



Published in final edited form as:

*Cogn Neuropsychol.* 2010 May ; 27(3): 230–244. doi:10.1080/02643294.2010.512285.

## Rasch Models of Aphasic Performance On Syntactic Comprehension Tests

Roe Gutman<sup>1</sup>, Gayle DeDe<sup>2</sup>, Jennifer Michaud<sup>3</sup>, Jun Liu<sup>1</sup>, and David Caplan<sup>3</sup>

<sup>1</sup> Department of Statistics, Harvard College

<sup>2</sup> Department of Speech, Language, and Hearing Sciences, University of Arizona

<sup>3</sup> Neuropsychology Laboratory, Massachusetts General Hospital

### Abstract

Responses of forty-two people with aphasia to eleven sentence types in enactment and sentence-picture matching tasks were characterized using Rasch models that varied in the inclusion of the factors of task, sentence type, and patient group. The best fitting models required the factors of task and patient group but not sentence type. The results provide evidence that aphasic syntactic comprehension is best accounted for by models that include different estimates of patient ability in different tasks and different difficulty of all sentences in different groups of patients, but that do not include different estimates of patient ability for different types of sentences.

---

This paper describes the use of Rasch models to analyze the performance of people with aphasia on tests of syntactically based comprehension. Rasch models have rarely been applied to the analysis of aphasic performance (one example is Donovan et al, 2007). There are, however, several ways they might be useful in understanding the nature of aphasic impairments.

Rasch models express the probability of a person correctly answering a test item as a linear function of the ability of the person and the difficulty of the item. In the simplest model, all individuals vary along a single ability dimension and all test items vary along a single difficulty dimension. In more complex models, the performance of an individual on an item differs for different individuals or groups of individuals and for different items or groups of items. Different models are assessed based on their goodness of fit to a dataset. Applying these methods to the performance of people with aphasia can help determine whether the best model groups all people with aphasia together with respect to how difficult they find items or divides them in some way. These methods can also help determine whether all test items fall along a single difficulty dimension or if they are best considered as being of different types with respect to how difficult they are. Thus, for instance, it would be possible to utilize Rasch models to ask whether dyslexia is best modeled as the result of individual differences in a single reading ability that applies to all words, or whether different words (say, regular and irregular words) have different levels of difficulty, possibly in different groups of people with dyslexia.

We applied Rasch models to the performances of 42 aphasic people with single left hemisphere strokes on two tests that have been used to assess syntactically based comprehension – sentence picture matching (SPM) and object manipulation (OM). The

performance of these patients has been previously reported in two publications: Caplan et al (2006, 2007). In those papers, we examined the performance of these patients both as individuals and as a group. At the individual level, no person with aphasia showed a selective deficit on sentences that required processing a particular syntactic structure on both tasks. At the group level, individuals in less well performing groups had disproportionately lower accuracy on sentences on which the group as a whole performed less well. Finally, rotated and unrotated Factor Analyses (FA) at the group level did not differentiate sentences with different syntactic structures. Together, these findings led to the view that syntactically based comprehension deficits in aphasia reduce the ability to apply parsing and interpretive processes to all sentences in a given task, with individual performance determined primarily by the level of a person's ability on a task and by the difficulty of a sentence type in that task.

This conclusion is inconsistent with much thinking about the nature of aphasic deficits of syntactically based comprehension. The most widely held view of these deficits sees them as specific structural deficits (loss of a particular syntactic structure; e.g., the claim that Broca's aphasics do not represent traces (Grodzinsky, 2000)), or specific processing deficits (inability to apply a type of parsing and interpretive operation; e.g., the claim that Broca's aphasics have abnormal abilities to activate more complex lexical argument structures (Thompson & Lee, 2009)). These theories claim that individuals with such deficits will have abnormal performance on sentences that contain the structure or require the operation in any comprehension task. They therefore predict that sentences will fall into different groups with respect with their difficulty, at least for certain aphasia types, and that these groups of sentences will be the same in all tasks.

Rasch models afford a new way to evaluate these different views of aphasic deficits in this domain. If the "specific deficit" approach is correct, models that do not introduce a task factor should be favored over those that do, and models that introduce patient and sentence type groupings should be favored over those that do not. If the view that aphasic deficits consist of reduced ability to apply parsing and interpretive processes to all sentences in a given task is correct, the opposite should be found: models that introduce a task factor should be favored over those that do not, and models that do not introduce patient and sentence type groupings should be favored over those that do. The analyses that follow examine Rasch models of aphasic syntactic comprehension performance with these questions in mind.

## Methods

Forty-two aphasic people with aphasia (mean age 60.3 years (range: 24.7 – 84.5); mean education, 14.7 years (range: 9 – 22); M:F = 26:16) were tested. Patients were required to be aphasic, right handed, have a single left hemisphere stroke, be able to perform the tests, and be judged to have adequate single word comprehension to not fail because of lexical semantic disturbances (patients were screened for disturbances of phoneme discrimination, auditory lexical decision and spoken word-picture matching, and ability to match words to pictures and objects in the syntactic comprehension tasks).

The ability to construct and understand three syntactic structures—passives, relative clauses, and sentences with reflexive pronouns—was tested by having patients respond to pairs of sentences in which the baseline sentence did not contain the construction/element in question or could be interpreted on the basis of a heuristic and the experimental sentence contained the structure/element and required the assignment of a complex syntactic structure to be understood. Each structure was tested with two experimental/baseline contrasts ("constructions"), with 10 examples of each sentence type (Table 1). Words in the sentences

were high frequency<sup>1</sup> and were identical in experimental and baseline versions of the sentences, thus controlling for frequency. Experimental and baseline versions of the sentences were synonymous (except for SO/SS, where this is impossible, and where they were thematically very similar), thus controlling for plausibility.

These structures were selected because they contain different syntactic structures that have been claimed to be selectively affected in comprehension in aphasia. In Chomsky's theory, passives and relative clauses contain phonologically "empty" referentially dependent noun phrases (NPs). Passive sentences contain one type of empty NP (an "NP-trace") and relative clauses contain a different type (a "wh-trace"). Some results indicate that some patients process these types of structures differently (e.g., Thompson and Shapiro (2007) reported that patients do not generalize successful training on one of these structures to the other). Reflexive pronouns are similar to "traces" in being obligatorily related to antecedents in a syntactic domain (contrasting with pronouns in this respect) but are overt, rather than phonologically empty, referentially dependent NPs. Several researchers have argued that deficits in some aphasic patients affect empty and spare overt referentially dependent NPs, such as reflexives, but affect empty ones, such as NP-traces (see Grodzinsky, 2000, for a summary). The selection of these structures thus provides an opportunity for structure-specific deficits in syntactically based comprehension that have been postulated to exist to be found.

Subjects were tested in enactment (object manipulation (OM)), picture matching (SPM) and grammaticality judgment (GJ) tasks, the latter two with both whole sentence and self-paced listening presentation conditions, using digitized computer-delivered auditory stimuli. In the enactment task, participants indicated thematic roles and co-indexation by manipulating paper dolls. People with aphasia were told that the purpose of the experiment was to test their abilities to understand "who did what to whom" in the sentences. They were instructed to indicate "who did what to whom" by acting out the sentence using the items provided. The experimenter emphasized that they did not need to show details of the action of the verb, but had to clearly demonstrate which item was accomplishing the action and which item was receiving it. In the sentence picture matching test, each sentence was played auditorily with the two drawings in full view of the participant, and the participant was required to choose the drawing that matched the sentence by pressing one of two buttons on a timer interfaced with the computer using fingers on the non-paretic hand. Accuracy and end-of-sentence reaction time (RT) (in SPM and GJ) were measured. We here report on the accuracy data from the OM and SPM tasks, which require comprehension, with whole sentence presentation, which is the most natural form of language presentation in a hearing subject.

Contingency tables for the Rasch models were created as follows. For each participant  $i$  ( $i \in 1, \dots, I$ ),  $I = 42$ , and each sentence type  $j$  ( $j \in 1, \dots, J$ ),  $J = 22$ , the number of examples administered ( $n_{ij}$ ) and the number of correct responses ( $y_{ij}$ ) was recorded, resulting in an  $I \times J$  contingency table ( $\mathbf{Y}$ ) with a maximum number of  $n_{ij}$  examples in each cell. The rows contained the participants and the columns contained different items (sentences in tasks) for which the response is either correct or not. In most cases  $n_{ij} = 10$ , but there are a significant number of cells for which  $n_{ij} = 9$  (62 cells), and one cell in which  $n_{ij} = 8$ .

We explored a series of Rasch models of these data. Rasch models describe the log-odds of the probability of a person providing the correct response to a sentence as a linear function of the ability of the person ( $\alpha_i$ ) and the difficulty of the sentence ( $\beta_j$ ), or in mathematical

<sup>1</sup>For nouns, the mean CELEX cumulative frequency was  $57 \pm 64.5$ ; range = 0–246 (lowered by *aunt* (3), *nephew* (2), and *niece* (0)); for verbs, the mean frequency was  $36 \pm 45$ ; range = 1 – 143 (lowered by *pinch* (2), *hug* (5) and *tickle* (1)).

terms,  $\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_i + \beta_j$ . In the simple Rasch model, the estimate of each individual's ability was based on the sum of correct responses for that individual in all of the sentences from both tasks (the row sum).

Models that introduce task or sentence type factors will be referred to as “extended models.” In the extended model with a task factor, each individual's ability was estimated separately for each task (the row sum for SPM and the row sum for OM). In the extended model with sentence type factors, the ability of each individual was estimated separately for groups of similar sentences. We grouped sentences according to syntactic structure—passives, relative clauses, and sentences with reflexive pronouns—including their baseline sentences.<sup>2</sup> In the extended model with both a task and sentence type factor, each participant's ability was estimated separately for groups of similar sentences in each task.

As noted, in the simple Rasch model, the probability of a correct response is modeled as a logistic function of the difference between the ability of a person and the difficulty of a sentence. The usual assumption is that  $y_{ij}$  is a dichotomous random variable. However, in our data  $y_{ij}$  has a binomial  $(n_{ij}, p_{ij})$  distribution (due to the forced choice procedure). As a result, the likelihood function has the form:

$$L(\tilde{\alpha}, \tilde{\beta} | Y_{ij}, n_{ij}) = \prod_{i=1}^I \prod_{j=1}^J \binom{n_{ij}}{y_{ij}} p_{ij}^{y_{ij}} (1-p_{ij})^{n_{ij}-y_{ij}}$$

where

$$\tilde{\alpha} = \{\alpha_i, i=1, \dots, I\} \quad \text{and} \quad \tilde{\beta} = \{\beta_j, j=1, \dots, J\}$$

Defining the row sums as  $r_i$  for  $i \in \mathbf{1}, \dots, I$  and the column sums as  $c_j$  for  $j \in \mathbf{1}, \dots, J$ ,

$$L(\tilde{\alpha}, \tilde{\beta} | Y_{ij}, n_{ij}) = \exp\left(\sum_{i=1}^I r_i \alpha_i - \sum_{j=1}^J c_j \beta_j\right) \prod_{i=1}^I \prod_{j=1}^J \binom{n_{ij}}{y_{ij}} \left(\frac{1}{1 + \exp(\alpha_i - \beta_j)}\right)^{n_{ij}}$$

The probability of the whole data matrix depends on the data only through the marginals (row and column sums) of the table and the parameters, not on the particular entries within the data matrix. Thus, the complete sufficient statistics for  $\alpha$  and  $\beta$  are  $(r_i, c_j)$  for  $i \in \mathbf{1}, \dots, I$  and  $j \in \mathbf{1}, \dots, J$ . This approach was extended to models incorporating task or sentence type groupings (note that from the computational point of view these are simply different ways of

<sup>2</sup>Because baseline sentences can be interpreted using linear heuristics, we also considered a grouping in which all baseline sentences formed one group and the experimental sentences were grouped by type. In the third sentence type grouping, we eliminated the truncated passive sentences because performance on these sentences strongly suggested that factors other than syntactic complexity (e.g., the presence of only one NP) greatly affected results, and formed two groups of sentences based upon the presence of a phonologically empty or phonologically overt referentially dependent noun phrase (“trace” (PF, CO, SO) and reflexives (RP, RG)). In the fourth sentence type grouping, all the experimental sentences were grouped together into one group. We also created groups of experimental sentences according to the syntactic features found in the experimental sentences ((PT, PF),(CO, SO) and (RP, RG)) and the corresponding sets of baseline sentences ((A), (CS, SS), (RPB, PGB)). The results did not differ from those reported here. Interested readers can obtain the details of these analyses from the corresponding author.

grouping the 22 columns). Thus, for example, for the extension of the simple Rasch model to include task effects, the likelihood for  $\mathbf{Y}$  is

$$\begin{aligned} L(\alpha, \beta | \mathbf{Y}) &= \prod_{i=1}^I \prod_{j=1}^J \frac{(\exp(\alpha_{i0} \delta\{j \in SPM\} + \alpha_{i1} \delta\{j \in OM\} - \beta_j))^{y_{ij}}}{(1 + \exp(\alpha_{i0} \delta\{j \in SPM\} + \alpha_{i1} \delta\{j \in OM\} - \beta_j))^{n_{ij}}} \\ &= \prod_{i=1}^I \prod_{j \in SPM} \frac{(\exp(\alpha_{i0} - \beta_j))^{y_{ij}}}{(1 + \exp(\alpha_{i0} - \beta_j))^{n_{ij}}} \prod_{i=1}^I \prod_{j \in OM} \frac{(\exp(\alpha_{i1} - \beta_j))^{y_{ij}}}{(1 + \exp(\alpha_{i1} - \beta_j))^{n_{ij}}} \\ &= \frac{\exp(\sum_{i=1}^I r_{i(SPM)} \alpha_{i0} - \sum_{i=1}^I r_{i(OM)} \alpha_{i1} - \sum_{j=1}^J c_j \beta_j)}{\prod_{i=1}^I \prod_{j \in SPM} (1 + \exp(\alpha_{i0} - \beta_j))^{n_{ij}} \prod_{i=1}^I \prod_{j \in OM} (1 + \exp(\alpha_{i1} - \beta_j))^{n_{ij}}} \end{aligned}$$

where  $r_{i(OM)}$  and  $r_{i(SPM)}$  are the row sums for OM and SPM respectively and  $c_j$  is the column sums. This likelihood function reveals that the sufficient statistics for this model are the row sums for the SPM task, the row sums for the OM task, and the column sums.<sup>3</sup>

A second issue is the application of Rasch model to multiple choice questions, since random guessing may not allow the probability of responding correctly to approach zero for very difficult items or for individuals with very low ability levels. This issue is not a major problem for our data, for the following reasons.

First, previous analyses of on-line performance in these aphasic individuals suggested that, contrary to the most commonly made assumption about aphasic performance (Grodzinsky, 2000), guessing was not a major factor in determining their responses. Analyses of self paced listening times in these people with aphasia for the words in the sentences showed normal effects of syntactic structure in sentences to which they responded correctly, and effects of syntax that differed from normal in sentences to which they responded incorrectly. In addition, the poorer performing individuals performed at below chance levels on more difficult sentences. These features of the data suggest that most responses are not guesses

<sup>3</sup>For the model with sentence groupings,

$$\text{logit}(p_{ij}) = \alpha_{i0} \delta\{j \in \text{Pass}\} + \alpha_{i1} \delta\{j \in \text{Reflex}\} + \alpha_{i2} \delta\{j \in \text{ObjRel}\} - \beta_j$$

For the model with task and sentence groupings,

$$\text{logit}(p_{ij}) = \sum_{k=1}^K \alpha_{ik} \delta\{j \in G_k\} - \beta_j$$

where

$$G = \{\{\text{Pass} \cap SPM\}, \{\text{Reflex} \cap SPM\}, \{\text{ObjRel} \cap SPM\}\} \cup \{\{\text{Pass} \cap OM\}, \{\text{Reflex} \cap OM\}, \{\text{ObjRel} \cap OM\}\}$$

and  $k$  is an index over the elements in  $G$ .

but rather true errors. Dickey et al. (2007) came to a similar conclusion on the basis of analysis of aphasic individuals' eye movements.

Second, the application of the Rasch model to binomially distributed data as opposed to dichotomous responses reduces the effects of guessing. When the Rasch model is applied to dichotomous responses, it is crucial to determine whether a correct response on the part of an individual on a specific item is due to guessing. This is important because correctly answering a difficult question may have large impact on the estimate of an individual's ability. When the data consist of the number of correct answers to items with similar structures, the effect of guessing diminishes, since the probability of responding correctly to a type of item ( $p_{ij}$ ) by guessing decreases as the number of questions ( $n_{ij}$ ) increases. Also, guessing is more likely to occur on difficult items. These two considerations result in a lower probability of correct responses for individuals with lower ability even if they are guessing randomly.

A second issue that applies to our data is the possibility that performance on particular sentences is influenced by factors such as plausibility or lexical frequency. This seems unlikely. The lexical items used in these tasks were high frequency nouns and verbs, and the frequency of lexical items did not differ among sentences of a given type or across sentences of different types. The tasks themselves – enactment and sentence picture matching – introduce a context that de-emphasizes real world plausibility (for instance, as with most such tests in use with aphasics, pictures of participants in actions were drawings, not photographs). Nonetheless, since the task was not entirely fictitious (as, for example, a test in which animals interact would be), real-world plausibility of the meanings of the sentences might have affected responses. To guard against that, all sentences were constructed so that both the correct and reversed thematic role assignments in each sentence were plausible. Previous work (Caplan et al, 1985) has shown that real world plausibility of actions in an enactment task did not affect responses (nor did location of items in arrays of manipulanda).

A final issue, related to the above, is whether presentation of multiple examples of each sentence type could have affected responses; that is, might responses reflect strategic operations conditioned by the presentation of several stimuli of a given type? There are three responses to this question. First, in general, strategic effects are not found in studies in which no more than 25% of the stimuli are of a given type. Second, strategically induced repeated similar responses to sentences of a given type would lead to large numbers of errors or large numbers of correct responses to particular sentence types in many patients. The data, however, show that large numbers of errors occurred on many sentence types in the poorest performing patients and large numbers of correct responses occurred on many sentence types in the best performing patients, a pattern that is more likely to reflect actual ability than the application of strategies. Third,  $r$  values for split-half reliability of accuracy on sentences of the same type were very close to, or lower than, those for the  $r$  values for correlations of accuracy on different sentence types within each task (for SPM mean split-half  $r = .53$ , mean cross-sentence  $r = .65$ ; for OM, mean split-half  $r = .73$ , mean cross-sentence  $r = .69$ ). This provides evidence that strategic factors did not lead to repeated similar responses to sentences of a given type.

For these reasons, the main issues associated with examining the Rasch model as a possible fit for the data were addressed, and we proceeded with such analyses. We then had to deal with several technical challenges.

One way to examine the validity of the Rasch model is by calculating the probability of certain test statistics under the assumption that the Rasch model is true. Using the idea of similar tests, one can obtain the exact distribution of a test statistic by conditioning on the

sufficient statistics of the Rasch model. However, obtaining the exact conditional distribution under the Rasch model and some of the extensions is a difficult task. It can be accomplished by (1) enumerating all the tables with the same row and column sums for which each cell is smaller than or equal to  $n_{ij}$ , (2) computing the test statistic for each of the enumerated tables, and (3) determining how often the value computed is equal to or exceeds the observed value of the test statistic using the Neyman-Pearson Lemma (Lehmann, 1998). If the Rasch model is inappropriate, we would expect that the frequency of obtaining the observed test statistic or exceeding it will be smaller than or equal to a predefined rejection level (by convention,  $p < .05$ ) (Ponocny, 2000). The enumeration of tables with fixed row sums and columns and cell values between 0 to  $n_{ij}$  is an intensive computational task. When  $n_{ij} = 1$ , each table with the same row and column sums is equally likely; however when  $n_{ij} > 1$  the distribution is no longer uniform over the space of tables. We used the MCMC approach proposed by Diaconis & Strumfels (1998) to sample from the space of available tables with fixed marginals. In this method, one iterates the following simple step: choosing two rows and two columns at random and changing the intersection cells according to one of the randomly selected patterns:

$$\begin{array}{cc} + & - \\ - & + \end{array} \quad \text{or} \quad \begin{array}{cc} - & + \\ + & - \end{array}$$

In cases where negative or more than  $n_{ij}$  correct answers had occurred in a cell, the random walk stayed at the current table.

We examined the performance of groups of people with aphasia using a mixture model that combines the theoretical strength of the Rasch model with the power of latent class analysis. The mixture model has the following form:

$$\sum_{k=1}^K \eta_k p(Y_{ij}=y_{ij} | \beta_k, \alpha_k, \sigma_\varepsilon^2)$$

where  $K$  is the number of groups,  $\eta_k$  is the probability of an individual belonging to group  $k$ ,  $\beta_k = (\beta_{1k}, \dots, \beta_{Jk})$  defines the sentence difficulty for group  $k$ ,  $\alpha_k$  is the common ability of the group, and  $\sigma_\varepsilon^2$  represent the deviation in an individual's ability from the average ability of the entire group. One type of model that allows for deviation from the common ability for individuals, while still allowing for sentence types to have different difficulty levels, defines  $p(Y_{ij}=y_{ij} | \beta_k, \alpha_k, \sigma_\varepsilon^2)$  as the following Generalized Linear Mixed Model (GLMM):

$$p(Y_{ij}=y_{ij} | \beta_k, \alpha_i, \varepsilon_i) = \binom{n_{ij}}{y_{ij}} \frac{\exp(\alpha_k + \beta_{kj} + \varepsilon_i)^{y_{ij}}}{(1 + \exp(\alpha_k + \beta_{kj} + \varepsilon_i))^{n_{ij}}}$$

This model allows each participant to have an ability parameter that differs from that of the whole group ( $\alpha_i \rightarrow \alpha_k$ ). The latent variable  $\varepsilon_i$  signifies the difference of each individual from the group's mean  $\alpha_k$ . We assumed that  $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ , where  $\sigma_\varepsilon^2$  is unknown. Since we were interested in the distribution of different test statistics described below as well as the ability to easily extend the model to incorporate additional latent structure, we decided to use a full Bayesian statistical analysis.

The distribution obtained from our Bayesian modeling does not have a closed analytical form, thus we chose to use Markov Chain Monte Carlo (MCMC) algorithms to perform the calculations. We used the Gibbs (Geman 1984) and Metropolis-Hasting procedures to

sample from the posterior distribution of the parameters and the latent structures. To conduct the MCMC sampling we introduced a latent indicator variable for each person indicating his group membership. Assuming that the number of clusters is known in advance and set to  $K$ , we let  $\gamma_i \in \{1, \dots, K\}$  define the group membership of participant  $i$ , resulting in the following model:

$$\begin{aligned} [y_{ij}|p_{ij}] &\sim \text{Bin}(p_{ij}, n_{ij}) \quad i \in \{1, \dots, I\}, j \in \{1, \dots, J\} \\ [p_{ij}|\gamma_i=k] &= \frac{\exp(\alpha_k + \beta_{kj} + \varepsilon_i)}{(1 + \exp(\alpha_k + \beta_{kj} + \varepsilon_i))} \quad k \in \{1, \dots, K\} \\ \varepsilon_i &\sim N(0, \sigma_\varepsilon^2) \end{aligned}$$

To complete the Bayesian modeling, we assumed the following prior distributions for the different model parameters:

$$\begin{aligned} \alpha_k &\sim N(0, \sigma_\alpha^2) \quad k \in \{1, \dots, K\} \\ \beta_{kj} &\sim N(0, \sigma_\beta^2) \quad k \in \{1, \dots, K\}, j \in \{1, \dots, J\} \\ \sigma_\varepsilon &\sim \Gamma(a, a) \quad i \in \{1, \dots, I\} \\ \gamma_i|\eta &\sim \text{multinom}(1, \eta) \quad i \in \{1, \dots, I\} \\ \eta &\sim \text{Dirich}(1, \dots, 1) \end{aligned}$$

To examine whether these prior distributions had any effect on the posterior analysis we fixed  $\sigma_\alpha^2$  and  $\sigma_\beta^2$  at different values in the range 6 – 10 in a 0.5 interval. This had little effect on the outcome, and both  $\sigma_\alpha^2$  and  $\sigma_\beta^2$  were set at five. For the prior distribution of  $\sigma_\varepsilon^2$  we set  $a$  in the  $\Gamma$  distribution to either 0.01 or 0.001. We also examined two other diffused priors ( $U[0, 100]$  and half-Cauchy) as suggested by Gelman (2006) and found that the different prior distributions had little effect on the outcome.

We monitored the convergence of the MCMC algorithm by both the Gelman & Rubin potential scale reduction ( $\hat{R}$ ) statistics (Gelman and Rubin, 1992) and visualization of the samples as a time series.  $\hat{R}$  measures how much improvement in the estimates would be possible by increasing the number of MCMC iterations. All estimated scalar parameters had  $\hat{R}$  values of less than 1, and the time series plots indicated convergence to a stable distribution.

In order to find the optimal number of participant clusters ( $K$ ), we explored  $K$  in the range of 2 – 7. Classifying the individuals using maximal posterior probability, we found that for  $K > 4$  all participants were classified into 4 clusters and therefore examined models where  $K \leq 4$ .

In order to examine both group and task effects, we applied the mixture GLMM (MGLMM) to each task independently as well as to both tasks jointly. Convergence was monitored in a fashion similar to the previous mixture model. The optimal number of clusters was decided based on the Bayesian Information Criteria (BIC) and by observing whether the posterior probability of any group was  $\approx 0$ .

Each model's adequacy was assessed in two ways. First, we calculated the goodness of fit of the model's performance to the observed data using the AIC (Akaike, 1974) and the BIC (Bayesian information criterion). Second, we calculated the probability of certain features of performance ("test statistics") occurring in the Monte Carlo simulations. For the Rasch model and some of the extensions, this probability can be estimated using the Diaconis & Strumfels (1998) method described earlier. This method cannot be applied to the MGLMM model, so we used the Bayesian idea of posterior predictive checks to obtain this distribution. The distribution of the test statistics is obtained by creating a set of predictive



simulations by sampling new data,  $y_{ij}$ , independently from the binomial distributions given  $p_{ij}$ . The  $p_{ij}$ s were calculated from independent samples of the variables appropriate for each model. These samples were obtained from the posterior distribution corresponding to each of the two models. Since we were also interested in comparing the adequacy of the different models we also applied the posterior predictive procedure to the simple and extended Rasch models (described above). This calculation provided a common measure with which to compare the probability of observing each of the test statistics under the MGLMM and the simple and extended Rasch models.

The first test statistic was the frequency and magnitude of instances in which there were more errors on baseline than on experimental sentences (“reversals”). This test statistic was considered a salient feature of the performance data for the following reason: If the assumptions about the operations needed to structure and interpret the experimental and baseline sentences are correct, reversals must reflect noise in the comprehension process. We measured both the frequency ( $F$ ) and magnitude ( $M$ ) of reversals for each sentence type:

$$F = \frac{1}{I \times |B^c|} \sum_{i=1}^I \sum_{j \in B^c} \delta(y_{ij} < y_{ij_B})$$

$$M = \frac{1}{I \times |B^c|} \sum_{i=1}^I \sum_{j \in B^c} \max(y_{ij_B} - y_{ij}, 0)$$

where  $\delta$  represents the indicator variable. We also measured the frequency ( $F_D$ ) and magnitude ( $M_D$ ) of reversals for each sentence type in each task.

$$F_D = \frac{1}{I \times |D \cap B^c|} \sum_{i=1}^I \sum_{j \in D \cap B^c} \delta(y_{ij} < y_{ij_B})$$

$$M_D = \frac{1}{I \times |D \cap B^c|} \sum_{i=1}^I \sum_{j \in D \cap B^c} \max(y_{ij_B} - y_{ij}, 0)$$

where  $D \in \{SPM, M\}$ .

The second test statistic was the correlation of performance of each sentence type across the SPM ( $j_{SPM} = j \in SPM$ ) and OM ( $j_{OM} = j \in OM$ ) tasks:

$$T_j = \frac{1}{I} \sum_{i=1}^I (y_{ij_{OM}} - \bar{y}_{\cdot j_{OM}})(y_{ij_{SPM}} - \bar{y}_{\cdot j_{SPM}})$$

where

$$\bar{y}_{\cdot j_{SPM}} = \frac{1}{I} \sum_{i=1}^I y_{ij_{SPM}} \quad \text{and} \quad \bar{y}_{\cdot j_{OM}} = \frac{1}{I} \sum_{i=1}^I y_{ij_{OM}}$$

This second test statistic is psychologically salient because it is related to the claim that patients have deficits that affect parsing or interpretive operations independently of task.

## Results

Table 2 reports the AIC and BIC for the simple Rasch model, the extended Rasch models with factors of task, sentence type, and both, and MGLMM models. The extended model with task and sentence type factors has the best AIC but the worst BIC. The model with three groups in OM and two in SPM tasks (MGLMM<sub>[OM  $K = 3$ , SPM  $K = 2$ ]</sub>) has the best BIC and the fourth best AIC.

Tables 3 and 4 report the analyses of the number and magnitude of the reversals of performance in these models (test statistic 1). Tables 5 and 6 show the results of the analysis of the correlation of the number of correct responses for the same sentence types across tasks (test statistic 2).

For the observed number and magnitude of better performances on baseline than experimental sentences, the simple model without any factors did not fit the observed data well, generating more reversals than occurred in the patients' performances. The addition of sentence type and task factors improved model performance marginally. When comparing the posterior predictive probabilities of observing the number and magnitude of "reversals," the MGLMM<sub>[OM  $K = 3$ , SPM  $K = 2$ ]</sub> model had the best performance. For the correlation of performances on the same sentences across tasks, all models at least partially simulate the observed data. When the posterior predictive p-values for these correlations were considered, models that introduced sentence type were inferior to those that did not, and models that introduced a task factor better simulated the observed correlations. The MGLMM<sub>[OM  $K = 3$ , SPM  $K = 2$ ]</sub> model best captured the between-task correlation statistic.

## Discussion

We consider these analyses for their implications for the nature of aphasic deficits in syntactically based sentence comprehension and for the grouping of aphasic individuals with respect to these abilities. Before turning to this question, we consider the issue of which model is best overall. As Table 2 shows, the model with the lowest AIC included task and sentence type factors, but this model produced the highest BIC. The model with the lowest BIC was the MGLMM<sub>[OM  $K = 3$ , SPM  $K = 2$ ]</sub> model. This model had a relatively low AIC and best matched the observed test statistics. Based on these observations, we believe the weight of evidence supports the MGLMM<sub>[OM  $K = 3$ , SPM  $K = 2$ ]</sub> model.

We emphasize that studying more people with aphasia is important to explore these models more fully. In the analyses reported here, two levels of latent structure were included which could be influenced by the sample size. The first latent structure was applied to the patients grouping. It is possible that there are more than 3 groups in the OM task and more than 2 groups in the SPM, but since there are only 42 participants with aphasia in the study only these number of groups were detected. The second latent structure is the variation of a person's ability from the total group's ability. In cases where the number of people in a group is small the estimation of this variation may not be reliable. This could be especially critical in the Group 2 and Group 3 in the OM task where the numbers of people in each group were 4 and 7 correspondingly. For these reasons, including more patients is highly desirable.

Based on the data available, the results show that grouping the sentence types by task leads to better fitting models, while dividing the sentence types into groups along linguistic lines either leads to no improvement or to worse fitting models. *This result is consistent with the view of aphasic deficits outlined in Caplan et al (2006, 2007), according to which deficits consist of reductions in a person with aphasia's ability to apply parsing and interpretive operations to sentences in order to perform particular comprehension tasks. As discussed in*

the Introduction, if people with aphasia had specific deficits affecting particular syntactic elements (e.g., “traces”) or particular types of operations (e.g., “co-indexation of traces”), Rasch models that attributed different abilities for particular sets of sentences in both tasks would be expected to provide the best fits to the data. Instead, dividing performance on the basis of task resulted in the best model.

Models that describe patients’ performances as resulting from specific deficits usually confine their claims to subsets of patients. For example, the claim that patients have lost the ability to “co-index traces” has been made for Broca’s aphasics (Grodzinsky, 2000). We explored the possibility that the best models might include clustering of patients, and found that a model that divided patients into three clusters in OM and two in SPM provided a good fit to the data (by most measures, the best fit). The fact that these patient groupings differed for OM and SPM is not predicted by “specific deficit” models, which lead to the expectation that there will be groups of patients whose abilities differ for particular sentence types on both tasks.

Despite the above results, we looked for other evidence from the analyses performed here that there are patients who performed similarly on OM and SPM. We identified patients who did not change between the better and worse performing groups on both tasks, and those who changed from one group to another. Twenty-two patients were in the better performing group (Group 1) in both tasks; three were in the worst performing group (Group 2) in both tasks; and seventeen changed from one group to another across tasks. We examined the correlation of performance of patients on each sentence type across the two tasks, separately for patients who did and who did not change from one group to another, to determine whether patients who remained in groups that performed at the same relative level showed more consistency of performance across tasks than those who changed from one group to another across tasks, and whether they showed high correlations of performance on particular sentence types across tasks. As expected because of the small  $n$ , for patients who fell in the low performing groups in both tasks, most correlations were not significant (only performance on SO sentences was significantly correlated across tasks). The results for the other two groups are shown in Table 7. There is no indication that patients who remained in the same relative group showed more consistency of performance across tasks than those who did not, or showed particularly high  $r$  values for certain sentences. Therefore, this analysis of the data does not provide support for structure- or construction-specific task-independent deficits.

A final point is that it is possible that deficits affecting parsing and interpretation of specific structures would interact with task demands, leading to different patterns of performance on specific sets of sentences in two tasks. It is possible that such interactions might differ for different groups of patients because of particular problems they have in accomplishing particular tasks. To see whether the groups in the  $MGLMM_{[OM\ K=3, SPM\ K=2]}$  model showed task-specific structure- or construction-specific deficits, we examined their performance on each sentence type (Table 8). In both tasks, the groups differed primarily in their overall level of performance, with some differences in how the groups performed on specific sentences types. For instance, patients in Group 2 in OM did well on truncated passives and had no correct responses to subject-extracted relative clauses; patients in Group 3 in OM did poorly on truncated passives and on both subject-and object-extracted relative clauses. There does not appear to any particular linguistic structure or psycholinguistic operation that makes for poor performance of the patients in these groups in either task.

Though not directly relevant to the issues above, we examined patients in the groups for clinical syndrome and common features of lesions. Patients were classified into clinical syndromes based on the BDAE and diagnosed as agrammatic based on criteria in the Oral

Production Test (Goodglass, Christiansen, & Gallagher, 1993). The groups did not contain patients with classical aphasic syndromes. Of the six Broca's aphasics in the sample, 4 were clustered into Group 1 and 2 were clustered into Group 2 in OM, and 1 was clustered into Group 1 and 5 into Group 2 in SPM. Where available, MR and FDG PET scans were examined (see Caplan et al, 2007b for methods); the clusters did not contain patients whose lesions shared site, size, or metabolic activity level.

In summary, the finding that sentence comprehension abilities are best estimated for different groups of patients in different tasks suggests that the abilities of patients are best understood as enabling them to understand sentences and demonstrate their understanding in a particular task. This is consistent with the view that sentence comprehension is "situated" (Tanenhaus et al, 1995; Kamide et al, 2003) -- that is, that it proceeds incrementally alongside of accomplishing the task to which it is applied -- and that deficits affect the integrated function of incrementally determining sentence meaning and accomplishing task demands. Two aspects of the results also support the "resource reduction" model of aphasic deficits in syntactically based comprehension (Caplan, 2009a,b; Caplan et al, 2006, 2007). First, the finding that introducing a sentence type factor did not improve, or worsened, the Rasch model supports the view that people with aphasia's performance results from a reduction in the ability to apply syntactic and interpretive operations to all sentences, with different sentence types making different overall demands on these resources. Second, the aphasic groups identified in the best-fitting model differed primarily in terms of their overall level of performance. The results do not rule out the possibility that patients with task-independent specific deficits exist, but they confirm previous analyses showing that such patients must be very rare. To date, there are no reported cases of a person with aphasia who has shown a selective deficit in one aspect of syntactically based comprehension on more than one task that requires comprehension.

Finally, these results suggest that an important research topic is the way in which syntactic structure and task demands interact to determine abnormal performances in people with aphasia. Recent work has begun to explore the relation between parsing and interpretation and task performance, and has shown that features of visual arrays incrementally affect parsing and interpretive operations (e.g., Tanenhaus et al, 1995) and that features of sentences incrementally influence aspects of task performance (Kamide et al, 2003). The analyses reported here and related results point to the potential importance of these interactions for understanding the nature of the disruption of syntactically based sentence comprehension found in aphasic patients.

## References

- Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 1974;19:716–723.
- Caplan, D. The Cognitive Neuroscience of Syntactic Processing. In: Gazzaniga, M., editor. *The Cognitive Neurosciences*. Vol. IV. MIT Press; Cambridge, MA: 2009a.
- Caplan, D. Neural Organization for Syntactic Processing as Determined by Effects of Lesions: Logic, Data, and Difficult Questions. In: Bickerton, D.; Szathmáry, E., editors. *Report of the Strungmann Forum on the Evolution of Language*. 2009b.
- Caplan, D.; Hildebrandt, N. *Disorders of Syntactic Comprehension*. Cambridge, Mass: MIT Press (Bradford Books); 1988.
- Caplan D, Waters G, Hildebrandt N. Syntactic determinants of sentence comprehension in aphasic patients in sentence-picture matching and enactment tasks. *Journal of Speech and Hearing Research* 1997;40:542–555.
- Caplan D, DeDeGMichaud J. Task-independent and task-specific syntactic deficits in aphasic comprehension. *Aphasiology* 2006;20:893–920.

- Caplan D, Waters G, Kennedy D, Alpert A, Makris N, DeDe G, Michaud J, Reddy A. A Study of Syntactic Processing in Aphasia I: Psycholinguistic Aspects. *Brain and Language* 2007;101:103–150. [PubMed: 1699989]
- Caplan D, Waters G, Kennedy D, Alpert N, Makris N, DeDe G, Michaud J, Reddy A. A study of syntactic processing in aphasia II: Neurological aspects. *Brain and Language* 2007b;101:151–177. [PubMed: 16997366]
- Caramazza A, Zurif ER. Dissociation of algorithmic and heuristic processes in language comprehension: Evidence from aphasia. *Brain and Language* 1976;3:572–582. [PubMed: 974731]
- Chen Y, Diaconis P, Holmes SP, Liu JS. Sequential monte carlo methods for statistical analysis of tables. *Journal of the American Statistical Association* 2005;469:109–119.
- Cupples L, Inglis AL. When task demands induce “asyntactic” comprehension: A study of sentence interpretation in aphasia. *Cognitive Neuropsychology* 1993;10:201–234.
- Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society Series B* 1977;39(1):1–38.
- Diaconis P, Sturmfels B. Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics* 1998;26(1):363–397.
- Dickey M, Choy J, Thompson C. Real-Time Comprehension of Wh-Movement in Aphasia: Evidence from Eyetracking while Listening. *Brain and Language* 2007;100:1–22. [PubMed: 16844211]
- Donovan NJ, Vellozo CA, Rosenbek JC. The communicative effectiveness survey: Investigating its item-level psychometric properties. *J Med Speech-LangPath* 2007;15:433–447.
- Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Statistics* 2006;1(3):515–533.
- Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science* 1992;7(4):457–472.
- Gelman A, Meng XL, Stern H. Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statist Sinica* 1996;6:733–807.
- Geman S, Geman D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1984;6 (6):721–741.
- Goodglass H, Christiansen JA, Gallagher RE. Comparison of morphology and syntax in free narrative and structured tests: Fluent vs. non-fluent aphasics. *Cortex* 1993;29:377–407. [PubMed: 8258282]
- Grodzinsky Y. The neurology of syntax: Language use without Broca’s area. *Behavioral and Brain Sciences* 2000;23:47–117.
- Hula W, Doyle P, McNeil M. Rasch Modeling of Revised Token Test Performance: Validity and Sensitivity to Change. *Journal of Speech, Language, and Hearing Research* 2006;49:27–46.
- Lehmann, EL. *Testing Statistical Hypotheses*. Wiley; 1986.
- Linebarger MC. Agrammatism as evidence about grammar. *Brain and Language* 1995;50:52–91. [PubMed: 7552230]
- Linebarger MC, Schwartz MF, Saffran EM. Sensitivity to grammatical structure in so-called agrammatic aphasics. *Cognition* 1983;13:361–392. [PubMed: 6683142]
- Kamide Y, Altmann GTM, Haywood S. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye-movements. *Journal of Memory and Language* 2003;49:133–159.
- Meng X-L. Posterior predictive  $p$ -values. *Ann Statist* 1994;22:1142–1160.
- Ponocny I. Nonparametric goodness-of-fit tests for the Rasch model. *Psychometrika* 2001;66(3):437–460.
- Ponocny I. Exact person fit for Rasch model for arbitrary alternatives. *Psychometrika* 2000;65(1):29–42.
- Rost J. Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement* 1990;14:271–282.
- Rubin DB. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann Statist* 1984;12:1151–1172.
- Thompson CK, Shapiro LP. Complexity in treatment of syntactic deficits. *American Journal of Speech Language Pathology* 2007;16(1):30–42. [PubMed: 17329673]

- Thompson C, Lee Y. Psych verb production and comprehension in agrammatic Broca's aphasia. *J Neurolinguistics* 2009;22:354–369. [PubMed: 20174592]
- Wagenmakers EJ, Farrell S. AIC model selection using Akaike weights. *Psychonomic Bulletin & Review* 2004;11:192–196. [PubMed: 15117008]

**Table 1**

Sentence types used in the Sentence-Picture Matching and Object Manipulation tasks (in SPM, sentences had sentence-final adverbs because of the additional SPM self-paced study).

<u>Passives (NP trace)</u>
Full Passive
Experimental Sentence (FP): The woman was chased by the man
Baseline Sentence (A): The man chased the woman
Truncated Passive
Experimental Sentence (TP): The woman was chased
Baseline Sentence (A): The man chased the woman
<u>Relative clauses (wh trace)</u>
Cleft Object
Experimental Sentence (CO): It was the woman who the man chased
Baseline Sentence (CS): It was the man who chased the woman
Subject Object Relative
Experimental Sentence (SO): The man who the woman chased followed the girl
Baseline Sentence (SS): The man who chased the woman followed the girl
<u>Reflexives</u>
Reflexive with Genitive
Experimental Sentence (RG): The brother of the king shaved himself
Baseline Sentence (RG-B): The brother of the king shaved the man
Reflexive Possessive (RP)
Experimental Sentence (RP): The king's brother shaved himself
Baseline Sentence (RP-B): The king's brother shaved the man

Table 2

Model Fit

Model	AIC	BIC
Simple Rasch	2900	3203
Rasch with task groupings	2785	3291
Rasch with sentence groupings	2866	3566
Rasch with task and sentence groupings	2730	4025
MGLMM All Data $K = 2$	2961	3193
MGLMM All Data $K = 3$	2830	3178
MGLMM All Data $K = 4$	2792	3255
MGLMM All Data $K = 5$	2846	3425
MGLMM OM $K = 3 + \text{SPM } K = 2$	2824	3143



**Table 3**

Probability ( $p$ -value) of observed occurrence of better performance on Baseline than Experimental sentences occurring in Rasch models with different factors. “Reversals (F)” indicates average number of such occurrences; “Reversal Magnitude (M)” indicates average magnitude of these occurrences. Observed reversals that fall within the range of reversals found in each model have non-significant  $p$ -values (at the 0.05 rejection level) and are shown in bold italic.

	Reversals Value	<i>Two-tailed p-value</i>					
		Simple Rasch model	Sentence Type Factor	Task Factor	Task and Sentence Type Factor	Task and Sentence Type Factor	Task and Sentence Type Factor
All Reversals (F)	2.214	0.00005	0.0001	0	0	0.0005	
All Reversal Magnitude (M)	4.52	0	0	0	0	0	
SPM Reversals (F <sub>SPM</sub> )	1.21	0.0005	0.0001	0.0004	0	0	
SPM Reversal Magnitude (M <sub>SPM</sub> )	2.21	0	0	0	0	0	
OM Reversals (F <sub>OM</sub> )	1	0.056	<i>0.11</i>	<i>0.18</i>	<i>0.3</i>		
OM Reversal Effect (M <sub>OM</sub> )	2.31	0.023	0.07	<i>0.1</i>	<i>.15</i>		
Total number non-significant $p$ values		0	1	2	2	2	

Table 4

*p*-values from posterior predictive distributions (Bayesian *p*-values) of better performance on Baseline than Experimental sentences in Rasch models with different factors. “Reversals (F)” indicates average number of such occurrences; “Reversal Magnitude (M)” indicates average magnitude of these occurrences. Observed reversals that fall within the range of posterior predictive distributions found in each model have non-significant *p*-values (at the 0.05 rejection level) and are shown in bold italic.

		Posterior Predictive Two Tailed <i>p</i> -value for Occurrence of Statistic					MGLMM OM <i>K</i> = 3 SPM <i>K</i> = 2
	Reversals Value	Simple Rasch model	Sentence Type Factor	Task Factor	Sentence Grouping and Task Factor		
All Reversals (F)	2.214	<b>0.8</b>	<i>0.91</i>	<b>1</b>	<i>0.59</i>	<b>0.23</b>	
All Reversal Magnitude (M)	4.52	<b>0.21</b>	<i>0.11</i>	<b>0.16</b>	<i>0.054</i>	<b>0.95</b>	
SPM Reversals (F <sub>SPM</sub> )	1.21	<b>0.46</b>	<i>0.69</i>	<b>0.38</b>	<i>0.63</i>	<b>0.19</b>	
SPM Reversal Magnitude (M <sub>SPM</sub> )	2.21	<b>0.85</b>	<i>0.99</i>	<b>0.77</b>	<i>0.998</i>	<b>0.52</b>	
OM Reversals (F <sub>OM</sub> )	1	<b>0.67</b>	<i>0.45</i>	<b>0.29</b>	<i>0.12</i>	<b>0.88</b>	
OM Reversal Effect (M <sub>OM</sub> )	2.31	0.03	0.014	0.01	0.003	<b>0.38</b>	
Total number non-significant <i>p</i> values		5	5	5	5	6	

**Table 5**

Probability (*p*-value) of the observed correlation between sentence types in SPM and OM under the assumption of Rasch models with different factors. When the model is well fit to the data, the observed correlations fall within the range of correlations found in each model. Such correlations have non-significant *p*-values (at the 0.05 rejection level) and are shown in bold italic.

Sentence Type	Observed Correlation Across SPM and OM	Two tailed <i>p</i> -value				
		Simple Rasch model	Sentence Grouping Factor	Task Factor	Sentence Grouping and Task Factor	
A	0.572	0.036	<b>0.073</b>	0.036	<b>0.053</b>	
PF	0.654	0.044	<b>0.08</b>	0.041	0.036	
PT	0.138	<b>0.073</b>	0.0053	<b>0.14</b>	0.026	
RG	0.635	<b>0.136</b>	<b>0.14</b>	<b>0.09</b>	<b>0.11</b>	
RGB	0.518	<b>0.82</b>	<b>0.86</b>	<b>0.66</b>	<b>0.73</b>	
RP	0.434	<b>0.44</b>	<b>0.52</b>	<b>0.4</b>	<b>0.44</b>	
RPB	0.52	<b>0.35</b>	<b>0.38</b>	<b>0.28</b>	<b>0.31</b>	
CO	0.70	0.019	0.0037	0.015	0.006	
CS	0.518	<b>0.06</b>	0.029	<b>0.058</b>	0.038	
SO	0.54	<b>0.94</b>	<b>0.41</b>	<b>0.73</b>	<b>0.48</b>	
SS	0.45	<b>0.79</b>	<b>0.60</b>	<b>0.97</b>	<b>0.48</b>	
Total number non-significant <i>p</i> values		8	8	8	7	

**Table 6**

Probabilities from posterior predictive distributions (Bayesian *p*-values) of the correlation between Sentence Types in SPM and OM in Rasch models with different factors. Observed correlations that fall within the range of posterior predictive distributions found in each model have non-significant *p*-values (at the 0.05 rejection level) and are shown in bold italic.

Sentence Type	Observed Correlation Across SPM and OM	Posterior Predictive Two Sided <i>p</i> -value for Correlation Statistic						MGLMM OM <i>K</i> = 3 SPM <i>K</i> = 2
		Simple Rasch models	Sentence Grouping Factor	Task Factor	Sentence Grouping and Task Factor			
A	0.572	<i>0.55</i>	<i>0.98</i>	<i>0.25</i>	<i>0.07</i>	<i>0.20</i>		
PF	0.654	<i>0.78</i>	<i>0.26</i>	<i>0.38</i>	0.047	<i>0.26</i>		
PT	0.138	0	0	0.00026	0.007	<i>0.06</i>		
RG	0.635	<i>0.24</i>	<i>0.082</i>	<i>0.63</i>	<i>0.28</i>	<i>0.62</i>		
RGB	0.518	0.0053	0.0008	<i>0.31</i>	<i>0.74</i>	<i>0.76</i>		
RP	0.434	0.055	0.013	<i>0.49</i>	<i>0.84</i>	<i>0.94</i>		
RPB	0.52	<i>0.064</i>	0.017	<i>0.68</i>	<i>0.86</i>	<i>0.89</i>		
CO	0.70	<i>0.94</i>	<i>0.9</i>	<i>0.16</i>	0.02	<i>0.23</i>		
CS	0.518	<i>0.64</i>	<i>0.63</i>	<i>0.3</i>	<i>0.14</i>	<i>0.21</i>		
SO	0.54	0.0007	0.0013	<i>0.33</i>	<i>0.84</i>	<i>0.66</i>		
SS	0.45	0.0009	0.0016	<i>0.10</i>	<i>0.68</i>	<i>0.97</i>		
Total number non-significant <i>p</i> values		6	5	9	8	11		

**Table 7**

Correlation (Pearson's  $r$ ) of performance of performance across tasks in patients who stayed within the higher performing group in both tasks and those who changed from groups across tasks.

	Group	
	High (n = 22)	Change (n = 17)
Active	0.11	.58**
Full Passive	.51**	.58**
Truncated Passive	-0.2	-0.2
Cleft Subject	.47*	.61**
Cleft Object	0.3	0.37
Subject Subject	0.01	.52*
Subject Object	-0.09	0.33
Reflexive Genitive	.56**	.58**
Reflexive Genitive Baseline	-0.11	0.44
Reflexive Possessive	.52**	-0.07
Reflexive Possessive Baseline	0.27	0.22

\*  $p < .05$ ;

\*\*  $p < .05$ ;

Table 8

Mean percent correct (sd) of patients in different groups

	Sentence-Picture Matching		Object Manipulation		
	Group 1 <i>n</i> = 26	Group 2 <i>n</i> = 16	Group 1 <i>n</i> = 31	Group 2 <i>n</i> = 4	Group 3 <i>N</i> = 7
Active	97.4 (6.5)	77.1 (13.1)	96.5 (7.1)	85 (19.1)	98.6 (3.8)
Full Passive	98.3 (5.2)	63.2 (21.7)	91.9 (16.8)	45 (35.1)	90 (18.3)
Truncated Passive	97 (5.9)	64.6 (24.4)	90.6 (15.7)	95 (5.8)	30 (34.6)
Cleft Subject	96.2 (7)	81.3 (10.9)	98.1 (6)	87.5 (12.6)	91.4 (14.6)
Cleft Object	91.5 (10.1)	54.4 (19.3)	87.4 (19.7)	57.5 (20.6)	85.7 (18.1)
Subject Subject	90 (10.6)	73.8 (17.1)	77.7 (31.1)	0 (0)	28.9 (40.1)
Subject Object	80.4 (16.4)	51.3 (18.2)	45.8 (31.7)	12.8 (9.5)	11.8 (10.7)
Reflexive Genitive	92.7 (12.2)	61.3 (19.6)	80.3 (30.2)	52.5 (34)	81.4 (20.4)
Reflexive Genitive Baseline	86.9 (16.2)	58.8 (19.6)	82.9 (24.8)	15.3 (5.5)	85.7 (21.5)
Reflexive Possessive	94.2 (7.1)	74.4 (21.3)	84.8 (27.6)	87.5 (15)	82.9 (12.5)
Reflexive Possessive Baseline	95 (7.1)	64.4 (21)	86.5 (21.2)	37.5 (15)	87.1 (16)