The Effects of Sample Size and Guessing on Parameter Recovery in IRT Modeling of Aphasia Test Data

Item response theory (IRT) models are increasingly being applied to the construction and psychometric evaluation of performance-based aphasia tests (Fergadiotis, Wright, & Capilouto, 2010; Hula, Donovan, Kendall, & Gonzalez-Rothi, 2010; Hula, Kalinyak-Fliszar, & Martin, 2009; Kendall et al., 2010). One important consideration in this work is sample size. Recommendations for the minimum necessary sample sizes for accurate estimation of model parameters range from 50 for the simplest models (1-parameter (1P) or Rasch models) to 1000+ for more complex models (2- and 3P models) (Embretson & Reise, 2000; Linacre, 1994). Because sample sizes available in aphasia research are typically small, all IRT modeling in aphasia to date has been restricted to 1P models, which may not always adequately fit the empirical data. Moreover, the existing sample size recommendations are generally drawn from the educational literature and rest on assumptions that may not hold for aphasia tests or the population of persons with aphasia.

A second issue concerns the appropriateness of the IRT model to the data. Although all four studies cited above included items where guessing may be a factor (e.g., as in the Pyramids and Palm Trees (PPT), where chance performance is nominally 50%), only one (Fergadiotis et al., 2010) used a model that accounted for this feature of the data. Previous research has suggested that failing to incorporate the assumption of correct guessing into IRT models can lead to less accurate estimation of item and person parameters (Barnes & Wise, 1991).

The purposes of this study were to (1) determine if a 1P or 2P model augmented with the assumption of correct guessing recovered item and/or person parameters with greater accuracy than a standard 1p model, and (2) whether samples characteristic of those available in aphasia test development are adequate for estimating these models. We accomplished this by simulating data for the three-pictures version of the PPT, with simulation parameters based on an empirical data set.

METHOD

The empirical data used to develop the simulation parameters were collected from a sample of 70 persons with aphasia who were given the PPT for research purposes. These 70 participants were part of a larger sample of 111 previously reported by Martin and colleagues (2006). In the current data set, 7 PPT items were answered correctly by all subjects and an additional 5 items obtained corrected item-total correlations ≤ 0, suggesting that they are minimally related to the target construct. These 12 items were excluded from the analyses used to establish the simulation parameters. Demographic data and descriptive statistics for PPT performance in this sample are presented in Table 1.

The 40 remaining items were analyzed using a modified version of the 2P model (Mod2P) that included the assumption that the chance of correct guessing was 45%. The simulated item and person parameter distributions were chosen based on the empirical estimates. By default, person ability was scaled to have a mean ≈ 0 and a standard deviation ≈ 1, and item difficulty was scaled in reference to the person mean. Descriptive statistics for item parameter (discrimination and difficulty) and person parameter (ability) estimates are presented in Table 2.

Data were simulated under a 3P model using WinGen 3.01. For each simulated person-item combination, the IRT model equation was used to determine the probability of a correct response. This probability was compared to a number drawn randomly from a uniform (0,1) distribution. If the number was higher than the calculated probability, an incorrect response was recorded, and if the number was lower, a correct response was recorded.

Each simulation included 40 items and 75, 200, or 2000 examinees, with 1000 replications for each sample size condition for item estimation, and 100 replications for ability estimation. The smaller number of replications for ability estimation was due to the greater computational demands of aggregating those results. The n=75 condition was chosen to represent the lower acceptable bound of sample sizes that would be maximally feasible to obtain from a population of persons with aphasia. The n=2000 condition was chosen to represent an ideal sample size that would demonstrate the best performance of all four models under consideration. The n = 200 condition was chosen to represent an intermediate sample size that is potentially obtainable from an aphasic population.

The discrimination parameters used to generate the simulated data were chosen from a lognormal distribution with mean=0.14 and sd=0.57. Item difficulty parameters were chosen from a beta distribution with α=1.8, β=3.6, with a constant of 0.76 subtracted from each value to give a mean of -1.60. Pseudo-guessing parameters, which represent the probability of a correct response by persons of infinitely low ability, were chosen from a normal distribution with mean=0.50, sd=0.03. Person ability parameters were chosen from a beta distribution with α=4.2, β=4.0. Descriptive statistics are presented for the generating item and person parameters in Table 2.

ANALYSIS AND RESULTS

MULTILOG 7.03 was used to analyze each simulated data set with four different IRT models: the standard 1P, Mod1P, Mod2P, and 3P models. The models are presented in Table 3. Item parameters were estimated first, using the default marginal maximum likelihood method. There were many simulation runs on which the IRT model failed to adequately converge on a stable estimate for one or more items, especially in the n=75 conditions. Also, even runs with adequate convergence obtained extreme item parameter estimates in many cases. For these reasons, we excluded from the item estimation error analyses all observations where the difficulty value was <-4 or >4 and observations where the discrimination estimate was <0 or >4. The proportion of observations included in the item error analysis for each condition is displayed in Table 4.

Person ability was estimated in a second step using maximum a posteriori estimation, with the item parameter estimates obtained in the previous step treated as known values. Extreme item estimates were included in the ability estimation step because excluding them on a replication-by-replication basis was not feasible.

Accuracy of item and person parameter recovery was evaluated by comparing the estimates obtained in each model-by-sample size condition to the generating parameters, averaged across replications. Specifically, we calculated the root-mean-square error (RMSE), constant error (bias), and the correlation between the generating and estimated parameters for each condition. These are shown in Tables 5-7.

DISCUSSION

In general, the results suggest that the Mod1P model performed best under sample size conditions realistic for aphasia test development. However, none of the models performed particularly well. Item difficulty RMSEs ≈ 0.25 and person ability RMSEs ≈ 0.40 have been taken as evidence of precise estimation in previous simulation studies (Barnes & Wise, 1991; Hulin, Lissak, & Drasgow, 1982). By comparison, the lowest RMSEs in the current study were 0.37 and 0.59 for difficulty and ability, respectively. Also, differences between the sample size and model conditions in ability estimation error were minimal.

Both of these aspects of the results are likely due to the easiness of the test and the potentially large influence of guessing on performance. Also, we are currently exploring the extent to which the inclusion of extreme item parameter estimates in the ability estimation process may have compromised those results. Other potentially important factors, and implications for aphasia test use and development will be discussed.

References

Barnes, L. L. B. & Wise, S. L. (1991). The utility of a modified one-parameter IRT model with small samples. *Applied Measurement in Education, 4,* 143-157.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Fergadiotis, G., Wright, H., & Capilouto, G. (2010). Psychometric properties of the Pyramids and Palm Trees tests. *Procedia Social and Behavioral Sciences, 6,* 33-34.

Hula, W. D., Donovan, N. J., Kendall, D. L., & Gonzalez-Rothi, L. J. (2010). Item response theory analysis of the Western Aphasia Battery. *Aphasiology, 24,* 1326-1341.

Hula, W. D., Kalinyak-Fliszar, M., & Martin, N. (2009). Using item response theory to examine the effects of short-term memory demands on minimal pair discrimination. Presented to the annual meeting of the Academy of Aphasia, Boston, MA, October.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovey of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6,* 249-260.

Kendall, D., del Toro, C., Nadeau, S. E., Johnson, J., Rosenbek, J. C., & Velozo, C. A. (2010). The development of a standardized assessment of phonology in aphasia. Presented to the Clinical Aphasiology Conference, Isle of Palms, SC, May.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7,* 328.

Martin, N., Schwartz, M. F., & Kohen, F. P. (2006). Assessment of the ability to process semantic and phonological aspects of words in aphasia: A multi-measurement approach. *Aphasiology, 20,* 154-166.

Table 1. Descriptive statistics for the empirical sample of participants with aphasia. Age, MPO, and education are for the larger sample of n=111 previously reported by Martin et al. (2006) from which the current sample of n =70 was drawn. The 41 excluded subjects were not given the PPT.

	Mean	SD	Min	Max
Age	58.57	14.37	22	86
Months Post-Onset	33.51	41.52	1	195
Education	12.82	2.57	7	20
PPT % correct	89	7.41	60	100
Etiology %				
L Stroke	72			
L AVM or Aneurysm	8			
Bil. Stroke	4			
R Stroke	3			
Other	2			
Unavailable	10			

Table 2. Descriptive statistics for the empirical item and person parameters obtained from the empirical data, and for the parameters used to generate the simulated data.

	Mean	SD	Skewness	Kurtosis	Min	Max
Item Parameters						
Empirical						
Difficulty	-1.58	1.09	.555	.491	-3.39	.96
Discrimination	1.32	1.07	0.863	-0.205	0.20	2.87
Simulated						
Difficulty	-1.6	1.11	0.382	-0.237	-3.47	1.23
Discrimination	1.24	0.63	1.42	2	0.4	3.1
Pseudo-Guessing	0.5	0.03	-0.03	-0.75	0.44	0.56
Person Ability by Sample Size Condition						
N=70 (empirical)	0.04	0.98	-0.11	-0.242	-2.32	2.08
75	0.06	1.01	-0.07	-0.61	-2.31	2.2
200	0.04	1.03	-0.12	-0.68	-2.49	2.21
2000	0.09	1.01	-0.06	-0.2	-2.71	2.78

Table 3. Presentation and description of the IRT models used to analyze the simulated data. The symbol *p* refers to probability, x_j is the vector of responses (correct = 1, incorrect = 0) to the *j* items, θ is person ability, *a* is item discrimination, *b* is person ability, and *c* defines the minimum probability of a correct response.

Model	Equation	Description
1P	$p(x_j = 1 \mid \theta, b_j) = \frac{e^{(\theta - b_j)}}{1 + e^{(\theta - b_j)}}$	The only parameter estimated for each item is difficulty. Discrimination is held equal across items. The probability of correct guessing is assumed to be zero.
Mod1P	$p(x_j = 1 \mid \theta, b_j) = 0.45 + (1 - 0.45)\frac{e^{(\theta - b_j)}}{1 + e^{(\theta - b_j)}}$	Same as the 1P model, except the minimum probability of a correct response is set to 45%.
Mod2P	$p(x_j = 1 \mid \theta, a_j, b_j) = 0.45 + (1 - 0.45)\frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}}$	Same as the Mod1P model, except that discrimination is permitted to vary across items.
3P	$p(x_j = 1 \mid \theta, a_j, b_j, c_j) = c_j + (1 - c_j)\frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}}$	Same as the Mod2P model, except that the minimum probability of a correct response is permitted to vary across items.

Table 4. The proportion of observations (across 40 items x 1000 replications) obtaining non-extreme parameter estimates. Only these observations were included in the comparisons of the generating and estimated item parameters.

Sample Size	IRT Model			
	1P	Mod1P	Mod2P	3P
75	0.58	0.85	0.57	0.39
200	0.64	0.88	0.80	0.57
2000	0.61	0.87	0.96	0.93

Table 5. RMSE, bias, and correlations for IRT model difficulty estimates. For each metric and sample size, the mean value for the best-performing model is bolded.

	IRT model							
	1P		Mod1P		Mod2P		3P	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
RMSE								
75	1.74	0.21	**0.71**	0.11	0.83	0.17	1.33	0.25
200	1.64	0.13	**0.57**	0.07	0.65	0.12	1.37	0.20
2000	1.75	0.05	0.57	0.04	**0.37**	0.05	1.18	0.13
Bias								
75	-1.65	0.22	-0.30	0.15	**-0.16**	0.19	-0.17	0.43
200	-1.58	0.14	-0.23	0.09	-0.15	0.10	**-0.06**	0.26
2000	-1.70	0.05	-0.30	0.03	-0.22	0.04	**0.16**	0.16
Correlation	Median	IQR	Median	IQR	Median	IQR	Median	IQR
75	0.81	0.08	**0.85**	0.05	0.70	0.17	0.32	0.37
200	0.87	0.04	**0.89**	0.03	0.83	0.08	0.34	0.28
2000	0.89	0.02	0.90	0.02	**0.95**	0.03	0.50	0.18

Table 6. RMSE, bias, and correlations for IRT model discrimination estimates. For each metric and sample size, the mean value for the best-performing model is bolded.

	IRT model							
	1P		Mod1P		Mod2P		3P	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
RMSE								
75	**0.49**	0.07	0.52	0.06	0.80	0.20	0.89	0.25
200	0.52	0.05	**0.52**	0.05	0.59	0.10	0.84	0.18
2000	0.55	0.04	0.52	0.05	**0.23**	0.05	0.56	0.11
Bias								
75	-0.27	0.06	**-0.10**	0.10	0.25	0.15	0.41	0.25
200	-0.28	0.04	**-0.10**	0.06	0.15	0.13	0.37	0.16
2000	-0.31	0.02	-0.14	0.03	**0.02**	0.03	0.25	0.08
Correlation	Median	IQR	Median	IQR	Median	IQR	Median	IQR
75	n/a	n/a	n/a	n/a	**0.37**	0.29	0.32	0.38
200	n/a	n/a	n/a	n/a	**0.63**	0.15	0.39	0.26
2000	n/a	n/a	n/a	n/a	**0.89**	0.11	0.71	0.13

Table 7. RMSE, bias, and correlations for IRT model ability estimates. For each metric and sample size, the mean value for the best-performing model is bolded.

	IRT model							
	1P		Mod1P		Mod2P		3P	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
RMSE								
75	0.64	0.04	**0.62**	0.04	0.73	0.07	0.76	0.09
200	0.63	0.03	**0.61**	0.03	0.62	0.03	0.65	0.03
2000	0.63	0.01	0.61	0.01	**0.59**	0.01	0.60	0.01
Bias								
75	-0.15	0.00	-0.11	0.00	-0.11	0.06	**0.09**	0.18
200	-0.10	0.00	**-0.06**	0.00	-0.08	0.03	0.02	0.05
2000	-0.11	0.00	-0.08	0.00	-0.11	0.00	**-0.07**	0.02
Correlation	Median	IQR	Median	IQR	Median	IQR	Median	IQR
75	0.79	0.05	**0.79**	0.05	0.71	0.07	0.69	0.07
200	0.80	0.02	**0.81**	0.03	0.80	0.03	0.77	0.04
2000	0.79	0.01	0.80	0.01	**0.82**	0.01	0.81	0.01