Disability and Rehabilitation

ISSN: 0963-8288 (Print) 1464-5165 (Online) Journal homepage: https://www.tandfonline.com/loi/idre20

Inter-rater reliability, intra-rater reliability and internal consistency of the Brisbane Evidence-Based Language Test

Alexia Rohde, Molly McCracken, Linda Worrall, Anna Farrell, Robyn O'Halloran, Erin Godecke, Michael David & Suhail A. Doi

To cite this article: Alexia Rohde, Molly McCracken, Linda Worrall, Anna Farrell, Robyn O'Halloran, Erin Godecke, Michael David & Suhail A. Doi (2020): Inter-rater reliability, intra-rater reliability and internal consistency of the Brisbane Evidence-Based Language Test, Disability and Rehabilitation, DOI: 10.1080/09638288.2020.1776774

To link to this article: https://doi.org/10.1080/09638288.2020.1776774

Published online: 22 Jun 2020.

DISABILITY AND REHABILITATION
https://doi.org/10.1080/09638288.2020.1776774

Taylor & Francis
Taylor & Francis Group

ORIGINAL ARTICLE

OPEN ACCESS Check for updates

Inter-rater reliability, intra-rater reliability and internal consistency of the Brisbane Evidence-Based Language Test

Alexia Rohde[a,b], Molly McCracken[b], Linda Worrall[b], Anna Farrell[c], Robyn O'Halloran[d], Erin Godecke[e], Michael David[f] and Suhail A. Doi[g]

[a]Speech Pathology Department, Southern Cross University, Bilinga, Australia; [b]Speech Pathology Department, The University of Queensland, Brisbane, Australia; [c]Department of Speech Pathology, Royal Brisbane and Women's Hospital, Brisbane, Australia; [d]Department of Community and Clinical Allied Health, La Trobe University, Melbourne, Australia; [e]Department of Speech Pathology, Edith Cowan University, Joondalup, Australia; [f]School of Public Health, The University of Queensland, Brisbane, Australia; [g]Department of Population Medicine, Qatar University, Doha, Qatar

ABSTRACT

Purpose: To examine the inter-rater reliability, intra-rater reliability, internal consistency and practice effects associated with a new test, the Brisbane Evidence-Based Language Test.

Methods: Reliability estimates were obtained in a repeated-measures design through analysis of clinician video ratings of stroke participants completing the Brisbane Evidence-Based Language Test. Inter-rater reliability was determined by comparing 15 independent clinicians' scores of 15 randomly selected videos. Intra-rater reliability was determined by comparing two clinicians' scores of 35 videos when re-scored after a two-week interval.

Results: Intraclass correlation coefficient (ICC) analysis demonstrated almost perfect inter-rater reliability (0.995; 95% confidence interval: 0.990–0.998), intra-rater reliability (0.994; 95% confidence interval: 0.989–0.997) and internal consistency (Cronbach's $\alpha = 0.940$ (95% confidence interval: 0.920–1.0)). Almost perfect correlations (0.998; 95% confidence interval: 0.995–0.999) between face-to-face and video ratings were obtained.

Conclusion: The Brisbane Evidence-Based Language Test demonstrates almost perfect inter-rater reliability, intra-rater reliability and internal consistency. High correlation coefficients and narrow confidence intervals demonstrated minimal practice effects with scoring or influence of years of clinical experience on test scores. Almost perfect correlations between face-to-face and video scoring methods indicate these reliability estimates have direct application to everyday practice. The test is available from brisbanetest.org.

➤ IMPLICATIONS FOR REHABILITATION

- The Brisbane Evidence-Based Language Test is a new measure for the assessment of acquired language disorders.
- The Brisbane Evidence-Based Language Test demonstrated almost perfect inter-rater reliability, intra-rater reliability and internal consistency.
- High reliability estimates and narrow confidence intervals indicated that test ratings vary minimally when administered by clinicians of different experience levels, or different levels of familiarity with the new measure.
- The test is a reliable measure of language performance for use in clinical practice and research.

ARTICLE HISTORY
Received 25 January 2019
Revised 27 May 2020
Accepted 28 May 2020

KEYWORDS
Aphasia; test; stroke; reliability; psychometric properties; outcome measures

Introduction

Reliable identification of acquired language disorders (aphasia) is a core component of healthcare [1]. Substantial functional disability caused by language impairment features prominently in healthcare decision-making [2]. During the recovery phase, reliable monitoring of language abilities provides an accurate gauge of patient recovery [2]. A deterioration in language performance may indicate a worsening medical condition, such as post-stroke haemorrhagic transformation [3], or conversely, a detected improvement in language skill may indicate betterment in functioning and a response to therapy or intervention. Reliability in language measurement is pivotal in determining treatment effectiveness in research trials, gauging individual patient recovery, and informing critical clinical decisions such as the need for medical intervention or determining the need for referral and assistance post-discharge. Such factors rely heavily upon accurate, reliable assessment of language performance and a patient's ability to communicate [4].

The Brisbane Evidence-Based Language Test (Brisbane EBLT) (brisbanetest.org) is a new adult language test [5]. The test is intended to provide an evidence-based, psychometrically robust

CONTACT Alexia Rohde alexia.rohde@scu.edu.au Department of Speech Pathology, Southern Cross University, Bilinga, QLD, 4225, Australia

alternative to informal or non-diagnostically validated language measures used in stroke care [6,7] and comprehensive formalised tests which are reported to be too lengthy for use in some clinical contexts (e.g., acute hospital ward). The Brisbane EBLT aims to provide a comprehensive, yet user-friendly and efficient new measure to assist in the identification of language deficits within a range of clinical contexts, including the hospital bedside [5]. The 49 subtest Brisbane EBLT is the full version of the assessment, evaluating language across the severity spectrum in the following language domains: verbal expression including repetition, automatic speech, spontaneous speech (picture description), naming, auditory comprehension, actions/gesture, reading, and writing. Certain subtests require the use of two of each of the following everyday objects: cup, spoon, pen and knife. An additional "Perceptual" subtest examines abilities not requiring a verbal or written response (e.g., object to picture matching). Adapted scores and shorter test versions allow the test to adjust to individual patient need and varying clinical settings. This study is the second of two psychometric investigations of this new measure. Test development and diagnostic accuracy analysis examining the test's ability to identify aphasia within acute stroke populations have been described elsewhere (brisbanetest.org) [5]. The aim of this study is to report on the inter-rater reliability, intra-rater reliability, internal consistency and practice effects associated with this new measure.

Materials and methods

Study design

Reliability analysis was completed in a concurrent inter-rater and intra-rater repeated measures study design. All clinician raters, stroke participants (or authorised next of kin) provided informed written consent prior to study participation. This study received ethical approval from The University of Queensland Behavioural & Social Sciences Ethical Review Committee (2013000948) and Metro South Human Research Ethics Committee (HREC/14/QPAH/138). This paper is written in accordance with published Guidelines for Reporting Reliability and Agreement Studies (GRRAS) [8]. The GRRAS guidelines are EQUATOR network guidelines (Enhancing the QUAlity and Transparency Of health Research) of widely accepted criteria for the rigorous reporting of sample selection, study design and statistical analysis in reliability research [8].

Sample size justification

No pilot data for inter-rater ICC existed therefore the expected ICC was assumed to be 0.8 [9,10]. As the amount of between-rater variance could not be estimated, the number of simulations (R) = 10 000 was used for inter-rater sample size calculation. When R is large, the highest precision of estimation of the ICC is achieved when the number of participants approximates the number of raters. Therefore, with an average 95% confidence interval (CI) of the ICC based on 10 000 simulations ($p = 0.8$) a total of 15 participants and 15 clinician raters were required to make the width of the CI less than 0.3 (lower bound 0.610; upper bound 0.898) width = 0.288 [9,10]. This equated to a total of 225 test ratings.

For intra-rater sample size calculation, the criterion value of 0.8 was used to determine the number of consecutive measurements required per clinician rater [9]. To obtain 80% power at the 5% significance level two clinician raters were required to complete 2 ratings on 35 participants [9] after a 2-week interval. This equated to 70 ratings per clinician and a total of 140 Brisbane EBLT ratings.

Participants

Inter-rater reliability analysis required 15 stroke patient participants and 15 clinician raters. Intra-rater analysis required 35 stroke patient participants and two clinician raters who were required to complete their ratings twice. In total, 15 clinicians were recruited as two of the 15 clinicians from the inter-rater reliability study (both with >5 years' experience) went on to complete a second round of ratings for the intra-rater analysis.

Stroke participants

Reliability participants were acute stroke patients randomly sampled from a larger cross-sectional diagnostic accuracy study of 100 study participants [5]. Patients in this larger diagnostic study were consecutive stroke admissions from 21 January to 15 December 2015 at two large tertiary hospitals in Brisbane, Australia. All patients were screened within 2 days of hospital admission. Participants were eligible to participate if they were admitted for ischaemic or haemorrhagic stroke management and deemed sufficiently medically and cognitively able to undergo language assessment if the following criteria were met: aged >14 years; native-level English language ability in both written and spoken language; sustained level of consciousness for >10 min; (cognitive functioning was pragmatically assessed based on a patient's ability to participate in, engage with and complete the required language tasks); absence of any precluding acute medical condition as per treating medical team; and with confirmed stroke site of lesion within the left frontal, parietal, temporal, occipital, limbic or insular lobes, internal capsule, thalamus (including thalamic nuclei), and basal ganglia (caudiate nucleus, putamen, globus pallidus, substantia nigra, nucleus accumbens, and subthalamic nucleus). To optimise test external validity, the presence of common post-stroke non-language but communication-related conditions (affecting vision, hearing, speaking, or writing) such as hemianopia, hemiparesis, dysarthria or apraxia of speech was not used as an exclusionary criterion. For these patients, the presence of these co-occurring conditions was noted, and language test items affected by these conditions were recorded as missing data. Patients with subarachnoid haemorrhage or lesions isolated to the right cerebral hemisphere, right midbrain or subcortical regions, or below were not included [5,11].

All 100 recruited stroke patients were video recorded as they were administered the full 49 subtest Brisbane EBLT. Participants wore lapel microphones and were audio-recorded during the assessment to ensure all patient responses were accurately captured. The test was administered by one of two new-graduate qualified clinicians (speech pathologists) both of whom were familiar with the Brisbane EBLT's administration guidelines (brisbanetest.org). A randomized sample of these 100 video recordings was selected for reliability analysis.

Participant video sampling method and strata size calculation

Videos used for reliability analysis were selected *via* stratified randomisation sampling [12]. The Brisbane EBLT total score obtained from the original face to face clinician ratings provided a single rating which demonstrated no floor or ceiling effects with scores ranging from 7 to 215 (out of a possible 0 to 258). This score was therefore used to provide a universal control for the covariate

influence of language test performance [13,14]. Proportional allocation was used to ensure the selected sample in each stratum level was representative of the larger 100 participant group [15]. The same strata levels were applied to both inter-rater and intra-rater reliability studies, however separate simple randomisation was applied to each. Selected videos within each stratum were then randomized. Selected videos and audio-recordings were checked for sound and video quality. If video positioning or poor recording quality impacted on the ability to accurately rate patient performance these videos were discarded and alternative videos were randomly selected from the sample *via* the same sampling method.

Clinician raters

A total of 15 clinicians were recruited to complete reliability analysis. Clinician raters (speech pathologists) were recruited *via* purposeful sampling based on their level of clinical expertise (5 with <5 years' experience; 5 with 5–10 years' experience and 5 with >10 years' experience). Raters were recruited *via* clinical and research contacts to include clinicians with experience within stroke and non-stroke clinical practice and research. All 15 clinicians participated in the inter-rater reliability analysis and two of these clinicians (both with >5 years' experience) went on further to participate in the intra-rater reliability analysis by completing each of their ratings twice (after a 2 week interval).

Procedure

Stroke participant Brisbane EBLT videos and audio recordings were collected and randomized prior to the commencement of the reliability analysis. Recruited clinician raters signed study consent forms and were given headphones, access to the participant video and audio recordings, paper copies of the Brisbane EBLT and a copy of the Brisbane EBLT test Administration and Scoring Guidelines (brisbanetest.org) [5]. A photocopy of the stroke participant's written responses to the Brisbane EBLT writing subtests was provided to each clinician as is reflective of a usual clinical environment and as these were difficult to visualise fully and score *via* video alone.

Prior to commencing the video ratings, all recruited clinicians were unfamiliar with the Brisbane EBLT. Each clinician was provided with one practice video to watch and score in order to familiarise themselves with the new test. These scores were not included within the analysis. The same practice video was given to all raters. After completing the video, clinicians were given the opportunity to ask questions about the general study procedure (e.g., questions relating to the procedure of watching the videos or factors relating to steps in completing the study). Clinician raters were given only the Brisbane EBLT test form (which includes information on scoring specific test items) and the test Administration and Scoring Guidelines form (which provides general scoring guidance) to assist their marking of patient responses. Clinicians were not provided with any specific Brisbane EBLT training or scoring guidance by the research staff prior to or during the reliability ratings (e.g., the research team did not provide any verbal suggestions of how to score items). The absence of any additional test-specific training (beyond that provided on the test forms) was to ensure the psychometric findings would replicate usual clinical practice, when clinicians would not have any specific training prior to using the test and have to rely on the Brisbane EBLT test form and Administration and Scoring Guidelines form to guide their marking of patient responses. To replicate the usual clinical environment, clinicians were asked to refrain from repeatedly re-watching sections of videos which may be ambiguous due to clinical reasons (i.e., ambiguous patient response). If, however reduced video or audio quality affected scoring, clinicians were instructed to re-watch that section as needed to obtain as accurate a rating as possible.

The 15 inter-rater reliability clinicians watched the same randomised 15 participant videos. The order of the videos was individually randomized for each clinician. Two clinicians went on to participate in the intra-rater reliability study, and watched an additional 20 videos each, bringing the total number to 35. After a two-week interval, these two clinicians each re-watched the same 35 videos in the same randomized order. The two-week interval was selected to ensure clinicians could logistically complete the 70 videos within a 2 month time period. As the schedule required each clinician to watch a minimum of 12 videos before returning and re-scoring the first participant video, any carry-over effect was considered minimal.

Reliability ratings were completed across four independent healthcare sites. No clinician rater knew or had met all other raters in the study. All clinicians completed their ratings independently, were blinded to the reference standard result, other clinicians' ratings and their own prior ratings (where applicable). Clinicians were instructed to score all administered test items as per the scoring guidelines. If test items were mistakenly left blank or missed, the forms were returned and clinicians were asked to score these items (e.g., one clinician accidently (unintentionally) left a whole section of the test unscored and this was returned to the clinician who was asked to score these items).

Statistical analysis

Reliability correlations were performed for the 45 language subtests, the four self-report questions, the five section totals and overall Brisbane EBLT score. While Brisbane EBLT test scores are discrete, the underlying construct being measured (language functioning) was considered a continuous variable. Data was examined for normality and homogeneity of variance to ensure it fulfilled the criteria for parametric tests. Ninety-five percent confidence intervals were calculated for each reliability coefficient.

Inter-rater analysis (degree of agreement among different raters) at the Brisbane EBLT subtest level involved different reliability coefficients dependent upon the number of possible participant responses. Binary questions and questions with up to 3 different possible answer types were analysed using Fleiss's kappa [16] as indicated when analysis involves only a few possible rating levels [17]. An Intraclass Correlation Coefficient (ICC) (two-way random-effects model) was used for questions with multiple possible rating categories and for ordinal variables with >4 possible outcome responses [17]. ICC scores range from 0 to 1 and represent the proportion of the variation in the ratings that is due to the performance of the participant under evaluation rather than factors such as how the rater interprets the rubric. An ICC of 1 indicates perfect agreement whereas a 0 indicates no agreement [17]. Mean inter-rater agreement, the probability for a randomly selected participant, that two randomly selected raters would agree was also calculated for each subtest. Complete percentage agreement across all 15 raters was also determined [17].

Intra-rater reliability (consistency of scoring by a single rater) for each Brisbane EBLT subtest was also examined using Intraclass Correlation Coefficient (ICC) measures of agreement. An ICC 3k (mixed effect model) was used to determine the consistency of clinician scoring over time. Binary questions (nominal variables)

(e.g., yes/no self-report questions) were analysed using a multilevel mixed-effect logistic regression for binomial responses. Ordinal variables (questions with >2 possible participant response types) were analysed using ICC mixed effect model. In addition, potential practice effects, manifested as changes in Brisbane EBLT clinician scoring performance due to increased familiarity with the assessment or potential fatigue effects were also examined [18].

Cronbach's alpha was used to determine the internal consistency of the Brisbane EBLT. Values range between 0 and 1 with highly correlated test items resulting in a higher value of alpha [19]. Finally, the mode of test administration was evaluated to assess for any potential difference between face-to-face scoring and scores obtained from clinicians' rating *via* participant video. An ICC 2,1 two-way random effects model was used to determine if scores obtained across the two mediums were comparable. All statistical analyses were completed using StataIC 13 and correlation index interpreted according to Landis and Koch [20] guidelines for reliability coefficients: slight agreement (0.0–0.20), fair agreement (0.21–0.40), moderate agreement (0.41–0.60), substantial agreement (0.61–0.80), and almost perfect agreement (0.81–1.00).

Results

Participants (stroke patient videos)

Fifteen inter-rater videos and 35 intra-rater participant videos were selected *via* randomised stratified sampling based on Brisbane EBLT language ability as per sample size requirements. Randomised participant videos were on average 48.09 min long and ranged from 31 to 71 min in length. Stratification levels and the number of allocated participants per strata are listed in Table 1. Characteristics of the randomised inter-rater and intra-rater reliability stroke participants are described in Table 2.

Clinician raters

Fifteen clinicians participated in the study of which five had <5 years' experience, five between 5 and 10 years' experience and five had >10 years clinical experience. All 15 clinicians were female, and all participated in the inter-rater analysis. Two clinicians (<5 years' experience) participated in both the inter-rater and intra-rater video ratings. Recruited clinicians included 7 acute hospital clinicians; 5 PhD research students and 3 research staff. Characteristics of the clinician raters are described in Table 3.

Normality of the data

The Brisbane EBLT contains a total of 49 subtests which vary in level of task difficulty. Questions range from simple tasks (where most participants achieved a full score) to difficult tasks (where a minority achieve a score). As such, data at the individual subtest level does not follow a normal distribution. While data

Table 3. Characteristics of inter-rater and intra-rater clinician raters ($n = 15$).

Experience level	Average age (μ) years (SD)	Average number of years since graduation (SD) (range)
<5 years (n = 5)[a]	31.6 (14)	0.5 (0.5) (range 0–5)
5–10 years (n = 5)	31.8 (4.81)	7.9 (1.75) (range 5.5–10)
>10 years (n = 5)	44 (12.35)	16.75 (3.77) (range 12–35+)

[a]Two inter-rater clinicians aged 23 and 30 with 0 (new graduate) and 2 years' clinical experience respectively also participated in the intra-rater study.

Table 1. Stratification levels of participant sample by Brisbane EBLT score.

	Strata Level			
Characteristic	Stratum 1	Stratum 2	Stratum 3	Stratum 4[a]
Participant performance as % of overall Brisbane EBLT score (n = 100)	0–25%	26–50%	51–75%	76–100%
Raw Brisbane EBLT score range within strata level	<42	48–90	94–135	>138
Strata size as representative of total participant sample (n = 100)	10	20	17	53
Intra-rater study strata size (n = 35)	3	7	6	19
Inter-rater study strata size (n = 15)	1	3	3	8

[a]Stratum 4 was the largest group within the sample. This group included participants both with and without mild language conditions as determined by the reference-standard language measure in the original diagnostic accuracy study [5].

Table 2. Characteristics of the stroke participant sample.

Characteristic	Inter-rater reliability study ($n = 15$)	Intra-rater reliability study ($n = 35$)[a]
Age	66.13 (SD 11.52) (range 44–83)	66.90 years (SD 15.74) (range 35–87)
Education (school and tertiary formal education only)	11.86 (SD 2.92) (range 7–17)	11.25 years (SD 3.74) (range 3–18)
Sex	Males 60% (9) Females 40% (6)	Males 63% (22) Females 37% (13)
Language	Monolingual (English) 87%) (13) Bi or multilingual 13% (2)	Monolingual (English) 77% (27) Bi or multilingual 23% (8)
Handedness	Right 80% (12) Left 20% (3) Ambidextrous 0% (0)	Right 83% (29) Left 11% (4) Ambidextrous 6% (2)
Average Brisbane EBLT score (possible range 0–258)	121.9 (SD 52.4) (range 37–197)	128 (SD 55.04) (range 24–202)
Presence of language impairment diagnosis (as per validation reference standard)	Impaired language 93% (14) Language intact 6% (1)	Impaired language 77% (27) Language intact 23% (8)
Stroke type	Infarct 87% (13) • 2 thrombolysis Haemorrhagic 13% (2)	Infarct 91% (32) • 1 clot retrieval • 1 clot retrieval and thrombolysis Haemorrhagic 9% (3)
Target lesion site	Left cerebral hemisphere 80% (12) Left subcortical 27% (4) Note: 1 had both cerebral hemisphere and subcortical involvement	Left cerebral hemisphere 86% (30) Left subcortical 31% (11) Note: 6 had both cerebral hemisphere and subcortical involvement

[a]Intra-rater participant sample included the 15 participants from the inter-rater reliability, plus an additional 20 randomized participants.

transformations were attempted this did not influence the normality of the subtest distributions or distributions of the residuals. However non-normal residuals in multilevel modelling with large sample sizes have been shown to have little or no effect on the parameter estimates [21]. Clinically, subtests are not interpreted in isolation and therefore the overall test normality and homogeneity of variance is instead used to ensure this dataset fills the criteria for parametric tests. The data consistently demonstrates almost perfect ICC correlations, consequently, despite the non-normality of the residual distribution, if there was spurious increase in the correlation estimates the data would still display significantly high correlations [21].

Missing data

Brisbane EBLT scoring guidelines direct clinicians not to penalise due to non-language related deficits. For test items where a co-occurring condition (e.g., severe apraxia of speech, dysarthria, hemianopia or hemiparesis) resulted in inability to determine language functioning, clinicians are directed to leave items blank. The decision as to whether test items were affected by severe co-occurring conditions and to leave test items blank was based on the clinical judgement of each clinician rater. These blank scores were statistically treated as missing data and not included in the analysis. As less than 5% of the data was missing this was considered to have negligible effect on correlation estimates [22].

Estimate of reliability including measures of statistical uncertainty

Inter-rater reliability analysis

Inter-class correlation coefficient (ICC) analysis demonstrated almost perfect agreement (0.995; 95%CI: 0.990–0.998) when comparing 15 clinician total Brisbane EBLT scores of 15 acute stroke subjects (total 225 test ratings) [20]. Inter-rater reliability analysis was also completed at the Brisbane EBLT subtest level. Subtest correlations are listed in Table 4. Fleiss's kappa was calculated for 30 Brisbane EBLT questions with <3 possible response types [16] and was found to demonstrate substantial agreement (0.7165) with an average mean percentage inter-rater agreement of 92% and complete agreement of 76%. Inter-rater ICC and complete and mean percentage agreement were calculated for subtests with >4 possible response types. Subtest ICC estimates ranged from substantial 0.704 to almost perfect 0.994 agreement. The average ICC correlation of 0.704 indicated substantial agreement across all relevant Brisbane EBLT subtests [20].

Intra-rater reliability analysis and practice effects

Intra-rater reliability involved the analysis of two clinicians' scores of 35 videorecorded participants when re-scored after a 2-week interval. ICC analysis demonstrated almost perfect intra-rater agreement (0.994; 95% CI: 0.989–0.997) of the test ratings over time (total 140 test ratings) [20]. Subtest level intra-rater correlations were all almost perfect ranging from 0.822 (95%CI: 0.721–0.892) to 1 (95%CI: NA) [20]. ICC intra-rater subtest results are shown in Table 5.

Clinician raters were unfamiliar with the Brisbane EBLT prior to completing test ratings. Intra-rater consistency estimates therefore can be interpreted in the context of practice effects in clinicians' scoring evidenced by changes in scoring style or method as a consequence of becoming familiar with the new test. The almost perfect consistency in test ratings between clinician results obtained from their first video rating, and their re-rating of the same video 35 participants later demonstrated there was limited clinician practice effect evident in Brisbane EBLT test scores.

Internal consistency

The Brisbane EBLT subtests demonstrated almost perfect internal consistency with a Cronbach's alpha of 0.940 (95%CI: 0.920–1.0) [23]. A high Cronbach's alpha is regarded as >0.80 which demonstrates each subtest is examining the same underlying construct and contributing additional information to the overall total score [23].

Mode of delivery

To ensure scores obtained from video ratings are comparable to typical clinical face-to-face scoring methods, a comparison between scores obtained across these modalities was completed. An ICC (2,1 two-way random effects model) was used to compare clinician face-to-face scores obtained from the previous diagnostic accuracy study with inter-rater video scores obtained in the present reliability analysis. Results indicated almost perfect agreement (ICC 0.998; 95%CI: 0.995–0.999) between test results obtained from these different scoring methods when scoring the same acute stroke participant [20].

Discussion

The aim of this study was to examine the inter-rater reliability, intra-rater reliability and internal consistency of the Brisbane EBLT. Practice effects and the impact of the mode of delivery of clinician ratings (video versus face-to-face scoring methods) were also evaluated. Results demonstrated the Brisbane EBLT total score has almost perfect inter-rater (0.995; 95%CI: 0.990–0.998) and intra-rater reliability (0.994; 95%CI: 0.989–0.997) [20]. Cronbach's alpha estimate was also high (0.940; 95%CI: 0.920–1.0), indicating strong internal consistency [23].

Clinicians with a range of experience levels participated in the study. The almost perfect inter-rater estimates and narrow confidence intervals found across all fifteen clinician scores (irrespective of expertise level) indicate that prior experience has negligible impact on test score. All raters were unfamiliar with the Brisbane EBLT prior to completing ratings. High intra-rater reliability estimates between initial and subsequent scores demonstrate there were minimal practice effects associated with clinicians becoming familiar with the new assessment. These results have direct implications for clinical practice and research and indicate that experienced and newly-qualified clinicians as well as clinicians new to the assessment and those highly familiar with the Brisbane EBLT will record similar scores when evaluating the same participant. Finally, comparison of clinician results of the same stroke participant obtained from face-to-face scoring and those obtained from watching participant videos also demonstrated almost perfect correlations (ICC 0.998; 95%CI: 0.995–0.999) [20], indicating the video reliability results obtained in this study have application for everyday face-to-face clinical practice.

Comparison with other research

The Brisbane EBLT is a new measure, and as yet there are no studies with which to compare this study's reliability estimates. Historically however, a number of existing published language tests are used with high frequency among stroke clinicians [6]. The Western Aphasia Battery (WAB) (and WAB-R) [24], Comprehensive Aphasia Test (CAT) [25], Measure for Cognitive Linguistic Abilities (MCLA) [26] and Boston Diagnostic Aphasia

Table 4. Inter-rater reliability per Brisbane EBLT subtest.

Brisbane EBLT Subtest	No. of output responses	Reliability coefficient	Correlation (Kappa or ICC)	95% Confidence Interval	Mean Inter-Rater Agreement	Complete Agreement
Perceptual subtests 1–6						
Copying gestures	3	Kappa[a]	0.199	0.028–0.271	92%	80%
Object to object matching	3	Kappa	0.964	0.933–1.000	100%	100%
Demonstrating object use	3	Kappa	0.672	0.587–1.000	93%	86%
Demonstrating gestures from pictures	3	Kappa	0.629	0.541–0.738	89%	73%
Object to picture matching	3	Kappa	0.738	0.015–0.777	97%	86%
Picture to picture matching (semantic links)	6	ICC[b]	0.874	0.781–0.946	88%	60%
Perception section total	16	ICC	0.704	0.546–0.859	70%	40%
Auditory comp subtests 7–14						
Following commands	9	ICC	0.947	0.903–0.978	70%	26%
Yes / No Questions	13	ICC	0.968	0.940–0.987	86%	66%
Identifying pictures by description	7	ICC	0.931	0.874–0.971	91%	73%
Identifying objects by function	3	Kappa	0.340	0.321–0.340	96%	93%
Odd one out	3	Kappa	0.778	0.640–0.829	85%	46%
Complex questions	7	ICC	0.977	0.957–0.991	85%	60%
Complex questions self-report	2	Kappa	0.971	0.960–1.000	98%	92%
Synonyms	3	Kappa	0.703	0.632–0.771	83%	46%
Auditory comp section total	41	ICC	0.982	0.966–0.993	41%	0%
Verbal expression subtests 15–29						
Counting 1 to 10	2	Kappa	0.877	0.699–1.000	94%	86%
Sentence completion	3	Kappa	0.960	0.912–1.000	99%	93%
Personal questions	7	ICC	0.833	0.709–0.933	76%	40%
Repetition	5	ICC	0.829	0.698–0.935	77%	46%
Object naming	3	Kappa	0.847	0.748–0.878	94%	66%
Naming actions (verbs)	3	Kappa	0.878	0.833–0.881	94%	80%
Picture naming	5	ICC	0.962	0.926–0.986	91%	73%
Naming objects from around the room	5	ICC	0.903	0.825–0.961	87%	60%
Naming gestures	3	Kappa	0.833	0.786–0.985	90%	66%
Verbal fluency (both items)	<40	ICC	0.991	0.982–0.996	58%	26%
Picture description	17	ICC	0.963	0.931–0.985	51%	13%
Picture description self-report	2	Kappa	0.958	0.941–1.000	97%	84%
Picture description self-report (new)	2	Kappa	0.026	0.054–0.012	97%	83%
Word definitions	5	ICC	0.913	0.844–0.963	72%	40%
Similarities and differences	3	Kappa	0.585	0.543–0.699	74%	40%
Proverbs	4	ICC	0.770	0.630–0.895	74%	46%
Verbal expression section total	51	ICC	0.994	0.989–0.997	23%	0.06%
Reading subtests 30–40						
Object to word matching	3	Kappa	0.810	0.707–0.947	92%	80%
Single word reading	3	Kappa	0.723	0.378–0.924	94%	86%
Word to picture matching	7	ICC	0.917	0.851–0.965	84%	66%
Following written commands	5	ICC	0.961	0.928–0.984	89%	66%
Sums	3	Kappa	0.810	0.780–0.845	91%	73%
Reading sentence aloud	2	Kappa	0.771	0.739–0.788	85%	46%
Medicine label	5	ICC	0.896	0.816–0.956	75%	46%
High level sentence comp.	3	Kappa	0.912	0.892–0.952	94%	73%
Written paragraph comp. self-report	2	Kappa	0.898	0.771–1.000	96%	80%
Written paragraph comp. self-report (new)[c]	2	Kappa	N/A	Missing data	94%	66%
Written paragraph comp. total	22	ICC	0.979	0.961–0.991	59%	40%
Written paragraph inference	2	Kappa	0.713	0.438–0.821	88%	66%
Reading section total	24	ICC	0.982	0.967–0.993	0.08%	0%
Writing subtests 41–40						
Drawing in mouth	2	Kappa	0.007	−0.032–0.00	98%	86%
Copying	3	Kappa	0.524	0.185–0.620	89%	66%
Writing name	3	Kappa	0.639	0.525–0.850	86%	60%
Writing gender and address	5	ICC	0.932	0.877–0.972	84%	53%
Writing to dictation	7	ICC	0.972	0.949–0.989	83%	46%
Written object naming	2	Kappa	0.891	0.865–0.931	94%	86%
Written gesture naming	2	Kappa	0.890	0.797–0.953	94%	86%
Sentence construction	14	ICC	0.954	0.915–0.981	53%	13%
Sentence construction self-report	2	Kappa	0.951	0.922–1.000	97%	83%
Writing section total	30	ICC	0.978	0.958–0.991	33%	0%

[a]Kappa adjusts for the level of agreement that can be expected to occur by chance alone. This index however is affected by prevalence totals and has difficulty making distinctions between participants of a population in which those distinctions are rare [35]. Kappa values therefore can be misleadingly low if a large majority of ratings are at the highest or lowest level [17,35]. In these circumstances, the mean percentage agreement and percentage of complete agreement represent more accurate indications of level of reliability [17].

[b]Subtests with high numbers of possible response categories result in reduced probability of raters recording exactly the same score. The mean percentage and complete agreement for the subtests is therefore substantially reduced, despite the majority of ICC estimates falling within the almost perfect range [20]. For these categories, ICC correlations should be interpreted as the true estimate of inter-rater reliability [17].

[c]Only participants who responded "yes, they had difficulty" with the previous question were asked this question (resulting in minimal data).

Table 5. Intra-rater reliability per Brisbane EBLT subtest.

Brisbane EBLT Subtest	Correlation (ICC)	95% Confidence Interval
Perceptual subtests 1–6		
Copying gestures	0.920	0.868–0.952
Object to object matching	1.0	N/A
Demonstrating object use	0.973	0.955–0.984
Demonstrating gestures from pictures	0.923	0.873–0.954
Object to picture matching	0.822	0.721–0.892
Picture to picture matching (semantic links)	0.984	0.973–0.990
Perception section total	0.979	0.963–0.987
Auditory comprehension subtests 7–14		
Following commands	0.983	0.970–0.990
Yes / No Questions	0.986	0.975–0.992
Identifying pictures by description	0.959	0.931–0.977
Identifying objects by function	0.958	0.929–0.976
Odd one out	0.971	0.951–0.983
Complex questions	0.974	0.955–0.985
Complex questions self-report	0.926	0.552–0.992
Synonyms	0.972	0.953–0.983
Auditory comprehension section total	0.994	0.990–0.996
Verbal expression subtests 15–29		
Counting 1 to 10	0.997	0.967–0.999
Sentence completion	1.0	N/A
Personal questions	0.979	0.964–0.988
Repetition	0.989	0.980–0.993
Object naming	0.966	0.943–0.981
Naming actions (verbs)	0.963	0.937–0.979
Picture naming	0.991	0.983–0.994
Naming objects from around the room	0.984	0.972–0.991
Naming gestures	0.956	0.927–0.974
Verbal fluency (both items)	0.934	0.888–0.962
Picture description	0.976	0.959–0.986
Picture description self-report	0.999	0.991–0.999
Picture description self-report (new)	0.927	0.781–0.978
Word definitions	0.959	0.929–0.976
Similarities and differences	0.913	0.856–0.949
Proverbs	0.873	0.794–0.925
Verbal exp. section total	0.983	0.970–0.990
Reading subtests 30–40		
Object to word matching	0.983	0.970–0.990
Single word reading	0.982	0.969–0.989
Word to picture matching	0.986	0.977–0.992
Following written commands	0.959	0.929–0.976
Sums	0.941	0.902–0.966
Reading sentence aloud	0.994	0.969–0.999
Medicine label	0.951	0.917–0.972
High level sentence comprehension	0.949	0.914–0.970
Written paragraph comprehension self-report	0.996	0.955–0.999
Written paragraph comprehension self-report (new)[a]	N/A	Missing data.
Written paragraph comprehension total	0.986	0.976–0.992
Written paragraph inference	0.995	0.969–0.999
Reading section total	0.994	0.990–0.997
Writing subtests 41–49		
Drawing in mouth	0.994	0.878–0.999
Copying	0.835	0.742–0.899
Writing name	0.891	0.825–0.935
Writing gender and address	0.954	0.923–0.972
Writing to dictation	0.989	0.981–0.993
Written object naming	0.999	0.999–0.999
Written gesture naming	0.974	0.883–0.994
Sentence construction	0.992	0.986–0.995
Sentence construction self-report	0.996	0.996–0.996
Writing section total	0.993	0.987–0.996

[a]Only participants who responded "yes, they had difficulty" with the previous question were asked this question (resulting in minimal data).

Examination (BDAE) [27] are some of the most commonly used language measures used in stroke care [6].

While the WAB-R [24] and BDAE [27] have no published reliability estimates with stroke populations, the WAB [24], CAT [25] and MCLA [26] have undergone this reliability analysis. Historically the WAB is one of the most frequently used language measures both within clinical practice and research [6]. WAB inter-rater reliability was examined through the analysis of eight judges (five speech pathologists; two psychometricians and one neurologist) scores of 10 participants of "various types and severities" [24, p.95] who had been videotaped while completing the WAB. Average intercorrelation of the judges' ratings was found to be extremely high ($\geq$ 0.98) [24]. WAB intra-rater reliability analysis also reported significantly high correlations ($\geq$ 0.79) when comparing three examiners' scores of 10 participants when re-assessed "several months" apart [24, p.94]. Similar inter-rater analysis was completed for the CAT [25]. In this study, videotapes of four participants representing "a range of severity and aphasia types" [25, p.111] were scored independently by five raters (two doctors; three speech pathologists). ICC analysis demonstrated excellent inter-rater agreement (0.722–1.00) for all subtest scores [25]. Inter-rater reliability of the MCLA [26] has also been analysed. In this study, scores of two different raters were compared for a subtest of a normative (non-brain damaged) population. Pearson correlation coefficients indicated high levels of reliability (0.90048–1.00) [26].

Methodologically however, these studies were completed prior to the publication of reliability reporting guidelines [8]. While the WAB and CAT inter-rater studies [24,25] documented the raters' professions, this was absent for the WAB intra-rater study [24] and for the MCLA [26]. The method of statistical analysis was not reported for the WAB, nor was the time interval between the intra-rater ratings [24]. Sampling methods for either the clinician raters or the study participants were not described for any study nor were the demographic characteristics of the participant samples (e.g., age, gender, stroke type). While the CAT reported that inter-rater reliability ratings were completed independently [25], this was absent for the WAB [24] or MCLA [26]. Reliability estimates were also based on limited study samples [28] of 20 test ratings (CAT inter-rater analysis) [25], 80 ratings (WAB inter-rater analysis) [24] and 30 ratings (WAB intra-rater analysis) [24]. All studies lacked reporting of *a priori* sample size calculation to ensure adequate statistical power [29]. Incomplete adherence to quality and reporting criteria means the true reliability of these measures is difficult to ascertain. Compromised methodological quality, such as the absence of blinding of assessors and use of small study sample sizes may spuriously inflate reliability estimates [29]. As such, true test reliability estimates could be substantially lower than those reported when applied within either clinical or research populations which differ from those used within the initial study conditions. This outcome may have significant implications, not only for clinical practice, but also for research, where excess in measurement errors adversely influences the sample size needed, overall study cost, and the power to detect a true treatment effect [30].

Strengths and weaknesses

A strength of this reliability study is in the methodology used and adherence to the published Guidelines for Reporting Reliability and Agreement Studies (GRRAS) [8]. *A priori* sample size calculations were completed for both inter-rater and intra-rater reliability analysis and equated to 225 and 140 test ratings respectively. Clinician raters were purposefully sampled to include clinicians from multiple centres with varying backgrounds and expertise and were blinded to their own, others' ratings and the reference standard. In addition, the participant sample was a randomly selected heterogeneous cohort, stratified based on language level

to represent a range of abilities, including those with and without language impairment as is typical of stroke populations. The high inter-rater reliability estimates found in the current study suggest that Brisbane EBLT test scores are not significantly altered by the location or experience level of clinicians. The generalisability of the result is strengthened by the varied clinical characteristics of the stroke participants, the diversity of clinician raters and the absence of any Brisbane EBLT scorer guidance or training, all of which reflect typical real-world everyday practice [30].

Findings of this study need to be interpreted in the context of a number of factors. Firstly, given the absence of an existing published reference standard language test which assesses language across the severity spectrum, stratification of participants' language ability was based on performance on the index measure, the Brisbane EBLT, the inherent reliability of which may have influenced the stratification process. Secondly, while clinicians were stratified for experience level, they were not randomly selected from the wider professional population. Finally, reliability estimates were obtained using ratings from videoed participant performance. While this method is considered one of the most realistic methods for collecting participant data for reliability studies and controls for the variation in clinician scoring alone [31], the mode of evaluation varies from that of a typical clinical setting. ICC scores obtained across these two rating methods demonstrated almost perfect correlation, a finding supported by previous research [32]. The impact of this mode of delivery on clinician test ratings was therefore considered to be minimal.

Inter-rater reliability estimates obtained at Brisbane EBLT subtest level demonstrated variable levels of reliability. These lower estimates however occurred due to limitations of the statistical characteristics of correlation estimates and do not reflect poor reliability of the Brisbane EBLT language measure. Subtests analysed using the kappa statistic were influenced by the prevalence of ratings within subtest samples, resulting in low estimates despite near perfect agreement [33]. This is a well-documented limitation of this reliability coefficient [33–35]. For these subtests, percentage agreement is a more accurate estimation of true correlation for these variables [17,34]. Conversely, lower percentage agreement for variables with multiple response options was more accurately reflected by ICC estimates [17]. Clinically, reliability estimates based at the subtest level are not typically examined in isolation and the overall test score provides a more representative portrayal of the reliability of the measure when used in practice.

Conclusion

The Brisbane EBLT was found to demonstrate almost perfect reliability when tested by a variety of different clinicians with a range of stroke participants. Findings of this study suggest that Brisbane EBLT test ratings of the same patient will vary minimally when scored by different clinicians, or by the same clinicians at different times. These findings have direct implications for clinical practice and indicate that when a change in test performance is detected, this likely reflects a true difference in patient language ability as test scores are minimally influenced by measurement error. These study results support the use of the Brisbane EBLT as an evidence-based alternative to existing language measures and provide a psychometrically robust assessment of language performance for use within clinical practice and research. The Brisbane EBLT is available for download from brisbanetest.org.

Acknowledgements

The speech pathology department at the Royal Brisbane and Women's Hospital made this study possible. The authors thank the stroke patients, clinicians and other study participants who contributed to this research. Full list of acknowledgements is available at brisbanetest.org.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Sources of funding and role of funders

This work was supported by the Australian Stroke Foundation; Equity Trustees Wealth Services Ltd.; Royal Brisbane and Women's Hospital; and Royal Brisbane and Women's Hospital Foundation. The funders had no role in design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

ORCID

Linda Worrall http://orcid.org/0000-0002-3283-7038
Robyn O'Halloran http://orcid.org/0000-0002-2772-2164

References

[1] National Stroke Foundation. Clinical guidelines for stroke management. Melbourne (VIC); 2010.
[2] Maas MB, Lev MJ, Ay H, et al. The prognosis for aphasia in stroke. J Stroke Cerebrovasc Disc. 2012;21:350–357.
[3] Feteke R, Jeevan D, Marks SJ, et al. Hemorrhagic transformation of ischaemic stroke in patient treated with rivaroxaban. J Hematol. 2013;2:48–50.
[4] Spreen O, Risser AH. Assessment of aphasia. New York (NY): Oxford University Press; 2003.
[5] Rohde A, Doi S, Worrall L, et al. Development and diagnostic validation of the Brisbane Evidence-Based Language Test. Disabil Rehabil. 2020. Available from: https://mc.manuscriptcentral.com/dandr
[6] Vogel AP, Maruff P, Morgan AT. Evaluation of communication assessment practices during the acute stages post stroke. J Eval Clin Pract. 2010;16:1183–1188.
[7] Rohde A, Worrall L, Godecke E, et al. Diagnosis of aphasia in stroke populations: a systematic review of language tests. PLoS One. 2018;13:e0194143.
[8] Kottner J, Audigé L, Brorson S, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. J Clin Epidemiol. 2011;64:96–106.
[9] Eliasziw M, Young SL, Woodbury MG, et al. Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. Phys Ther. 1994;74:777–788.
[10] Saito Y, Sozu T, Hamada C, et al. Effective number of subjects and number of raters for inter-rater reliability studies. Stat Med. 2006;25:1547–1560.
[11] Binder JR, Frost JA, Hammeke TA, et al. Human brain language areas identified by functional magnetic resonance imaging. J Neurosci. 1997;17:353–362.
[12] Altman DG, Bland JM. Statistics notes: how to randomise. BMJ. 1999;319:703–704.

[13] Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. Statist Med. 1999;18:2987–3003.
[14] Kernan WN, Viscoli CM, Makuch RW, et al. Stratified randomization for clinical trials. J Clin Epidemiol. 1999;52: 19–26.
[15] Särndal CE, Swensson B, Wretman J. Model assisted survey sampling. New York (NY): Springer; 2003.
[16] Fleiss JL. Reliability of measurement. In: Fleiss JL. editor. Design and analysis of clinical experiments. New York (NY): John Wiley & Sons; 1986. p.1–32.
[17] Graham M, Milanowski A, Miller J. Measuring and promoting inter-rater agreement of teacher and principal performance ratings. Center for Educator Compensation Reform; 2012. p. 1–33. Available from: files.eric.ed.gov/fulltext/ED532068.pdf
[18] Cohen JA, Fischer JS, Bolibrush DM, et al. Intrarater and interrater reliability of the MS functional composite outcome measure. Neurology. 2000;54:802–806.
[19] Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ. 2011;2:53–55.
[20] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–174.
[21] Maas CJM, Hox JJ. The influence of violations of assumptions on multilevel parameter estimates and their standard errors. Comput Stat Data Anal. 2004;46:427–440.
[22] Schafer JL. Multiple imputation: a primer. Stat Methods Med Res. 1999;8:3–15.
[23] Bland JM, Altman DG. Cronbach's alpha. BMJ. 1997;314:572.
[24] Kertesz A. Western aphasia battery – revised. San Antonio (TX): Hardcourt Assessment; 2007.
[25] Howard D, Swinburn K, Porter G. Comprehensive aphasia test. New York (NY): Psychology Press; 2004.
[26] Ellmo W, Graser J, Krchnavek B. Measure of cognitive-linguistic abilities (MCLA). Norcross (GA): The Speech Bin, Incorporated; 1995.
[27] Goodglass H, Kaplan E, Barresi B. Boston diagnostic aphasia examination. 3rd ed. Baltimore (MD): Lippincott Williams & Wilkins; 2001.
[28] Kline P. The handbook of psychological testing. 2nd ed. London (UK): Routledge; 2000.
[29] Button KS, Ioannidis JPA, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013;14:365–376.
[30] Berg K, Wood-Dauphinee S, Williams JI. The balance scale: reliability assessment with elderly residents and patients with an acute stroke. Scand J Rehab Med. 1995;27:27–36.
[31] Fawcett AL. Principles of assessment and outcome measurement for occupational therapists and physiotherapists: theory, skills and application. Chichester (UK): Wiley; 2007.
[32] Theodoros D, Hill A, Russell T, et al. Assessing acquired language disorders in adults via the internet. Telemed J E Health. 2008;14:552–559.
[33] Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. BMJ. 1992;304: 1491–1494.
[34] Hand PJ, Haisma JA, Kwan J, et al. Interobserver agreement for the bedside clinical assessment of suspected stroke. Stroke. 2006;37:776–780.
[35] Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol. 1990; 43:543–549.