INTRODUCTION

Impairment of naming ability is ubiquitous in aphasia and assessment of naming is central to clinical assessment (Nickels, 2002). One prominent naming test is the Philadelphia Naming Test (PNT) (Roach et al., 1996), which has favorable psychometric properties and has been used in many investigations of the theoretical nature of aphasic naming deficits (e.g., Dell et al., 1997; Schwartz et al, 2006). However, the PNT is a long test, limiting its usefulness in clinical settings. Recently, Walker & Schwartz (2012) published two 30-item PNT short forms (PNT30-A, PNT30-B) along with data supporting their reliability and validity. These short forms were developed using classical test theory methods, with attention to items' lexical characteristics, their overall difficulty, and error type distributions.

An alternative approach to shortening the PNT would employ item response theory (IRT) and computerized adaptive testing (CAT) methods. The major advantage of this approach is that it could provide better measurement precision than static short forms. The purpose of this study was to develop an IRT-based CAT version of the PNT and compare it to the static short forms developed by Walker and Schwartz (2012).

The simplest IRT model, the 1-parameter logistic (1-PL) model, predicts responses to test items as a function of two parameters: item difficulty and person ability (Baylor et al., 2011). These parameters are estimated as latent variables under the assumptions that (1) all of the items respond to a single common underlying factor, i.e., the test is unidimensional, (2) the items are all related to the underlying factor with equal strength, i.e., all items are equally discriminating, and (3) that all responses are independent, conditional on the underlying trait. IRT-based CAT proceeds by updating the examinee's ability estimate after each response, and administering items that are best matched to the current estimate.

We asked four questions:

1. Does the PNT demonstrate adequate fit to a 1-PL model?
2. How well does a 30-item computerized adaptive PNT (PNT-CAT) predict scores on the full PNT relative to the static short forms (PNT30-A, PNT30-B) developed by Walker and Schwartz ?
3. How does the measurement precision provided by the PNT-CAT compare to the PNT30-A & PNT30-B?
4. Does the PNT-CAT predict the proportions of naming error types with equal or better accuracy than the PNT30-A & PNT30-B?

METHOD, ANALYSES, & RESULTS

The data used in this study were taken from the Moss Aphasia Psycholinguistic Project Database (MAPPD) (Mirman et al., 2010). We analyzed item-level PNT data for 251 individuals with aphasia who comprised all cases with a complete first administration of the PNT available on

May 6, 2012. Descriptive data are provided in Table 1 for the full sample and the two sub-samples used in evaluating the PNT-CAT.

We began by fitting the dichotomized (correct/incorrect) data for all 251 cases to a unidimensional item-level factor model using NOHARM (Fraser & McDonald, 2003), and found that fit was good (see Table 2). We also evaluated the assumption of equal item discrimination by fitting a factor model with item loadings constrained to be equal. The constrained model showed significant misfit according to the likelihood ratio chi-square, though other fit indices were still within acceptable ranges.

Next, we fit the dichotomous data to a 1-PL IRT model using Multilog (Thissen, 2003) and evaluated item fit using the information-weighted mean-square statistic. All items obtained values < 1.4. To evaluate local independence, we used Yen's (1984) Q_3 statistic, which is based on residual inter-item correlations. The observed Q_3 mean and variance closely approximated the values expected under the assumption of local independence, and <5% of item pairs obtained residual correlations >2SD from the mean. Based on these results, we cautiously concluded that data-model fit was acceptable. The data showed significantly better fit to a 2-parameter logistic model, which relaxes the assumption of equal item discrimination, but the literature suggests that the current sample size is inadequate for estimating unique item discriminations (de Ayala, 2009). The 1-PL model also has theoretical and practical advantages arising from its simplicity (Hambleton & Swaminathan, 1985).

We then conducted real-data CAT simulations retaining all 175 PNT items in the item pool. We obtained 1-PL item parameter estimates using the dichotomous data from the first 200 cases in the database. We used these estimates and the responses of the remaining 51 cases to simulate administration of the PNT-CAT, PNT30A, and the PNT30-B, and compared the resulting score estimates to estimates obtained from the full 175-item test. By using non-overlapping patient samples to estimate the item parameters on the one hand and test the ability of the CAT to recover the full 175-item score on the other, we were able to conduct a robust test of the PNT-CAT. The procedure underlying CAT is illustrated in Figure 1. Also, in order to provide context for the simulation results, descriptive data for the scaled naming ability scores based on the full 175-item PNT are provided in Table 3 for the full sample and the two subsamples.

Results of the simulations (See Table 4) indicated that the PNT-CAT and the two PNT30 short forms had similarly high correlations with the full PNT score. The PNT-CAT showed less bias and a smaller root-mean-square difference with the full PNT than either of the static short forms. Examination of scatter plots, shown in Figure 2, of the PNT-CAT, PNT30-A, and PNT30-B scores over the full PNT score suggested a ceiling effect for the static short forms that was absent for the PNT-CAT.

1-PL model standard error curves for the PNT-CAT, PNT30-A, and PNT30-B, and the full PNT are shown in Figure 3. These plots demonstrate that the PNT-CAT provided better measurement precision across a wider range of naming ability than either static short form. The difference was most pronounced for ability levels >0 and <-1, encompassing 86% of the test sample. Reliability was correspondingly better for the PNT-CAT (0.95) than either PNT30-A (0.89) or PNT30-B (0.90).

For the test sample, we also calculated the number of five types of naming errors observed on each of the four versions of the PNT. We subjected these counts to an empirical logit transformation and calculated the correlation between the full PNT and the three short versions, displayed in Table 5. The PNT-CAT performed comparably to or better than the static short forms for all error types.

DISCUSSION

Using archival data, we demonstrated that the PNT-CAT provided ability estimates that were as accurate as two previously developed static short forms, with substantially better measurement precision for most respondents. Tests of model fit indicated that the PNT met the 1-PL model assumptions of unidimensionality and local independence, but not equal item discrimination. However, the good performance PNT-CAT suggests that the observed misfit did not have material consequences for the current purpose. We also found that the PNT-CAT performed as well as carefully constructed short forms in terms of estimating error type proportions, perhaps because the CAT algorithm produced on average a slightly larger number of errors than the static short forms. Future steps in the development of the PNT-CAT will include testing with prospective independent administrations of the full PNT and PNT-CAT, and further investigation of item and person fit, including tests of differential item functioning.

REFERENCES

Baylor, C., Hula, W.D., Donovan, N.J., Doyle, P.J., Kendall, D., Yorkston K., (2011). An introduction to Item Response Theory and Rasch models for speech-language pathologists. American Journal of Speech-Language Pathology, 20, 243-259.

de Ayala, R.J. (2009). The theory and practice of item response theory. New York: Guilford Press.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. Psychological Review, 104, 801–838.

Fraser, C. & McDonald, R.P. (2003). NOHARM: A windows program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory, Welland, ON: Niagara College.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1985). Fundamentals of item response theory. Newbury Park, CA: Sage.

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

Nickels, L. (2002). Therapy for naming disorders: Revisiting, revising,and reviewing. Aphasiology, 16, 935-979.

Roach, A., Schwartz, M. F., Martin, N., Grewal, R. S., & Brecher, A. (1996). The Philadelphia Naming Test: Scoring and rationale. Clinical Aphasiology, 24, 121–133.

Schwartz, M. F., Dell, G. S., Martin, N., Gahl, S., & Sobel, P. (2006). A case-series test of the interactive two-step model of lexical access: Evidence from picture naming. Journal of Memory and Language, 54, 228–264.

Thissen, D. (2003). Multilog 7.03: A computer program for multiple, categorical item analysis and test scoring using item response theory [Computer software]. Chicago, IL: Scientific Software.

Walker, G.M. & Schwartz, M.F. (2012). Short-form Philadelphia Naming Test: Rationale and empirical evaluation

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement, 8(2), 125–145.

Table 1. Demographic and Clinical Characteristics of the Patient Sample

	Total Sample (N=251)	**Calibration Sample** (N=200)	**Test Sample** (N=51)
Ethnicity, %			
African American	34%	38%	16%
Asian	0.4%	0.5%	0
Hispanic	1.2%	1.5%	0
Caucasian	44%	49%	24%
Missing	20%	10%	61%
Education, Years			
Mean	13.6	13.6	14.1
SD	2.8	2.7	3.2
Min	7	7	12
Max	21	21	21
Missing,%	20%	10%	61%
Age, Years			
Mean	58.8	58.7	60.4
SD	13.2	13.2	12.9
Min	22	22	34
Max	86	86	79
Missing,%	20%	10%	61%
Months Post-Onset			
Mean	32.9	30.6	53.9
SD	51.0	49.6	59.7
Min	1	1	1
Max	381	381	185
Missing,%	20%	10%	61%
Western Aphasia Battery AQ			
Mean	73.4	73.0	75.3
SD	16.6	16.4	17.9
Min	27.2	27.2	38
Max	97.8	97.8	96.3
Missing,%	51%	54%	39%
Philadelphia Naming Test, % Correct			
Mean	61%	61%	61%
SD	28%	27%	31%
Min	1%	1%	1%
Max	98%	98%	97%

Table 2. Fit statistics for the assessment of dimensionality using NOHARM (Fraser & McDonald, 2003).

Model	Root mean square of residuals	Tanaka Goodness of Fit Index	Approximate Chi-Square	df	p-value
1-Factor, unique item loadings	0.0101	0.9841	8182	15050	1
1- factor, item loadings constrained to be equal	0.0164	0.9578	22087	15224	<0.001
Criterion for acceptable fit	<0.25	>0.95			>0.05

Table 3. Descriptive statistics for scaled naming ability scores on the 175-item PNT, scaled according to the item parameter estimates for the calibration sample.

	Full Sample	Calibration Sample	Test Sample
	(N=251)	(N=200)	(N=51)
Mean	0.13	0.12	0.15
SD	1.17	1.14	1.32
Min	-3.04	-3.04	-3.04
Max	2.44	2.44	2.18
Avg. Standard Error	0.15	0.15	0.16
Reliability	0.98	0.98	0.99

Table 4. Comparison of the simulated Computerized Adaptive PNT (PNT-CAT) and two static short forms (PNT30-A, PNT30-B) with the full 175-item PNT.

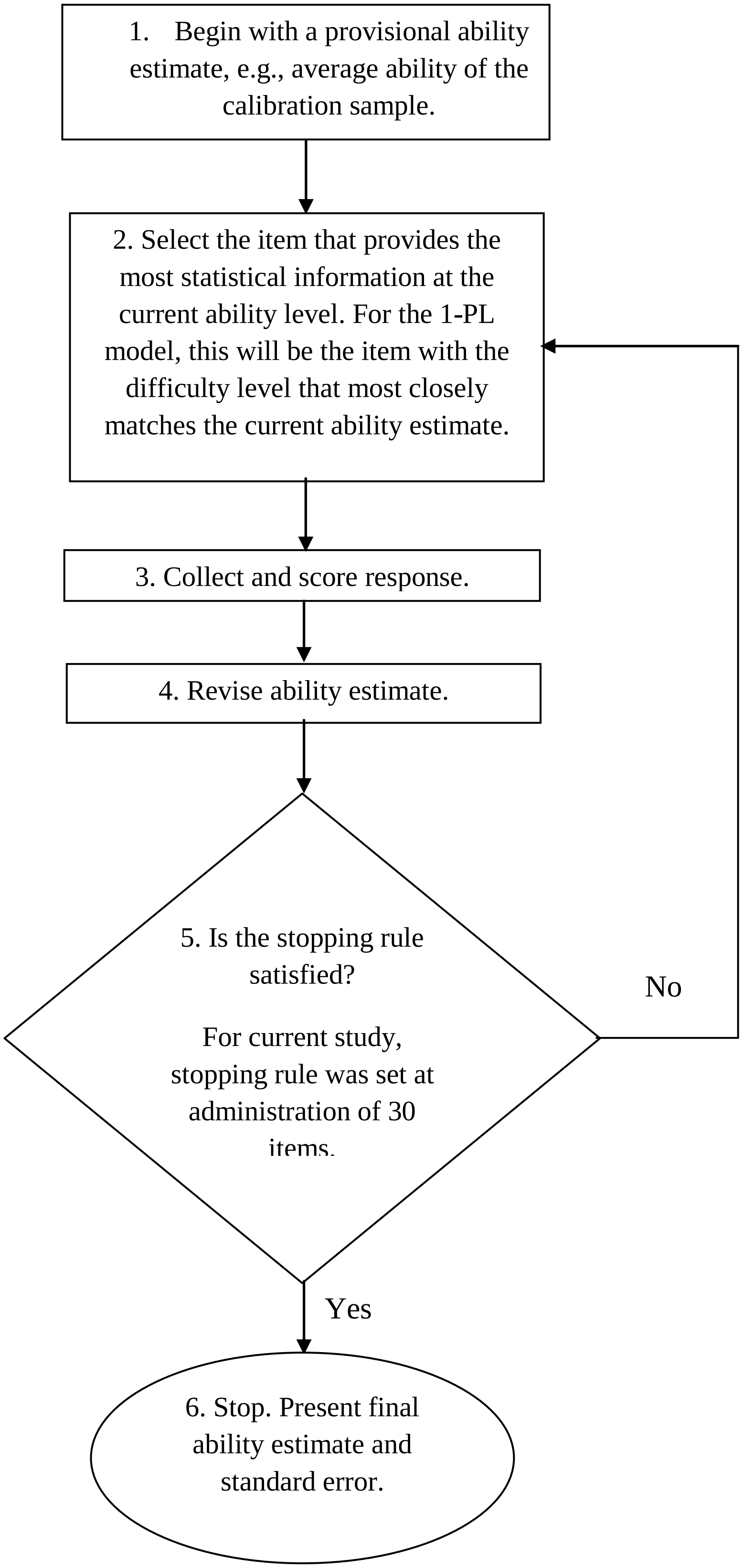
	PNT-CAT	PNT30-A	PNT30-B
Correlation	0.985	0.977	0.979
Bias	-0.008	0.089*	0.036
Root-mean-square difference	0.243	0.294	0.270

*Significantly different from 0, p<0.05

Table 5. Correlations for logit-transformed naming error proportions between the three simulated shortened versions of the PNT and the full 175-item PNT.

Error Type	PNT-CAT	PNT30-A	PNT30-B
Semantic	0.76	0.51	0.44
Formal	0.80	0.81	0.74
Mixed	0.51	0.54	0.19
Unrelated	0.83	0.88	0.88
Nonword	0.90	0.87	0.87

Figure 1. Schematic diagram of a computerized adaptive test.


1. Begin with a provisional ability estimate, e.g., average ability of the calibration sample.
2. Select the item that provides the most statistical information at the current ability level. For the 1-PL model, this will be the item with the difficulty level that most closely matches the current ability estimate.
3. Collect and score response.
4. Revise ability estimate.
5. Is the stopping rule satisfied?
For current study, stopping rule was set at administration of 30 items.
No
Yes
6. Stop. Present final ability estimate and standard error.

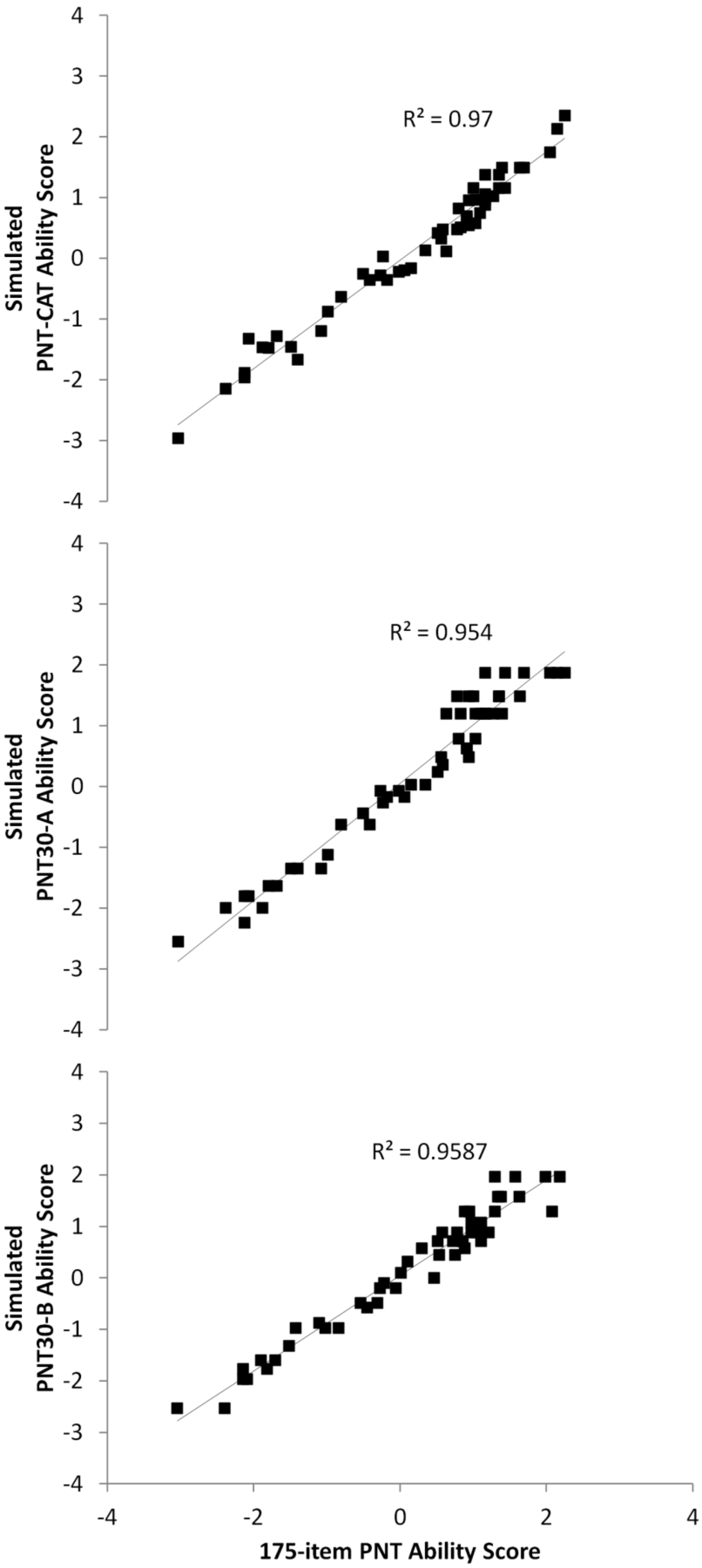


Figure 1. Scatterplots of simulated computerized adaptive PNT (PNT-CAT) and static short form (PNT30-A, PNT30-B) ability scores over scores estimated from the full 175-item PNT.

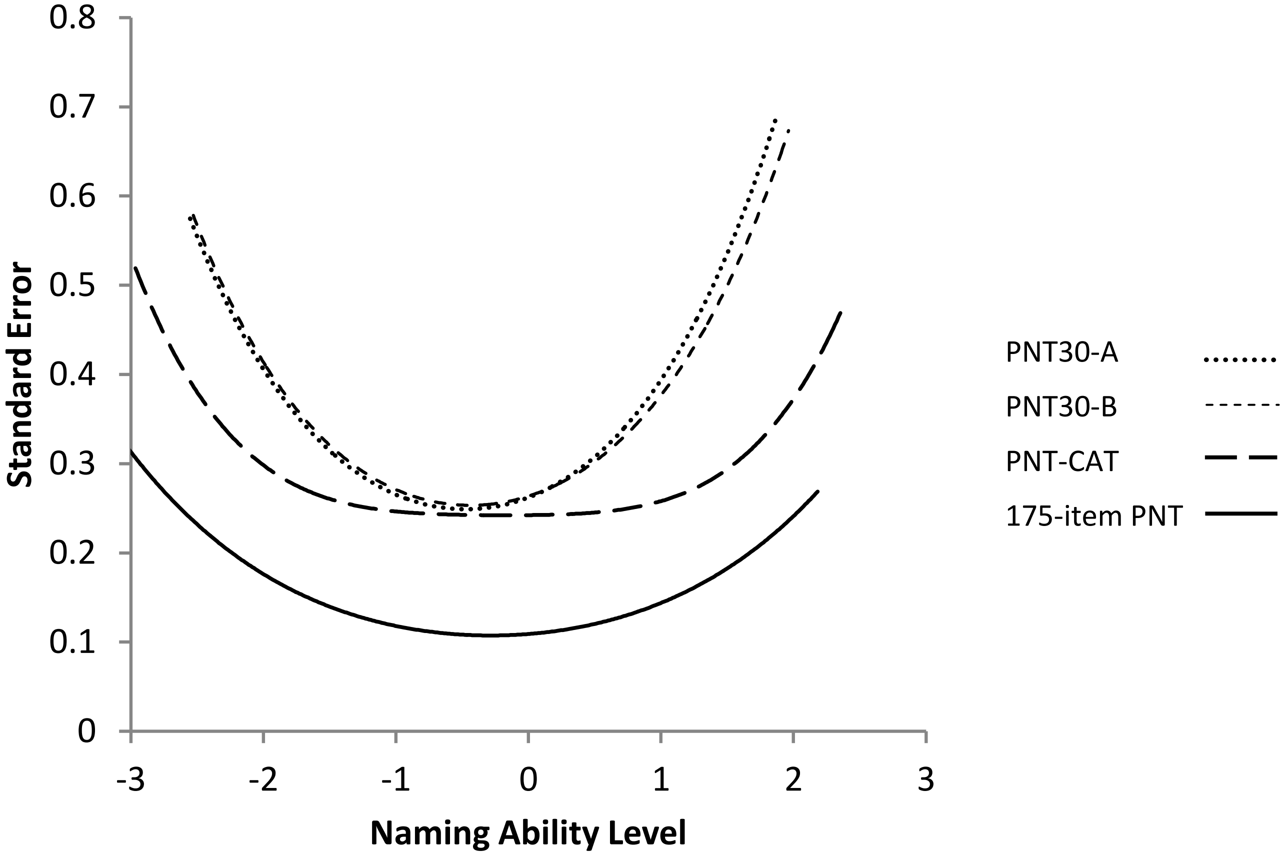


Figure 2. Plot of 1-PL model standard errors as a function of naming ability for the static PNT short forms (PNT30-A, PNT30-B), the computerized adaptive PNT (PNT-CAT), and the full 175-item PNT.