The Western Aphasia Battery (WAB) (Kertesz, 1982) is used to classify aphasia by classical type, measure overall severity, and measure change over time. Despite its near-ubiquitousness, it has significant psychometric shortcomings. The purpose of this investigation was to improve the psychometric properties of the WAB, specifically its reliability and validity for measuring aphasia severity and change over time.

The WAB overall score, known as the Aphasia Quotient (AQ) is a weighted sum of five subtest scores rendered on a 1-100 scale. The Fluency and Information Content subtests are each measured by a single clinician rating of questionable reliability (Trupe, 1984). Because these two subtests comprise 40% of the WAB AQ, small changes in either rating can lead to substantial changes in the overall score. A second shortcoming of the WAB is one that it shares with other aphasia tests: the lack of intervality of the measurement scale. Interval scales are comprised of equally-sized units, and are necessary for validly describing differences between people or change over time. Although an ordinal scale (such as the WAB AQ) can validly indicate that one score is higher than another, it cannot validly describe the amount of difference.

Current aphasia tests are based, implicitly or explicitly in classical test theory (CTT) but test development has shifted in recent years to rely on item response theory (IRT) (Embretson and Reise, 2000). IRT offers a number of advantages over CTT. One advantage is that IRT-based scores have stronger claim to interval status than CTT scores. A second advantage is that IRT offers the potential to separate the individuals' test scores from the particular items used to derive those scores. Currently, an AQ can be calculated only if the entire WAB is administered, and an AQ is only interpretable in relation to other AQs. The AQ cannot be directly compared to scores from other aphasia tests, although they may correlate highly with one another and also measure aphasia severity. Separation of test scores from the individual items used to derive them makes possible adaptive test administration that could permit the same amount of information to be obtained with fewer items.

A third advantage concerns the reporting of reliability of individual test scores. A single standard error of measurement is typically used to construct confidence intervals about individual scores. This practice assumes that measurement error is distributed normally and equally for all levels of aphasia severity. By contrast, standard errors derived from IRT models depend on the number of items administered and the degree to which the difficulty of the items administered matches the ability of the individuals tested. Thus, tests that are too easy or too hard for a given individual yield scores that are less reliable (i.e., with larger SEs).

These and other advantages of IRT models come at the cost of strong assumptions about the data being modeled. One key assumption concerns the dimensionality of the item set. The most commonly and easily applied models assume that performance on test items is related to only one underlying dimension or latent trait. The purpose of the current investigation was to determine whether the WAB can be productively fit to an IRT model, which could improve its reliability and validity for measuring aphasia severity, and lead to other potential benefits described above.

METHOD

This study used an archival data set collected at an aphasia research laboratory whose standard protocol includes the WAB. Inclusion criteria for the current analyses

were left hemisphere stroke ≥6 months prior to examination and AQ<94.7. Naming, repetition, and command items were coded as 2 for fully correct, 1 for partially correct, and 0 for fully incorrect. Otherwise, responses were coded according to standard WAB scoring.

ANALYSIS AND RESULTS

Exploratory factor analysis (EFA) was conducted using Mplus 1.04. The first factor extracted accounted for 66% of the variance and the ratio of first-factor to second-factor eigenvalues was 11.9, suggesting that the WAB measures a single unidimensional construct.

IRT analyses were conducted using Winsteps 3.67. In general, IRT models describe patient responses as a function of the difference between person ability and item difficulty. Because the WAB contains a mix of dichotomous and rating scale items, a Rasch partial credit model was estimated. This model estimates a single difficulty value for dichotomous items and a series of difficulty thresholds for items with 3 or more response categories, such as the fluency rating scale, naming, command, and repetition items. One Yes-No item, "Is your name (real name)?" had no incorrect responses and could not be estimated. Standard criteria were used to evaluate the functioning of the rating scales (Linacre, 2004). The Information Content scale was collapsed into five categories to achieve adequate psychometric properties, and the Fluency scale was collapsed into three. The Command, Naming, and single-word Repetition items were recoded as dichotomies because in each case the number of partially correct responses observed was insufficient to make the middle category viable. The Semantic Fluency item was collapsed into 3 categories.

Once adequate rating scale functioning was achieved, information-weighted mean-square fit statistics were used to evaluate model fit. Eleven items had fit values ≥1.4, indicating that they elicited a large number of unexpected responses, i.e, correct responses from lower-ability patients or incorrect responses from higher-ability patients. These items are listed in Table 1.

After exclusion of misfitting items, the final model was estimated with the person scores scaled such that the lowest and highest possible scores are 0 and 100, respectively, to facilitate comparison with the AQ. Descriptive statistics for item difficulty, person ability, and WAB AQ are presented in Table 2. A scatterplot of Rasch person ability scores and WAB AQs is presented in Figure 1. Histograms for these variables are presented in Figures 2 and 3. In Figure 4, a plot of the ability-conditioned standard errors provided by the model is presented. Finally, in Figure 5, a scatterplot of selected cases is presented with 95% confidence intervals about both scores.

DISCUSSION

The WAB demonstrates reasonable fit to a Rasch model, suggesting that item difficulty and person ability adequately predict patient responses. This is an important finding because Rasch-based scores provide more valid indices of severity and change, and Rasch models can support adaptive testing.

Despite overall good model fit, we found several items that misfit, suggesting that they are measuring some construct other than aphasia severity (assuming that the WAB in fact measures aphasia severity) and making them candidates for exclusion from the

test. For example, the misfitting Sentence Completion items are common fixed expressions, suggesting that such items are not necessarily valid indicators of aphasia severity.

We also found that measurement precision varied widely across the ability range. Comparison of Figures 3 and 4 reveals that the top half of the distribution is measured with considerably less precision than the bottom half. This implies that, when using the WAB to measure change, patients with milder aphasia must show greater improvement than those with more severe aphasia to achieve reliable score differences. The current practice of using a single standard error value across the entire ability range does not take this into account. The implications of IRT for aphasia testing will be discussed.

References

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

Kertesz, A. (1982). *Western Aphasia Battery*. New York: Grune & Stratton.

Linacre, J. M. (2004). Optimizing rating scale category effectiveness. In E.V.Jr.Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 258-278). Chicago: JAM Press.

Shewan, C. M. & Kertesz, A. (1980). Reliability and validity characteristics of the Western Aphasia Battery (WAB). *J Speech Hear.Disord., 45,* 308-324.

Trupe, E. H. (1984). Reliability of rating spontaneous speech in the Western Aphasia Battery: Implications for classification. In Minneapolis, MN: BRK.

Table 1. Items demonstrating misfit to the Rasch model

Subtest and Item Number	Item Content
Yes-No 9	Am I a man/woman? (y)
Yes-No 10	Are the lights on in this room? (y)
Yes-No 15	Will paper burn in fire? (y)
Yes-No 17	Do you eat a banana before you peel it? (n)
Yes-No 19	Is a horse larger than a dog? (y)
Yes-No 20	Do you cut the grass with an ax? (n)
Auditory Word Recognition 16	Arrow
Auditory Word Recognition 21	B
Auditory Word Recognition 30	5000
Repetition 3	Pipe
Sentence Completion 3	Roses are red, violets are _____?
Sentence Completion 4	They fought like cats and _____?

Table 2. Desciptive statistics for Rasch-based WAB item and person measures, and WAB AQ.

	Mean	SD
Item Difficulty	46.90	11.9
Person Ability	63.61	14.5
WAB AQ	65.64	25.0

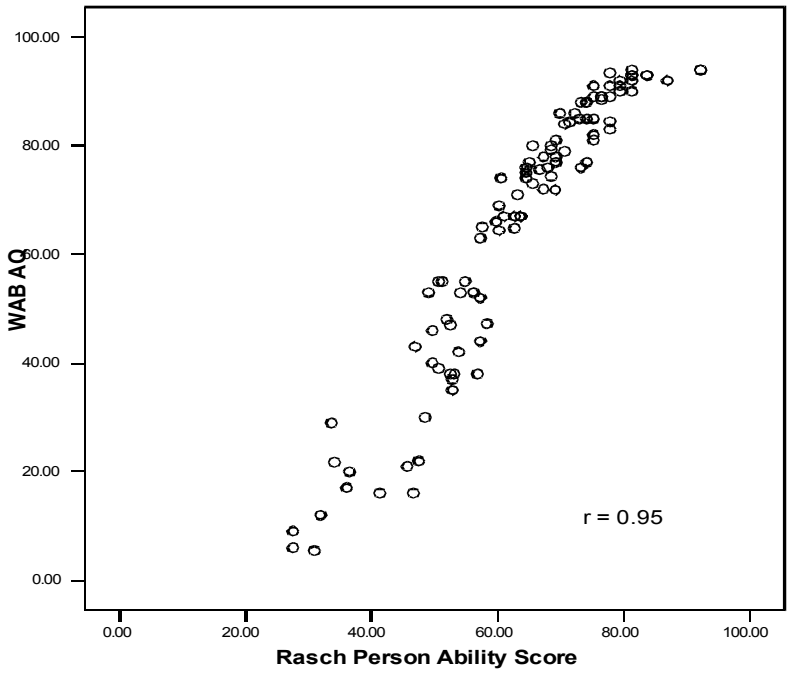

Figure 1. Scatterplot of WAB AQs over Rasch-based WAB scores.

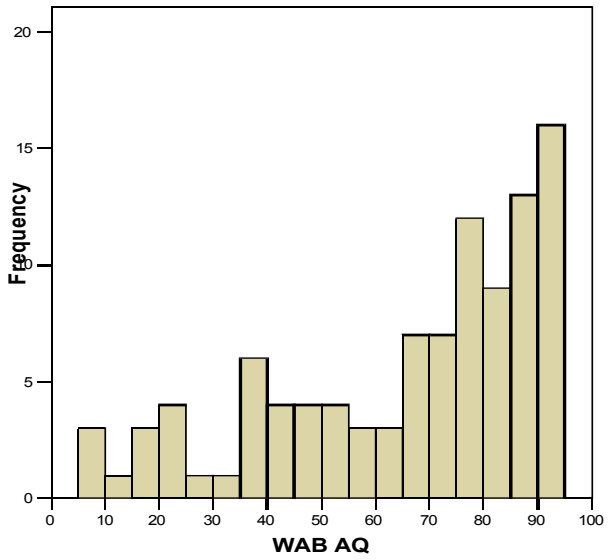

Figure 2. Histogram of WAB AQs.

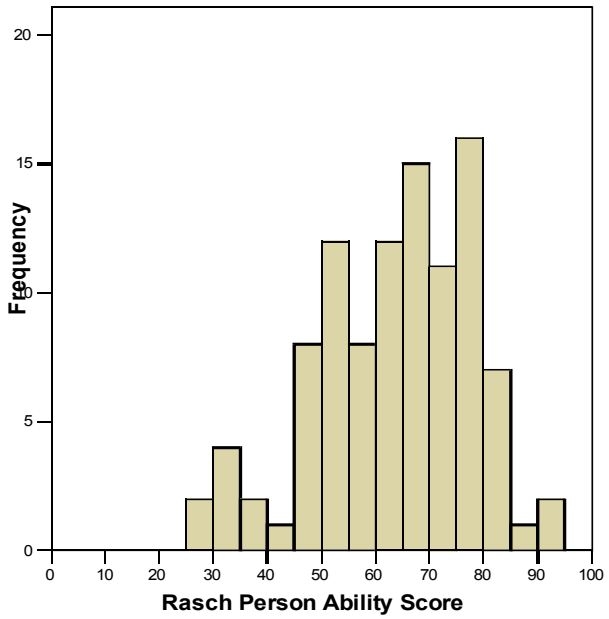

Figure 3. Histogram of Rasch-based WAB scores.

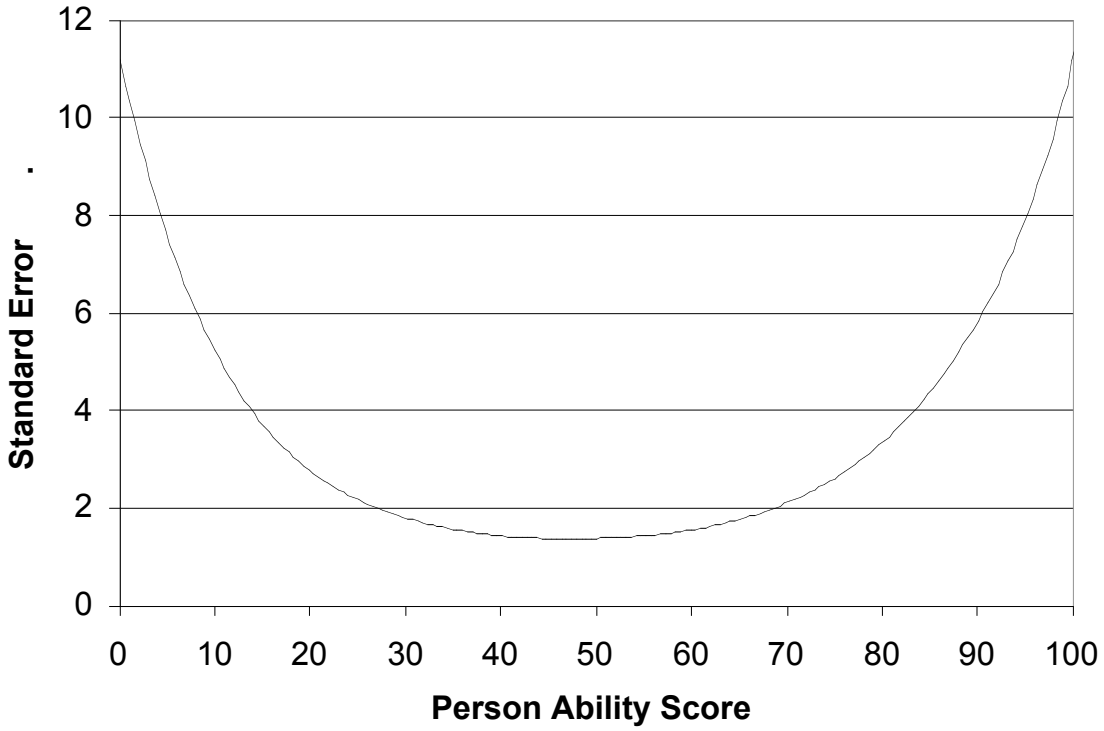

Figure 4. Plot of Rasch-modeled standard error by ability score. Person scores closer to the center of the item distribution (~ 47) are measured more reliably, with lower SEs, while scores toward either end of the scale are measured less reliably, with higher SEs. This is because there are fewer items with difficulty values toward either end of the scale, and when item difficulty and person ability are far apart, each response provides less statistical information.