The p-values given from the T statistics for the MPT model are very high. This suggests the posterior draws are fitting the observed data more closely than the posterior predicted data sets.

The credible intervals for participant abilities on the logit scale show increasing variability with higher point estimates. For a subject with high ability estimates, the point estimate may not be a reliable indicator for assessing recovery progress. Additionally, the credible intervals show the narrowest intervals, and therefore the highest confidence, in the point estimates for the Sem parameter and lower confidence for the estimates on the logit scale. Users should be aware of the difference in the width of the credible intervals for each cognitive step when considering the point estimates.

One potential area for improvement for the MPT model is the prior specification. The architecture of the MPT model allows for great flexibility in priors. Currently, the prior distribution for each bifurcation probability is standard normal or standard uniform. A priori, each bifurcation point has 50% probability on average of successful completion. While this appears non-informative at first, this prior specification heavily biases the outcome toward the Non-naming Attempt and Unrelated categories, see Table 3. The category counts predicted from the prior distribution do not resemble the pattern observed in the data, particularly for the Correct category. This discrepancy between the prior specification and the observed data is contributing to the poor recovery of the less common response types in the data.

A simple change to these bifurcation probabilities could account for Correct being the most

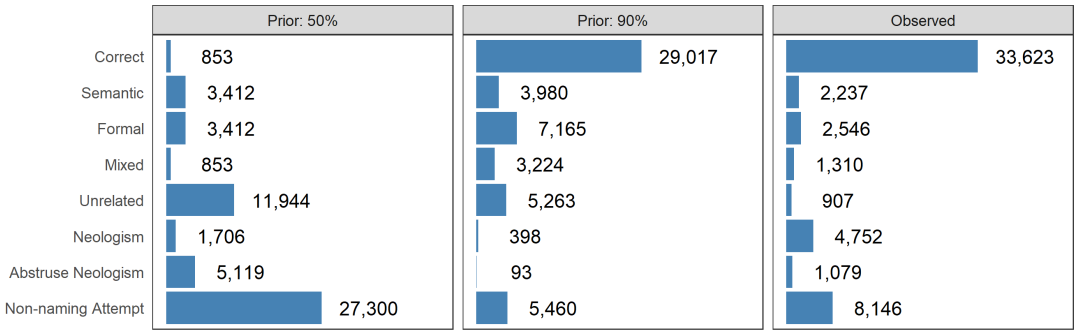


Figure 6: Prior Predictive Analysis

likely category and reduce the number of expected Non-naming Attempts. Setting the prior bifurcation probability to 90% for each process gives expected counts which more closely match the observed data. The expected counts a priori setting the bifurcation probabilities to 50% and 90% are shown in Figure 6.

Currently, the model assumes standard normal distributions for subject ability. This assumes the subject ability at each cognitive step is independent of the others. However, it seems possible that all of the cognitive steps are impacted by the subject's aphasia severity. Ideally, the model would account for the subject's overall naming ability and aphasia severity by using a multivariate prior distribution.

With some changes to prior specification, the goodness-of-fit and accuracy of the MPT model could be improved. Due to the item-specific effects and multiple latent parameters, the MPT model is a significant advancement in aphasia modeling. The theoretical cognitive steps represented in the model could potentially provide valuable and precise information to clinicians about a subject's particular condition. However, using the model with the current prior and parameter specifications may not perform reliably, and users should be cautious before making any decisions on patient treatment based off information obtained from the MPT model.

References

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

Fergadiotis, G., Kellough, S., and Hula, W. D. (2015). Item response theory modeling of the philadelphia naming test. *Journal of Speech, Language, and Hearing Research*, 58(3):865–877.

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, 75(1):70–98.

Mirman, D., Strauss, T. J., Brecher, A., Walker, G. M., Sobel, P., Dell, G. S., and Schwartz, M. F. (2010). A large, searchable, web-based database of aphasic performance on picture naming and other tests of cognitive function. *Cognitive neuropsychology*, 27(6):495–504.

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2.