Published in final edited form as:

Int J Speech Lang Pathol. 2018 April ; 20(2): 274–283. doi:10.1080/17549507.2016.1276962.

Validity, Reliability, and Sensitivity of the NORLA-6: Naming and Oral Reading for Language in Aphasia 6-Point Scale

Laura L. Pitts, Ph.D., CCC-SLP^{1,2,3}, Rosalind Hurwitz, M.S., CCC-SLP¹, Jaime Lee, M.S., CCC-SLP¹, Julia Carpenter, M.A., CCC-SLP¹, and Leora R. Cherney, Ph.D., CCC-SLP, BC-ANCDs^{1,2}

¹Rehabilitation Institute of Chicago

²Northwestern University

³University of Northern Iowa

Abstract

Purpose—Evaluation of the Naming and Oral Reading for Language in Aphasia 6-Point scale (NORLA-6), a scoring system of oral reading and naming performance in aphasia.

Method—Data were drawn from 91 participants with non-fluent aphasia secondary to left-hemisphere stroke across four treatment studies. To assess validity, Spearman's correlations were calculated between the NORLA-6 and the Gray Oral Reading Test-Fourth Edition (GORT-4) Accuracy score, GORT-4 Rate score, and the Boston Naming Test (BNT). Inter-rater and test-retest reliability were evaluated using correlations. Sensitivity to change following oral reading intervention was analyzed using Wilcoxon Signed Rank tests between pre- and post-treatment NORLA-6 scores.

Result—NORLA-6 performance was significantly correlated ($p < .001$) with all reference tests (GORT-4 Accuracy, $r_s = .84$; GORT-4 Rate, $r_s = .61$; and BNT, $r_s = .92$). Inter-rater (ICC = .90) and test-retest ($r > .92$) reliability were both excellent. Sensitivity following oral reading intervention was demonstrated in both oral reading accuracy and rate ($p < .004$).

Conclusion—The NORLA-6 is a valid and reliable measure of oral reading and naming performance. It also demonstrates sensitivity to change in treatment-targeted behaviours. Therefore, the NORLA-6 scale may enhance outcome measurement in both clinical practice and aphasia research.

Keywords

Aphasia; outcomes; oral reading; assessment; validity; reliability

Introduction

Quantifying the language abilities of individuals with aphasia is increasingly vital for speech language pathologists (SLPs) providing clinical services. It is also fundamental to the research process to measure efficacy and facilitate meta-analyses of outcomes. The World Report on Disability (World Health Organisation, 2011) calls for the development of research methodologies, which could be robust and permit international comparison of data.

Unfortunately, there is a lack of consistent use of and uniformity in aphasia assessment across and within health care systems (Katz et al., 2000; Simmons-Mackie, Threats, & Kagan, 2005; Wallace, Worrall, Rose, & Dorze, 2014). One multi-national study across five English-speaking countries reported that SLPs applied 679 different aphasia assessment tools, with 62% of professionals relying heavily on informal language assessment or screening measures (Katz et al., 2000). SLPs in Australia reportedly employ approximately 180 different aphasia assessments with an estimated 77% of professionals relying heavily on screening tools (Verna, Davidson, & Rose, 2009). In the USA, only 40% of health care-based SLPs report including outcome measures in their documentation despite the increasingly critical role measurements have in determining medical coverage and reimbursement (American Speech-Language-Hearing Association, 2013; Centers for Medicare and Medicaid Services, 2014). The lack of resources (including materials and time) has been specifically cited as a major barrier to effective service provision (Rose, Ferguson, Power, Togher, & Worrall, 2014). This overwhelming lack of consistency in aphasia assessment may indicate ongoing challenges with existing measurement tools such as availability, feasibility, and repeatability.

There are a number of useful standardised assessment tools to describe and quantify the language symptoms associated with aphasia (e.g. Western Aphasia Battery (WAB; Kertesz, 1982), Boston Diagnostic Aphasia Examination (BDAE; Goodglass, Kaplan & Barresi, 2001), Boston Naming Test (BNT; Kaplan, Goodglass, & Weintraub, 1983), and the Porch Index of Communicative Ability (PICA; Porch, 2001)). However, none of the many extant aphasia assessments are universally considered to be the “gold standard”. Reviews of the literature have also found little consistency in the use of standardised assessment instruments across randomised controlled trials of neurogenic communication disorders (Simmons-Mackie, Threats, & Kagan, 2005; Xiong, Bunning, Horton, & Hartley, 2011).

Additionally, few standardised tools are appropriate and feasible for frequent, repeated measurement, as is needed to monitor progress throughout treatment. Comprehensive aphasia assessment measures such as the WAB and the BDAE can take up to several hours to administer, since they aim to evaluate various communication modalities (Kertesz, 2007; Goodglass, Kaplan & Barresi, 2001). Given the time constraints within clinical practice settings, such comprehensive assessments cannot be administered within a typical therapy session.

Moreover, comprehensive and standardised assessments cannot be given frequently due to learning effects. Standardised assessments are susceptible to practice effects if used repeatedly within a short time frame, particularly in the absence of alternate forms (Duff,

Westervelt, McCaffrey, & Haase, 2001). Of the standardised tests referenced in the present article, only the Gray Oral Reading Test-Fourth Edition (GORT-4; Wiederholt & Bryant, 2001), discussed below, uses a Form A and Form B. Thus, there is need for supplemental outcome measures that can frequently quantify and chart language recovery throughout treatment.

Ways of quantifying aphasia and its recovery include binary systems (i.e. correct versus incorrect scoring) and assigning numerical values through the use of rating scales and multidimensional scoring systems. Binary systems often capture only one dimension (e.g. accuracy), and are therefore limited in the degree of information they provide. In contrast, multidimensional systems can account for several features simultaneously, such as responsiveness, completeness and immediacy. For example, the well-established multidimensional scoring system utilised in the PICA (Porch, 2001) is a 16-point scale with the following five dimensions: accuracy, responsiveness, completeness, promptness and efficiency. Other examples include a similar 16-point scale from the Revised Token Test (McNeil & Prescott, 1978), and a 15-point scale employed by Wambaugh and Wright (2007) to measure word-retrieval abilities.

These multidimensional metrics provide reliable methods for describing language performance; however, debate continues on the practicality of using extensive scales that have such wide ranges of scores in clinical settings. Some researchers have suggested that binary categories are strict and neglect valuable information regarding specific language behaviours and deficits underlying those behaviours (McNeil & Prescott, 1978; Porch, 1971, 2007). Others contend that extensive scales may be too cumbersome, require considerable training, may not be efficient and/or cost-effective, and may be most feasible when applied only to a limited number of test items (Odekar & Hallowell, 2005, 2007; Park, McNeil, & Tompkins, 2000). Nevertheless, there is wide agreement that further research is needed to evaluate our tools and develop efficacious scoring methods that demonstrate our quality of care (Odekar & Hallowell, 2005, 2007; Porch, 2007).

Thus, researchers and clinicians have continued to design idiosyncratic scales for their own specific tasks and purposes. One application of this type of scale is to score regularly repeated (e.g. weekly) language probes for items targeted in treatment as well as similar untrained targets. This, in turn, provides a mechanism for capturing potential changes resulting from clinical or experimental interventions. Probe measures can easily be obtained at the beginning of a therapy session due to their succinctness, and have the advantage of specificity through the use of treatment-related targets. However, in order to be valuable for clinical documentation or research purposes, all measurement tools must prove to be valid, reliable, and sensitive to change (Rogers, 2013).

The Naming and Oral Reading for Language in Aphasia 6-Point scale (NORLA-6) is a multidimensional scoring system developed to quantify naming and oral reading performance during repeated testing in a series of aphasia intervention studies. The scale was designed to capture multiple dimensions of language performance including responsiveness, completeness, accuracy and immediacy. It was based on clinically recognised characteristics of oral production errors that are typical of aphasia. Each word

produced during either oral reading of sentences or confrontation naming is scored from zero to five, where zero = no response and five = an accurate and immediate response. Individual word scores can then be summed within a sentence or across naming stimuli to give an overall score of oral reading accuracy, naming accuracy, and when combined with temporal measures of oral sentence reading, a metric of oral reading rate.

The NORLA-6 demonstrates promise of clinical utility. In contrast to other multidimensional scoring systems, the brevity of a six-point scale simplifies clinician training and permits greater feasibility and efficiency by quantitatively probing language performance within the time constraints of a single session. The purpose of the current study was to examine the validity, inter-rater and test-retest reliability, and sensitivity of the NORLA-6 scale.

Method

The NORLA-6 Scale

The Naming and Oral Reading for Language in Aphasia 6-Point scale (NORLA-6) was designed to quantify oral sentence reading and naming performance on probe measures during aphasia intervention studies that utilised a treatment technique called Oral Reading for Language in Aphasia (ORLA[®]; Cherney, 2010a, 2010b; Cherney, Merbitz, & Grip, 1986). Each word produced by the participant during either oral reading of sentences or confrontation naming tasks is assigned a score of zero through five, as illustrated in Appendix A. Participants' productions of sentences during oral reading are also timed to provide a measure of oral reading rate.

Reference Measures

Three reference measures were selected for the validation of NORLA-6 scores. Two of the GORT-4 converted scores (Accuracy and Rate) were chosen to validate the NORLA-6 measures of Oral Reading Accuracy and Oral Reading Rate, respectively. The GORT-4 was selected as a reference measure since it is the only available commercial test of oral reading fluency and comprehension. Although the GORT-4 is normed for healthy persons aged 6 to 19, it has been applied to adults with aphasia in several treatment research studies (e.g. Beeson & Insalaco, 1998; Orjada & Beeson, 2005). The Boston Naming Test (BNT) was chosen to validate Naming Percent Accuracy scores as it is a widely used, standardised confrontational naming test for aphasia.

Participants

Data from 91 participants in four aphasia treatment studies from 2005 to 2012 were evaluated. Each study provided six weeks of computerised aphasia treatment involving reading aloud sentences presented by a virtual therapist (ORLA[®]). All four studies were approved by the Northwestern University Institutional Review Board.

Inclusion criteria across studies included single unilateral stroke resulting in chronic aphasia of greater than six months post-onset; at least 12th grade education; premorbidly proficient in English; premorbidly right handed; and sufficient visual and auditory acuity to engage in

computerised treatment tasks. Exclusion criteria included any other neurological condition that could potentially affect language or cognition; any significant psychiatric history prior to the stroke; and active substance abuse.

Participants varied widely in age (24 – 76 years old), chronicity (6 – 160 months post-CVA onset) and years of formal education (12 – 21 years). The wide range in age, chronicity, and years of formal education is reflective of clinical populations with chronic aphasia treated in post-acute rehabilitation. Aphasia severity as measured by the WAB Aphasia Quotient (WAB AQ) ranged from 14 – 91 out of a possible 100. All participants were classified as nonfluent according to the WAB AQ subtest scores.

Standardised assessments and probe measures of oral reading and naming using the NORLA-6 scale were conducted at pre- and post-treatment. All participants completed the WAB. However, each study had its own separate study procedures, and not all 91 participants were administered the same battery of standardised aphasia assessments or oral reading and naming probes. Twelve participants had missing data for both oral reading and naming probes and were therefore excluded. The remaining sample ($n = 79$) was divided into three subsets: (1) a subset of 68 participants who completed both Oral Reading Probes and the GORT-4 at baseline assessment, (2) a subset of 61 participants who completed Oral Reading Probes and the GORT-4 at baseline and following completion of the computerised language intervention, and (3) a subset of 42 participants who completed Naming Probes and the BNT at baseline and following completion of the computerised language intervention. Demographic data (mean and standard deviations) for the total sample of participants and participants in the three subsets are provided in Table I. Table I also shows which probes and tests were completed by each subset of participants and Figure 1 provides a Venn diagram of the number of participants within each of the three subsets.

Across the four studies, the 68 participants in subset 1 were included in the assessment of validity and reliability of the NORLA-6 as a measure of oral reading performance. The 61 participants in subset 2 were included in the assessment of sensitivity of the NORLA-6 for oral reading following language treatment. The 42 participants in subset 3 permitted the exploration of the validity, reliability and sensitivity of the NORLA-6 as a measure of naming performance; these participants had completed 6 weeks of computerised aphasia treatment as well as the NORLA-6 naming probes and the BNT at both baseline and post-treatment. In our analyses of validity, NORLA-6 scores were obtained from an initial baseline evaluation. Additionally, a second baseline probe was taken within one week of the initial probe to assess test-retest reliability for both Oral Reading ($n = 68$) and Naming ($n = 42$). Finally, NORLA-6 probes obtained at the initial baseline evaluation and immediately following the end of treatment were used to analyse sensitivity to change following intervention.

Oral Reading Probes

Baseline and post-treatment probe stimuli were comprised of ten sentences that, for each participant and for each test session, were randomly selected from the same pool of untrained sentences. These untrained sentences were matched on length and grade level to the sentences used in treatment. For those with more severe aphasia (WAB AQ < 55), both

the 50 probe and 150 treatment stimuli were 3–5 words long at a Flesch-Kincaid first grade reading level. For those with less severe aphasia (WAB AQ > 55), both the 50 probe and 150 treatment stimuli were 8–12 words long at a Flesch-Kincaid third grade reading level. Flesch-Kincaid Reading Grade Level (FKRGL) is a widely used and well-recognised metric of readability particularly in the USA, but also abroad and across disciplines (Ley & Florio, 1996; Thomas, Hartley, & Kincaid, 1975). Probe stimuli utilised American English spelling and were presented in print to the participant. Examples of the trained and untrained probe stimuli are included in Appendix B. Participants were instructed to read the sentences aloud and were not provided any cues or support from the evaluating therapist. All productions were audio-recorded. A trained examiner reviewed the recorded productions and scored each word using the NORLA-6 ratings of zero through five. The scores for individual words were summed as a total accuracy score, divided by the total possible score (i.e. the total number of target words in all ten sentences multiplied by five), then multiplied by 100 to calculate the *Oral Reading Percent Accuracy* score.

Timing the sentences enabled calculation of correct words per minute. Using the audio recordings, each sentence was independently timed from the end of the stimulus presentation to the participant's last utterance. Results were summed across the ten sentences to obtain a total time. *Oral Reading Rate* (correct words per minute) was then calculated by counting the number of words across the ten sentences that received a NORLA-6 score of three or greater and dividing that number by the total time.

Naming Probes

Ten stimulus items were randomly selected from a library of 60 black-and-white line drawings of objects that were of relatively high frequency and available commercially for communication board use. The drawings were placed on individual 3×5 note cards. Examples of the stimulus items are included in Appendix B. Participants were asked to name the items as they were presented one at a time. Each response was audio-recorded and assigned a NORLA-6 rating by a trained examiner. The ten scores were summed, divided by 50 (i.e. the total possible score), then multiplied by 100 to calculate a *Naming Percent Accuracy* score.

Raters

A research speech language pathologist with ten years of experience had scored the oral reading and naming probes when the treatment studies were originally conducted. Two additional research speech language pathologists, with five and eight years of experience respectively, were trained as raters for inter-rater reliability in the present investigation. Their training consisted of an hour-long briefing on the NORLA-6 rules followed by independent practice scoring 100 previously recorded oral reading sentences. Practice results were compared with the scores of the original rater and feedback was given. In instances where a participant's idiosyncratic patterns led to difficulty in determining a score on the 6-point scale, decisions were made by consensus among the raters to ensure consistency in scoring for that particular subject. Raters passed training when their scores agreed with those of the original rater at a level of 90% or above. Agreement was reached within two to three hours of practice.

Data Analyses

Analyses were performed using SPSS 21.0 with alpha set at 0.05. The assumption of normality in range of performance was not met for the GORT-4, BNT, nor for the NORLA-6 scores by visual analysis of P-P plots and histograms, kurtosis z-scores less than 1.96, or Shapiro-Wilk tests and therefore non-parametric analyses were conducted. Group analyses were collapsed across age, chronicity, gender and years of premorbid education as there were no significant correlations with or significant group differences in (i.e. Mann-Whitney *U* and Kruskal-Wallis tests) oral reading and naming performance ($p > .05$). Collapsing the groups across these demographic variables allowed the major analyses to be conducted on larger samples.

Validity

For subset 1, with 68 participants completing baseline measures of oral reading performance, Spearman's rank-order correlations were conducted between both Oral Reading Accuracy, and Oral Reading Rate and the WAB AQ to assess the relation between probe performance and overall aphasia severity. To assess content validity for measurement of oral reading, Spearman's rank-order correlations were calculated between both Oral Reading Accuracy and Oral Reading Rate and the converted GORT-4 Accuracy and Rate scores. Additionally, first-order partial correlations were conducted to explore the validity of the NORLA-6 scale when controlling for participant age, years of education, and months post stroke onset between 1. Oral Reading Percent Accuracy and GORT-4 Accuracy, and 2. Oral Reading Rate and GORT-4 Rate.

For subset 3, with 42 participants completing baseline measures of naming performance, Spearman's rank-order correlations were calculated between Naming Percent Accuracy scores and both WAB AQ and BNT raw scores. First order partial correlations were also conducted to explore the validity of Naming Percent Accuracy and BNT scores when controlling for participant age, years of education, and months post stroke onset.

Reliability

Inter-rater reliability for Oral Reading Accuracy and Rate was prospectively calculated from a randomly selected 25% of the NORLA-6 oral reading probe sentences, 289 randomly selected words from the oral reading probes for point-to-point inter-rater reliability, and from a randomly selected 25% of the NORLA-6 naming probes that were scored by a second rater. Inter-rater reliability was analysed using Interclass Correlation Coefficients (ICCs) with a two-way random effects, single measure model and absolute agreement. ICC values > 0.75 indicate satisfactory reliability (Portney & Watkins, 2000).

Probes were repeated during a second evaluation occurring within one week of the first evaluation. Test-retest reliability was calculated as Pearson's correlation coefficient across the total sample of NORLA-6 probes for Oral Reading Accuracy and Oral Reading Rate ($n = 68$) and Naming Accuracy ($n = 42$).

Sensitivity

Finally, we assessed the sensitivity of the NORLA-6 scale to detect change in oral reading and naming performance following 6-weeks of oral reading treatment. The mean change and standard deviations of the NORLA-6 scale were calculated between the initial baseline evaluation and the post-treatment evaluation for the measures of Oral Reading Accuracy, Oral Reading Rate, and Naming Accuracy. Wilcoxon Signed Rank tests were also conducted for Oral Reading Accuracy, Oral Reading Rate, and Naming Accuracy between the initial baseline evaluation and the post-treatment evaluation following 6-weeks of computerised oral reading treatment.

Result

Validity

The WAB AQ scores were significantly and positively correlated with NORLA-6 performance for measures of: Oral Reading Percent Accuracy ($r_s = 0.863$, $p < 0.001$) and Oral Reading Rate ($r_s = 0.834$, $p < 0.001$). There was a significant, positive correlation between the Oral Reading Accuracy on the NORLA-6 and GORT-4 Accuracy scores, $r_s = 0.837$, $p < 0.001$ (Figure 2) and between the Oral Reading Rate on the NORLA-6 and GORT-4 Rate score, $r_s = 0.605$, $p < 0.001$ (Figure 3). When controlling for age, education and months post stroke onset, first-order partial correlations between 1. Oral Reading Percent Accuracy and GORT-4 Accuracy, and 2. Oral Reading Rate and GORT-4 Rate all remained significant at the $p < 0.001$ level with strong positive relation ($r = 0.71$).

The WAB AQ scores were significantly and positively correlated with NORLA-6 performance of Naming Percent Accuracy ($r_s = 0.895$, $p < 0.001$). Additionally, there was a significant, positive correlation between the Naming Percent Accuracy on the NORLA-6 and BNT raw score, $r_s = 0.918$, $p < 0.001$ (Figure 4). When controlling for age, education and months post stroke onset, first-order partial correlations between Naming Percent Accuracy and BNT scores all remained significant at the $p < 0.001$ level with strong positive correlations ($r = 0.84$).

Inter-rater and Test-retest Reliability

A high degree of reliability was found between trained raters using the NORLA-6 to rate Oral Reading Accuracy and Oral Reading Rate across 25% of the sample (170 sentences). The single measure ICC for Oral Reading Accuracy and Oral Reading Rate were 0.993 ($p < 0.001$) and 0.942 ($p < 0.001$) with a 95% confidence interval from 0.988 to 0.995 and 0.922 to 0.957, respectively. Point-to-point inter-rater reliability for a sample of 289 single words produced a single measure ICC of 0.945 ($p < 0.001$) with a 95% confidence interval from 0.923 to 0.960. Inter-rater reliability in Naming Percent Accuracy across 25% of the sample (105 words) produced single measure ICC of 0.957 ($p < 0.001$) with a 95% confidence interval from 0.937 to 0.970.

A high degree of test-retest reliability was found in the NORLA-6 ratings of Oral Reading Accuracy and Oral Reading Rate between two baseline testing sessions with Pearson

correlation coefficients (r) of 0.95 and 0.93, respectively. Naming Percent Accuracy also demonstrated a high degree of test-retest reliability with $r = 0.92$.

Sensitivity

Oral Reading Performance—Mean differences in performance between pre-treatment and post-treatment evaluations were 8.1% (SD = 10.52) for Oral Reading Accuracy and 4.44 (SD = 12.65) correct wpm for Oral Reading Rate. Related-samples Wilcoxon signed rank tests indicated a significant change from pre- ($M = 44.1$; $Mdn = 34.78$) to post-treatment ($M = 52.2$; $Mdn = 53.86$) in Oral Reading Accuracy ($T = 1629$, $z = 5.256$, $p < 0.001$, $r = 0.48$). Oral Reading Rate also improved significantly following treatment (pre-treatment $M = 15.0$; $Mdn = 7.03$; post-treatment $M = 19.4$; $Mdn = 10.53$; $T = 1204$, $z = 2.999$, $p = 0.003$, $r = 0.27$).

Naming Performance—The mean difference in Naming Percent Accuracy between pre-treatment and post-treatment evaluations was 3.48% (SD = 12.19). A significant difference in naming performance existed between pre- ($M = 53.6$; $Mdn = 55$) and post-treatment ($M = 57.1$; $Mdn = 63$; $T = 448$, $z = 2.182$, $p = 0.029$, $r = 0.24$).

Discussion

Establishing objective measures of performance in aphasia is critical for documenting and monitoring language performance over time and there is a need for repeatable and feasible outcome measures. The present study provides evidence that the NORLA-6 scale is a valid outcome measure for both oral reading and naming performance through strong correlations with established, standardised tests (i.e. GORT-4 and BNT) even when controlling for the factors of age, education, and chronicity. Additionally, the NORLA-6 scale demonstrates a strong positive correlation with overall aphasia severity (i.e. WAB AQ). Interestingly, while the NORLA-6, BNT, and GORT-4 did not demonstrate a ceiling effect for naming performance, floor effects for the BNT and GORT-4 exceeded 16% and 61% of the sample, respectively. Therefore, since our sample included participants with a mean WAB score of 50.1 (standard deviation of 19.4), it is possible that persons with more severe aphasia may exacerbate such floor effects when evaluated by the BNT or GORT-4 alone. It remains to be determined if the lack of floor effect by the NORLA-6 is due to its multidimensionality and therefore an ability to capture components of naming performance beyond dichotomous accuracy scores and/or the characteristics of the stimuli. Nevertheless, the NORLA-6 maintains integrity in quantifying overall oral reading and naming performance relative to well-established standardised measures while still exhibiting practicality.

The NORLA-6 scale was applied across a diverse clinical population of right-handed persons with nonfluent aphasia as the sample included a range of severities, chronicity, and participant ages. Furthermore, the scale was established as reliable both between raters and during test-re-test conditions. Therefore, Oral Reading Percent Accuracy, Oral Reading Rate and Naming Accuracy as measured by the NORLA-6 scale were both stable at baseline and able to demonstrate clinical gains following aphasia treatment. Overall, the results support the NORLA-6 as a valid, reliable, and stable measure that is also sensitive to change in

targeted oral reading measures following aphasia intervention within a diverse population of persons with nonfluent aphasia.

Findings suggest the NORLA-6 scoring system demonstrates clinical utility. It is quick to use within the clinical session, and can be applied repeatedly and frequently to measure the impact of treatment on a constrained set of stimuli. It can also be routinely used without practice effects to measure generalisation when applied to matched untrained stimuli. Repeated NORLA-6 scores may be compared across treatment sessions, both within and across the continuum of care, when obtained from the same stimuli or stimuli equated for difficulty. The scoring system also provides the clinician with information on the frequency of various error types (e.g. frequency counts of each scaled score zero through five), which may be valuable for classification and treatment planning.

Overall, the NORLA-6 scale is feasible for clinicians to establish baseline performance in oral reading and naming, monitor changes resulting from intervention, document improvement, and justify reimbursement for services. Currently our data support the use of the NORLA-6 with an impairment-based treatment. However, the potential availability, feasibility, and repeatability of the NORLA-6 may promote consistency not only to measure the effects of impairment-based therapies such as computerised oral reading treatment, but with further investigation, have applications as an outcome measure for other treatments that target language use during participation and activities of daily life.

In contrast to other multidimensional scoring systems that have received criticism for their complexity and burden of training (Odekar & Hallowell, 2005, 2007), the NORLA-6 is relatively easy to learn and apply. Our study has shown that speech language pathologists can become proficient with it within two–three hours. Appendix C includes scoring examples and a decision tree which were developed to facilitate clinicians' use of the scale. The ease and success of training are evident in the strong inter-rater reliability, which is comparable to other established multidimensional metrics (Park et al., 2000). The rules are clear and descriptive, ensuring accurate application of the scale, without prior training, to the range of productions typical of people with aphasia.

Limitations and future directions

Despite its potential utility in both clinical and research settings, this preliminary evaluation of the NORLA-6 scale has several limitations. First, its psychometric properties apply only to the sentence and picture stimuli used in the studies described herein; although it has the potential to provide valid and reliable measurement of responses to other stimuli including single word oral reading, this requires further evaluation. Secondly, the system is a measure of verbal responsiveness; therefore, it does not capture the often meaningful nonverbal communication utilised by many people with aphasia. Thirdly, we limited our sample to a homogeneous group of individuals with chronic nonfluent aphasia resulting from stroke because of the potential for those with a fluent aphasia to produce different kinds of error responses during oral reading and naming. We have applied the NORLA-6 scale to individuals with fluent aphasia in prior research and it appears feasible and useful (e.g. Cherney, Kaye, & Van Vuuren, 2014); however, validation remains to be established for people with fluent aphasia and other etiologies of aphasia, such as acquired brain injury and

other neurological disorders. Likewise, validation of the scale is currently limited to oral reading and naming tasks as described, but could theoretically be expanded to encompass other language modalities, such as written expression. Future research regarding probe stimuli selection and psycholinguistic variables such as word regularity and frequency may also be of particular interest. Furthermore, a limitation of this and other multi-dimensional scales of aphasia is the lack of empirical establishment of ordinality of the clinical ratings, and although averaging ordinal scores is widely practised to produce interpretable results, statistical confirmation is warranted.

The present study validates the NORLA-6 scale with the GORT-4 measures of accuracy and rate. Although the GORT-4 was developed to quantify oral reading and comprehension skills of school aged children, it has been applied to persons with aphasia in research contexts (Beeson & Insalaco, 1998; Orjada & Beeson, 2005). While the GORT-4 has adequate psychometric properties, there are no psychometrics on the aphasia population. Therefore, a limitation of the present study is the lack of standardised assessments of oral reading in persons with aphasia for which to establish validity.

Conclusion

The NORLA-6 is a metric to quantify and characterise both oral reading and naming performance in aphasia. It may serve as a useful tool to address current barriers to outcome reporting in clinical and research settings.

Acknowledgments

Supported by Grants H133G040269, H133G060055 and H133G070074 (to L.R.C.) from the National Institute on Disability and Rehabilitation Research, Department of Education, and by Grants 5R21DC9876 and 1R01DC011754 (to L.R.C.) from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health.

Appendix A: Naming and Oral Reading for Language in Aphasia 6-Point Scale (NORLA-6)

5	Accurate and immediate response
4	<p>Accurate but delayed response (A delay is defined as: a pause, with or without fillers, preceding the target word that is atypical of the participant's oral reading and naming performance. Some participants may show consistent, delayed responses in the context of slow and effortful oral reading and naming. In such cases, all words are rated a 4.)</p> <p>Self-corrected response</p> <p>Minor distortion of a recognizable phoneme</p> <p>Prosodic variation (e.g., stress on incorrect syllable)</p> <p>Correct word produced in the wrong order of the sentence</p>
3	<p>Omission, addition, or substitution of a grammatical morpheme</p> <p>Omission, addition, or substitution of a single phoneme (regardless of the length of the word)</p> <p>Incorrect use of verb tense for irregular verbs</p>

2	Related verbal paraphasia (includes in-class substitutions of function words such as determiners, prepositions, pronouns) Verbal paraphasia with a minor phonemic error Phonemic paraphasia in which more than half of the word is correct
1	Unrelated verbal paraphasia as a substitution for the target word Unrelated response as a substitution for the target word Unintelligible response as a substitution for the target word Phonemic paraphasia in which half or less of the word is correct
0	No verbal response
<p>Additional rules:</p> <p>Contractions</p> <p>If two words are contracted, score as 5 and 3. e.g. “it is” → “it’s”: score 5 for “it” and 3 for “is”</p> <p>If a contraction is expanded, score it as a 3. e.g. “it’s” → “it is” receives a single score of 3</p> <p>Fillers</p> <p>If the filler is a substitution for the target word, score as 1. e.g. effortful “um/uh” when attempting the target or “Oh boy” when attempting the target</p> <p>If the production is commentary that is not intended for a target word, it does not receive a score. e.g. “Okay” produced before attempting the sentence</p> <p>If the filler precedes an accurate production of the target word, score the target word as 4.</p> <p>Multiple attempts</p> <p>If multiple attempts are made, score the best production.</p> <p>Two-phoneme words</p> <p>If there is one phonemic error in a two-phoneme word, score as 3. e.g. “ma” for “me”</p>	

Notes:

¹ Dialectal differences and dysarthric distortions are not penalized.

² Do not score commentary before and after the attempt.

³ If two or more rules apply, apply the lowest score.

Appendix B: Examples of Probe Stimuli

1. Oral Reading Probe Stimuli

Level 1 Sample Probes. 3–5 word sentences

Flesch-Kinkaid grade level for individual sentences ranged from 0.5 to 2.8 with an overall Flesch-Kinkaid Grade Level of about 1.5

He is at the doctor.

Give me a minute.

I’d like some orange juice.

Please sit over there.

His new car was expensive.

What happened last night?

Call for more information.

I'm all finished.

I'd love some chicken soup.

It's raining hard.

Level 2 Sample Probes: 8–12 word sentences

Flesch-Kincaid grade level for individual sentences ranged from 2.4 to 4.8 with an overall Flesch-Kincaid Grade Level of about 3.5.

In a few years, we will be able to afford a house.

The one hour long documentary took months to film.

She bought a new dress on sale for twenty dollars.

It is a beautiful evening to take a walk.

We should arrange our summer vacation; where should we go?

We need to buy milk, cheese and eggs at the grocery store.

He must practice every day if he wants to improve his game.

I can meet you at the restaurant at about noon.

These are the most comfortable shoes that I have.

My winter coat is hanging in the closet.

2. Naming Probe Stimuli

Each set of probes comprised ten pictures randomly selected from a library of 60 pictures. Below, in alphabetical order, are some examples of the items in the 60-picture library.

Apple Beer Bell Bicycle Book Boots Broom Chair Clown Corn Desk Dress Envelope Ghost Hammer Horse House Iron King Knife Leaf Lion Scarf Scissors Train

Appendix C: Scoring Examples: NORLA-6 and Decision Tree

Target Sentence: <i>Chicago is the best city.</i>						
Production	Scoring					Rationale for Scoring
Chicago is the [3 second pause] best city.	Chicago	is	the	best	city	<i>best</i> : delayed but accurate response
	5	5	5	4	5	
Chi Chicago is the best city	Chicago	is	the	best	city	<i>Chicago</i> : self-corrected response

Target Sentence: <i>Chicago is the best city.</i>						
Production	Scoring					Rationale for Scoring
	4	5	5	5	5	
dChipago is the best city	Chicago	is	the	best	city	<i>Chicago</i> : Substitution of a single phoneme
	3	5	5	5	5	
Chicago is ba deft city	Chicago	is	the	best	city	<i>the</i> : substitution of a single phoneme <i>best</i> : half of the word is incorrect
	5	5	3	1	5	
Chicago is my favorite city	Chicago	is	the	best	city	<i>the</i> : unrelated response <i>best</i> : verbal paraphasia
	5	5	1	2	5	
Okay Chi ah (pointing to the word <i>the</i>) berst er ah ah cit oh I don't know	Chicago	is	the	best	city	<i>Chicago</i> : less than half of the word is produced correctly <i>is</i> : no response <i>the</i> : "ah" intended as a target word ("the") <i>best</i> : addition of a single phoneme <i>city</i> : omission of a single phoneme
	1	0	1	3	3	
Chicago is the best the best city	Chicago	is	the	best	city	<i>city</i> : delayed by repetition of previous target
	5	5	5	5	4	
Well it's a good one it's a good it's Chicago is a great city it's good it's a good one a best one	Chicago	is	the	best	city	<i>Chicago, is, best, city</i> : produces targets on best productions <i>the</i> : related verbal paraphasia (within class substitution of function word)
	4	4	2	4	4	

Target Word: <i>Spoon</i>		
Production	Scoring	Rationale
fork no uh spoon	4	self-corrected response
Shpoon	3	substitution of a single phoneme
Fork	2	verbal paraphasia
Flork	2	verbal paraphasia with minor error
ah ah (gestures for eating soup)	1	production intended as the target word
I don't know	0	Commentary on performance
gesture with no verbal attempt	0	no verbal response

NORLA-6 Decision Tree

References

- American Speech-Language-Hearing Association. Outcome Measures Report. SLP Health Care Survey 2013. 2013a. Retrieved from <http://www.asha.org/uploadedFiles/2013-SLP-Health-Care-Survey-Outcome-Measures.pdf>
- American Speech-Language-Hearing Association. Medicare Fee Schedule for Speech-Language Pathologists. 2013b. Retrieved from <http://www.asha.org/uploadedFiles/2013-Medicare-Fee-Schedule-SLP.pdf>
- Beeson PM, Insalaco D. Acquired alexia: Lessons from successful treatment. *Journal of the International Neuropsychological Society*. 1998; 4:621–635. [PubMed: 10050367]
- Centers for Medicare and Medicaid Services. Covered Medical and Other Health Services, Section 220.3 - Documentation Requirements for Therapy Services. Medicare Benefit Policy Manual. 2014. Retrieved from <http://www.cms.gov/Regulations-and-Guidance/Guidance/Manuals/downloads/bp102c15.pdf>
- Cherney LR. Oral Reading for Language in Aphasia (ORLA): Impact of Aphasia Severity on Cross-Modal Outcomes in Chronic Nonfluent Aphasia. *Seminars in Speech-Language Pathology*. 2010a; 31:42–51.
- Cherney LR. Oral Reading for Language in Aphasia (ORLA): Evaluating the Efficacy of Computer-Delivered Therapy in Chronic Nonfluent Aphasia. *Topics in Stroke Rehabilitation*. 2010b; 17(6): 423–431. [PubMed: 21239366]
- Cherney LR, Kaye RC, Van Vuuren S. Acquisition and maintenance of scripts in aphasia: A comparison of two cuing conditions. *American Journal of Speech Language Pathology*. 2014; 23:S343–S360. [PubMed: 24686911]
- Cherney LR, Merbitz CT, Grip JC. Efficacy of oral reading in aphasia treatment outcome. *Rehabilitation Literature*. 1986; 47:112–118. [PubMed: 3749589]
- Duff K, Westervelt HJ, McCaffrey RJ, Haase RF. Practice effects, test-retest stability, and dual baseline assessments with the California Verbal Learning Test in an HIV sample. *Archives of Clinical Neuropsychology*. 2001; 16:461–476. [PubMed: 14590160]
- Goodglass, H., Kaplan, E., Barresi, B. The assessment of aphasia and related disorders. Philadelphia: Lippincott Williams & Wilkins; 2001.
- Kaplan, E., Goodglass, H., Weintraub, S. Boston Naming Test. Philadelphia: Lee & Febiger; 1983.
- Katz R, Hallowell B, Code C, Armstrong E, Roberts P, Pound C, Katz L. A multinational comparison of aphasia management practices. *International Journal of Language and Communication Disorders*. 2000; 35(2):303–314. [PubMed: 10912257]
- Kertesz, A. Western Aphasia Battery. New York: Grune & Stratton; 1982.
- Kertesz, A. Western Aphasia Battery Revised Examiner's Manual. San Antonio, TX: Harcourt Assessment, Inc.; 2007.
- Ley P, Florio T. The use of readability formulas in health care. *Psychology, Health & Medicine*. 1996; 1(1):7–28.
- McNeil, MM., Prescott, TE. Revised Token Test. Austin, TX: Pro-Ed, Inc; 1978.
- Odekar A, Hallowell B. Comparison of Alternatives to Multidimensional Scoring in the Assessment of Language Comprehension in Aphasia. *American Journal of Speech-Language Pathology*. 2005; 14:337–345. [PubMed: 16396616]
- Odekar A, Hallowell B. Response to Porch. *American Journal of Speech-Language Pathology*. 2007; 16:87.
- Orjada S, Beeson P. Concurrent treatment for reading and spelling in aphasia. *Aphasiology*. 2005; 19:341–351.
- Park GH, McNeil MR, Tompkins CA. Reliability of the Five-Item Revised Token Test for individuals with aphasia. *Aphasiology*. 2000; 14:527–535.
- Porch BE. Multidimensional scoring in aphasia testing. *Journal of Speech, Language, and Hearing Research*. 1971; 14:776–792.
- Porch, BE. Porch Index of Communicative Ability – Revised (PICA-R). Albuquerque, NM: Pica Programs; 2001.

- Porch BE. Comments on “Comparison of alternatives to multidimensional scoring in the assessment of language comprehension in aphasia” by Odekar and Hallowell (2005). *American Journal of Speech-Language Pathology*. 2007; 16:84–86. [PubMed: 17329679]
- Portney, L., Watkins, M. *Foundation of Clinical Research: Application to Clinical Practice*. 2nd. Upper Saddle River, NJ: Prentice Hall Health; 2000.
- Rogers, M. Mapping the emerging landscape in health care. 2013. Retrieved from http://www.asha.org/Academic/questions/Mapping-the-Emerging-Landscape/?utm_source=asha&utm_medium=newsletter&utm_campaign=accessar040413
- Rose M, Ferguson A, Power E, Togher L, Worrall L. Aphasia rehabilitation in Australia: Current practices, challenges and future directions. *International Journal of Speech-Language Pathology*. 2014; 16(2):169–180. [PubMed: 23777446]
- Simmons-Mackie N, Threats T, Kagan A. Outcome assessment in aphasia: A survey. *Journal of Communication Disorders*. 2005; 38:1–27. [PubMed: 15475012]
- Thomas G, Hartley RD, Kincaid JP. Test-retest and inter-analyst reliability of the Automated Readability Index, Flesch Reading Ease Score, and the Fog Count. *Journal of Reading Behavior*. 1975; 7(2):149–154.
- Verna A, Davidson B, Rose R. Speech-language pathology services for people with aphasia: A survey of current practice in Australia. *International Journal of Speech-Language Pathology*. 2009; 11(3): 191–205.
- Wallace S, Worrall L, Rose T, Dorze G. Measuring outcomes in aphasia research: A review of current practice and an agenda for standardization. *Aphasiology*. 2014; 28:1364–1384.
- Wambaugh JL, Wright S. Improved effects of word-retrieval treatments subsequent to addition of the orthographic form. *Aphasiology*. 2007; 21:632–642.
- Wiederholt, J., Bryant, B. *Gray Oral Reading Tests*. 4th. Austin, TX: Pro-Ed; 2001.
- World Health Organisation. The world report on disability. 2011. Retrieved from http://www.who.int/disabilities/world_report/2011/report.pdf
- Xiong T, Bunning K, Horton S, Hartley S. Assessing and comparing the outcome measures for the rehabilitation of adults with communication disorders in randomized controlled trials: An International Classification of Functioning, Disability and Health approach. *Disability and Rehabilitation*. 2011; 33:2272–2290. [PubMed: 21473703]

Figure 1.

Venn diagram of the distribution of the total sample of 79 participants across 3 subsets. Subset 1 analysed the validity and reliability of oral reading probes. Subset 2 analysed the sensitivity of oral reading probes. Subset 3 analysed the validity, reliability and sensitivity of the naming probes. 31 participants contributed data to all three subsets. 30 participants contributed data to two subsets (i.e. 19 participants contributed data to subsets 1 and 2, while 11 participants contributed data to subsets 2 and 3). 18 participants contributed data to one subset, and that was subset 1.

Figure 2.

Scatterplot between NORLA-6 Oral Reading Percent Accuracy and GORT-4 Accuracy at baseline ($r_s=0.837$). Ceiling effects for the NORLA-6 and GORT-4 Accuracy scores were not found. NORLA-6 scores ranged from 0%–93% accuracy with 1 person scoring 0%; however, floor effects for the GORT-4 exceed 61% of the sample (i.e. 42 of 68 participants scoring 0).

Figure 3.
Scatterplot between NORLA-6 Oral Reading Rate and GORT-4 Rate at baseline ($r_s=0.605$).

Figure 4.

Scatterplot between NORLA-6 Naming Percent Accuracy and BNT score at baseline ($r_s=0.918$). Ceiling effects for the NORLA-6 and BNT Accuracy scores were not found. NORLA-6 scores ranged from 8%–98% accuracy; however, floor effects for the BNT exceed 16% of the sample (i.e. 7 of 42 participants scoring 0).

Table 1

Participant Demographics across Total Sample and Subset

	Demographics					Items Completed					
	N	Age	Sex	WAB AQ	MPO	Education (years)	Oral Reading Probes Baseline	Oral Reading Probes Post-Tx	Naming Probes	GORT	BNT
Total Sample	91	54.8 ±11.9	54M 37F	54.9 ±19.7	38.9 ±36.7	13.1 ±5.8					
Oral Reading: Validity & Reliability	68	54.6 ±12.8	40M 28F	55.7 ±19.6	39.1 ±35.1	14.3 ±4.7	X			X	
Oral Reading: Sensitivity	61	55.3 ±11.2	34M 27F	52.7 ±19.6	36.5 ±36.1	14.8 ±3.0	X	X		X	
Naming	42	54.5 ±10.6	23M 19F	50.1 ±19.4	32.2 ±33.5	14.8 ±2.0			X		X

Note. WAB AQ = Western Aphasia Battery Aphasia Quotient (Kertesz, 1982); MPO = months post onset; GORT = Gray Oral Reading Tests-4th edition (Wiederholt & Bryant, 2001); BNT = Boston Naming Test (Kaplan, Goodglass, & Weintraub, 1983); M = Male, F = Female.