Development and Simulation Testing of a Computerized Adaptive Measure of Communicative Functioning in Aphasia

INTRODUCTION

Computerized adaptive testing (CAT), based on the mathematical framework of item response theory (IRT), has increasingly been implemented in patient reported outcome measures over the past decade (Fries, Bruce, & Cella, 2005). Given a calibrated item pool fit by an appropriate IRT measurement model, a CAT can produce reliable ability estimates more efficiently than traditional paper-and-pencil tests by administering items that are most informative given the examinee’s estimated ability level (Wainer, 2000). As conventional measures employed in the measurement of aphasia were developed under traditional measurement theory, many of these measures are long and inefficient, and are consequently unsuitable for regular clinical care. In addition, these conventional measures often fail to meet the needs of many community-dwelling stroke survivors whose impairments falls outside the range reliably measured by these tests (Doyle et al. 2012). IRT-based and in particular CAT patient reported outcome measures offer the possibility of substantial improvements in measurement technology for persons with aphasia.

Communicative functioning in aphasia may be usefully described by a general factor with contributions from additional specific factors reflecting unique aspects or skills related to particular sub-domains of communicative functioning (Doyle & Hula, 2012). One concern in the development of a multidimensional test is selecting items that best meet test content specifications. The most common item selection method associated with IRT, the maximum Fisher information method, selects items based on the value of their mathematical parameters alone. This method may result test content unrepresentative of full item bank (Leung, Chang, & Hau, 2003; Zheng et al., 2012). Strategies for increasing control of test content are called content-balancing strategies (Leung et al., 2003; Nering & Ostini, 2010). Research on content balancing suggests that content-balancing allows greater control of test content specification without impacting test efficiency (Leung et al., 2003).

Recently, Doyle and colleagues (2012; Doyle & Hula, 2012) reported on the Aphasia Communication Outcome Measure (ACOM), a patient-reported outcome measure developed using IRT-based methods. The ACOM demonstrated acceptable IRT model fit, good reliability, and concurrent validity suggesting that the ACOM item pool might be suitable for CAT administration. A computerized adaptive ACOM (CAT-ACOM) has the potential to not only decrease test length, but also significantly increase measurement precision for a wider range of individuals with aphasia.

In the current study, we aim to evaluate the performance of a CAT-ACOM, compared with a short form ACOM (SF-ACOM), with and without a content balancing strategy. Specifically, we predict that:

1) A CAT-ACOM will produce more accurate ability estimates than a SF-ACOM.
2) Estimates obtained from a content-balanced CAT (CAT-BAL) will be more accurate than those obtained from a standard non-content-balanced CAT (CAT-STD).

3) CAT-BAL will not significantly increase test length relative to the CAT-STD or SF-ACOM.

METHODS/RESULTS

Participants were 329 person with aphasia who met the following inclusion criteria: diagnosis of aphasia ≥1 MPO; community dwelling; self-reported normal pre-morbid speech-language function; pre-morbid literacy with English as a first language; negative self-reported history of progressive neurological disease, psychopathology, and substance abuse; ≥0.6 delayed/immediate ratio on ABCD Story Retell (Bayles & Tomoeda, 1993); ≤5 self-reported depressive symptoms on the GDRS-15 (Sheikh & Yesavage, 1986); and BDAE severity rating ≥1. Demographic and clinical characteristics of the sample are summarized in Tables 1 and 2.

The initial ACOM item pool was comprised of 177 items describing various communication activities. Participants were asked to rate on a 4-point scale how effectively they perform each activity. "Effectively" was defined as "accomplishing what you want to, without help, and without too much time or effort." Responses were collected with interviewer-assist by study staff experienced in the assessment of aphasia.

Item response data were examined with exploratory and confirmatory factor analyses. Based on these analyses, we reduced the item pool to a set of 35 items that demonstrated good fit to a bi-factor model, which proposes that the response to each item is determined by a common general factor plus one of four specific factors related to sub-domains of communicative functioning: Conversation (n = 21 items), Naming (n = 6), Writing (n = 4), and Comprehension (n = 4). Model fit information is provided in Table 3. The superior fit of the bi-factor model, combined with low percentage of variance accounted for by the specific factors, suggested that the 35 items were sufficiently related to one common, general factor to justify summarizing participants' responses with a single overall score.

The item parameter estimates (factor loadings and item thresholds) obtained from the bi-factor model were transformed to IRT graded response model parameters (item discriminations and thresholds) for the general factor (Wirth & Edwards, 2007). Then we conducted real-data simulations in which we compared administration of the full 35-item ACOM to four shortened versions: Content-balanced CAT (CAT-BAL), standard CAT (CAT-STD), content-balanced 10-item short form (SF-BAL), and standard 10-item short form (SF-STD). In the CAT-BAL condition, we implemented maximum Fisher information item selection along with a content-balancing method developed by Kingsbury and Zara (CCAT; 1989). In the CAT-STD condition, we simulated CAT with the maximum information item selection criterion alone. The CAT stopping rule was set at a maximum standard error of 0.3 (roughly equivalent to reliability of 0.90) or administration of 20 items. In the two short form conditions, we simulated administration of a content-balanced short form (SF-BAL) and a short form designed only to provide maximum statistical information for ability estimates between -3 and +3 (SF-STD).

Statistics for the distribution of ability estimates across all testing conditions were comparable (see Table 4 for a summary). We evaluated test performance across the four experimental conditions in four ways. First, we compared correlations between the four shortened versions and the full 35-item ACOM. In all four conditions, the correlations were similarly high (0.96-0.97)

Next, we calculated differences between the 35-item ACOM and the four shortened versions. We conducted two 2x2 repeated measures ANOVAs with test type (CAT, SF) and content balancing (balanced, unbalanced) as independent variables and the signed difference (bias) and the squared difference (error) as dependent variables. For bias, no effects were significant. For error, only content balancing had a significant effect (p=0.03). Content balancing was associated with small decreases in error. Descriptive data for the comparisons with the 35-item ACOM are presented in Table 5.

We also evaluated the number of items administered by the CAT-BAL and CAT-STD and compared them to the number of items administered by the short forms (n=10). Results are presented in Table 5.

DISCUSSION

The results of this simulation study failed to demonstrate advantages for CAT administration of the ACOM over short form administration. Content balancing was associated with more accurate score estimation, but the differences were small. These results suggest that CAT administration of the ACOM item bank may not offer practical benefits relative to short form administration. They also suggest that content balancing may slightly increase measurement accuracy with minimal sacrifice of test efficiency or reliability.

REFERENCES

Bayles, K. A. & Tomoeda, C. K. (1993). *Arizona Battery for Communication Disorders of Dementia*. Tucson, AZ: Canyonlands Publishing, Inc.

Doyle, P.J. & Hula, W.D. (November 17, 2012). The Aphasia Communication Outcome Measure. Presented to the Research Symposium at the American Speech-Language-Hearing Convention, Atlanta, GA.

Doyle, P.J., Hula, W.D., Austermann Hula, S.N., Stone, C.A., Wambaugh, J.L., Ross, K.B., Schumacher, J.G. (2012). Self- and Surrogate-Reported Communication Functioning in Aphasia, Quality of Life Research. doi: 10.1007/s11136-012-0224-5

Fries, J., Bruce, B., & Cella, D. (2005). The promise of PROMIS: using item response theory to improve assessment of patient-reported outcomes. *Clinical and experimental rheumatology, 23*(5), 53.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359-375.

Leung, C.-K., Chang, H.-H., & Hau, K.-T. (2003). Computerized adaptive testing: a comparison of three content balancing methods. *The Journal of Technology, Learning, and Assessment, 2*(5).

Nering, L. M., & Ostini, R. (Eds.). (2010). *Handbook of polytomous item response theory models*. New York, NY: Routledge.

Sheikh, J. I. & Yesavage, J. A. (1986). Geriatric Depression Scale (GDS) Recent Evidence and Development of a Shorter Version. In T.L.Brink (Ed.), *Clinical Gerontology: A Guide to Assessment and Intervention* (pp. 165-173). New York: Hawthorn Press.

Wainer, H. (2000). *Computerized adaptive testing: a primer* (2 ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. Psychological methods, 12(1), 58.

Zheng, Y., Chang, C.-H., & Chang, H.-H. (2012). Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement. *Quality of Life Research*, 1-9. doi: 10.1007/s11136-012-0179-6

Table 1. Demographic characteristics of the study sample, n = 329 persons with aphasia.

Age in Years, mean (sd)	60 (14)
Gender, % male	65.2%
Race	
Caucasian	84.6%
African American	6.9%
Hispanic	6.2%
Mixed	1.3%
Asian or Pacific Islander	0.7%
Aleutian, Eskimo, or Native American	0.3%
Education	
Primary/Middle School	6%
High School	26%
Some College	34%
College Graduate	23%
Post-Graduate Degree	12%
Marital Status	
Currently Married or Cohabitating	68%
Divorced or Separated	22%
Never Married	7%
Widowed	4%

Table 2. Clinical characteristics of the study sample.

Months Post-Onset of Aphasia, median (min-max)	33 (1-506)
Etiology of Aphasia	
Ischemic Stroke	71%
Hemorrhagic Stroke	19%
Stroke, undetermined type	9%
Other (TBI, tumor, radiation necrosis)	1%
PICA Overall score, median (min-max)	12.31 (7.24-14.82)
BDAE Severity Rating	
0	0%
1	23%
2	17%
3	23%
4	29%
5	7%
Missing	2%
Motor Speech Diagnosis	
Aphasia Only (no motor speech disorder)	51%
Apraxia of Speech	38%
Dysarthria	11%
Undetermined Motor Speech Disorder	1%

Table 3. Factor model fit results.

	Chi-square value	Degrees of freedom	p-value	Root Mean Square Error of Approximation (90% CI)	Comparative Fit Index	Chi-square Difference Test with Bi-factor model
Bi-factor	658.061	525	0.001	0.028 (0.02, 0.034)	0.992	na
Multiple-factors with correlated dimensions	787.687	554	<0.0001	0.036 (0.03, 0.041)	0.987	<0.0001
One-factor	1363.218	560	<0.0001	0.066 (0.062, 0.070)	0.954	<0.0001
Criterion for Acceptable Fit	na	na	> 0.05	< 0.05	>0.95	>0.05

Table 4. Descriptive statistics for ability estimates by test version.

	Full 35 Item ACOM	CAT-BAL	CAT-STD	SF-BAL	SF-STD
Mean	0.007	0.018	0.007	0.005	0.003
SD	1.05	1.07	1.06	1.09	1.09
Min	-3.24	-3.23	-3.23	-3.20	-3.20
Max	3.18	3.17	3.18	2.97	3.20
Avg. Standard Error	0.19	0.30	0.30	0.32	0.31
Reliability	0.96	0.92	0.92	0.91	0.92

Table 5. Comparison of content-balanced and standard (unbalanced) computerized adaptive test (CAT-BAL, CAT-STD) and short form (SF-BAL, SF-STD) versions with the full 35-item ACOM.

	CAT-BAL	CAT-STD	SF-BAL	SF-STD
Correlation	0.97	0.96	0.97	0.96
Bias	0.010	0.000	-0.003	-0.004
Root-mean-square error	0.267	0.300	0.275	0.291
Mean (SD) length in items	10.29 (2.96)*	9.35 (2.98)*	10(0)	10(0)

* significantly different from 10, per 1-sample t-test, p < 0.001.