APHASIOLOGY
AN INTERNATIONAL, INTERDISCIPLINARY JOURNAL
VOLUME 37 • NUMBER 7 • JULY 2023
Routledge

Aphasiology

Routledge
Taylor & Francis Group

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/paph20

Statistics and psychometrics for the CAT-N: Documenting the Comprehensive Aphasia Test for Norwegian

Bård Uri Jensen, Monica I. Norvik & Hanne Gram Simonsen

To cite this article: Bård Uri Jensen, Monica I. Norvik & Hanne Gram Simonsen (2023): Statistics and psychometrics for the CAT-N: Documenting the Comprehensive Aphasia Test for Norwegian, Aphasiology, DOI: 10.1080/02687038.2023.2200132

To link to this article: https://doi.org/10.1080/02687038.2023.2200132

Published online: 27 Apr 2023.

APHASIOLOGY
https://doi.org/10.1080/02687038.2023.2200132

Routledge
Taylor & Francis Group

Statistics and psychometrics for the CAT-N: Documenting the Comprehensive Aphasia Test for Norwegian

Bård Uri Jensen [a], Monica I. Norvik [b,c,d] and Hanne Gram Simonsen [b]

[a]Faculty of Education, Inland Norway University of Applied Sciences, Hamar, Norway; [b]MultiLing Center for Multilingualism in Society across the Lifespan, Department of Linguistics and Scandinavian studies, University of Oslo, Oslo, Norway; [c]Department of Acquired Brain Injury, Statped, Oslo, Norway; [d]Department of Language and Literature, NTNU – The Norwegian University of Science and Technology, Trondheim, Norway

ABSTRACT

Background: Ivanova and Hallowell 2013 emphasise the importance of reporting on test development and psychometric properties of tests in international journals. Such documentation may serve as references for other test developers and enable researchers and clinicians to assess reliability and validity issues in tests made for a language unknown to them.

The CAT (Comprehensive Aphasia Test) is a general aphasia test which examines linguistic skills broadly, within the cognitive neuropsychological tradition; it has been and is being adapted to a number of languages.

Aims: The aim of this article is to document the statistical procedures used in the development and standardisation of the Norwegian adaptation of the CAT (CAT-N), to document its psychometric properties, and to discuss validity and reliability issues.

Methods & procedures: The adaptation of the CAT-N involved careful design of subtests and test items, taking into account features like word frequency, imageability and phonological and other language-specific linguistic variables. The prototype was tested on a normative sample of 85 persons with aphasia and a control group of 84 persons without aphasia. The items of some subtests were reordered based on the norming. A new scoring scheme was developed for two subtests of Picture description. The CAT-N includes the Aphasia Impact Questionnaire (AIQ), which is a new patient reported outcome measure developed for the CAT.

Outcomes & Results: Statistical methods are documented and discussed. Descriptive statistics for subtests and linguistic domains are presented. Internal consistency and partial inter-rater and intra-rater reliability aspects are investigated and documented. Construct validity is investigated and documented by factor analysis. Sensitivity and specificity are investigated through pairwise comparisons for subtests and domains and the use of normal-language cutoff values.

ARTICLE HISTORY
Received 09 Jul 2022
Revised 31 Mar 2023
Accepted 04 Apr 2023

KEYWORDS
aphasia assessment; Comprehensive Aphasia Test (CAT-N); statistical analysis; psychometrics; Norwegian language

CONTACT Bård Uri Jensen bard.jensen@inn.no Faculty of Education, Inland Norway University of Applied Sciences, Postboks 400, 2418 Elverum, Norway

Concurrent validity is investigated through comparisons with results from an existing aphasia test for Norwegian (NBAA).

Conclusions: The CAT-N is shown to have good reliability and validity, and it distinguishes well between persons with and without aphasia. The article provides explicit documentation of design decisions which may be useful in future adaptations of the CAT.

Introduction

There is a widely acknowledged need for valid and comparable language assessment tools for people with aphasia (PWA), for research purposes as well as for clinical use. In aphasia research, there is huge variability across countries and disciplines (Gitterman et al., 2012), making comparison across languages and groups difficult. This restricts the possibilities for large-scale investigations, for both monolingual and multilingual individuals with aphasia. For clinical purposes, there is again variation between languages: for English, for instance, there are many language assessment tools to choose from (see e.g. Howard et al., 2010), while for other languages, e.g. Croatian, Turkish, Hungarian, there are few and of varying quality (Kuvač Kraljević et al., 2020; Maviş et al., 2021; Zakariás & Lukács, 2022). For Norwegian, there are several aphasia tests of different types, both general and more specific ones. However, *Norsk Grunntest for Afasi* (the Norwegian Basic Aphasia Assessment, NBAA; Reinvang & Engvik, 1980) is the only comprehensive and standardised aphasia test and primarily aimed at classifying different aphasia syndromes. Furthermore, it is relatively old and slightly outdated.

The need of comparable assessment tools between different languages has been explicitly addressed in the COST network Collaboration of Aphasia Trialists (CATs; IS1208, 2013-2023, since 2017 funded by the Tavistock Trust for Aphasia). The network includes participants representing 40 countries and 43 languages from across the world (https://www.aphasiatrials.org/). Within this network, the working group on "Aphasia Assessment and Outcomes" decided in 2013 to adapt the Comprehensive Aphasia Test (CAT; Swinburn et al., 2004) to the different languages. This decision was based on the need for a comprehensive, yet relatively short test, including cognitive, linguistic and psychosocial aspects. The CAT is widely used in English-speaking countries, and validated and normed for English (Howard et al., 2010). Furthermore, the linguistic parts of the CAT are explicitly based on several psycholinguistic and linguistic variables (e.g. frequency, imageability, word and sentence length and complexity) facilitating an adaptation into different languages (Fyndanis et al., 2017). Presently, the test is under adaptation into 16 languages, in addition to the five that are already published (Croatian: Swinburn et al., 2020; Dutch: Visch-Brink et al., 2014; Hungarian: Zakariás & Lukács, 2022; Norwegian: Swinburn et al., 2021; Turkish: Maviş et al., 2021).

It is well established that a simple translation of an assessment instrument is never appropriate to obtain comparability (Bates et al., 1991; Paradis & Libben, 1987). However, neither is an adaptation from one language to another straightforward. Languages differ in many aspects, and the more a language in an adapted version differs from the original, the more difficult it is to assess important aspects of that language, yet maintain comparability. Such challenges and solutions in the adaptation of the CAT are discussed in detail in Fyndanis et al. (2017).

In addition to the linguistic challenges, a great number of design decisions pertaining to statistical procedures go into the development and standardisation of a test. Unfortunately, these often remain undocumented and unpublished (Ivanova & Hallowell, 2013). This lack of available documentation places a heavier burden than necessary on future test developers, and it increases the risk of introducing unnecessary discrepancies between test adaptations and hence reduces their intended comparability. The main aim of the present paper is to provide thorough documentation of the statistical procedures used in the development of the CAT-N, which may then function as guidelines for future adaptations of the CAT. In addition, we briefly outline aspects of the adaptation process, describe innovations of the CAT-N and present and discuss its psychometric properties, including reliability and validity aspects.

The adaptation of the CAT-N

The adaptation to Norwegian was conducted by a team of speech and language therapists (SLTs) and linguists (the Norwegian adaptation team). To ensure linguistic and cultural comparability across the different language versions, a set of criteria and guidelines were developed within the working group of "Aphasia Assessment and Outcomes" mentioned above. First, the fundamental decision was made to use the same number of subtests and items in all versions of the CAT. Furthermore, since the CAT is explicitly based on several psycholinguistic variables, ways to establish those within each language had to be agreed upon. With respect to frequency (how often a word is used), measures should ideally be taken from spoken corpora, but written corpora could also be used since spoken and written frequency measures generally correlate well (Pastizzo & Carbone, 2007). For Norwegian, we used the web-based written corpus *NoWaC*, based on 700 million words taken from the .no domain (Guevara, 2010). As for imageability (how easily a word evokes a mental image), values have to be based on ratings from native speakers (Paivio et al., 1968). For Norwegian, we used imageability values from the database *Norwegian Words* (Simonsen et al., 2013; Lind et al., 2015). This is a searchable lexical database containing approximately 1650 nouns, verbs, and adjectives, for which one can get information about (psycho)linguistic variables that are known to affect language acquisition, storage, and processing of words (e.g. imageability, frequency, age of acquisition, sound structure, and word length).

The words used in the test could most often not be directly translated from English. For example: in one subtest, the English monosyllabic word *pear* was used. In Norwegian, the corresponding word is the disyllabic *pære* /^{2}pæːɾə/. If the point was to see whether persons with aphasia understood the word for *pear* in their language, a translation would have been appropriate. However, in the subtest in question only single-syllable words should be included. Furthermore, one distractor word was supposed to sound nearly the same, to see whether the person could distinguish phonologically similar words. In the English version of the CAT this word is *bear*, but in Norwegian the translation is *bjørn* /bjøːɳ/, which would not work. Thus, in order to fill the linguistic criteria and maintain comparability across languages, actual test words in the various languages are often very different.

Many of the subtests in the CAT are picture-based. For the new versions, the Croatian artist Marko Belić was engaged to draw new black-and-white illustrations. Both for linguistic and cultural reasons, many new pictures were needed. To make sure that a picture actually evoked the right word, naming agreement tests were conducted in all languages, where only pictures rated with the same word by at least 80 % of the (20+) raters were accepted. Many pictures had to be redrawn many times – for example, the word for *mouse* in Norwegian (*mus*) needed several rounds with different sizes and addition of cheese so as not to be confused with *rat* (*rotte*). And the Norwegian word for *waffle* (*vaffel*) needed to have its local, heart-shaped form to be recognised as such – although the cognate word in English is very similar, the shape of the item is different.

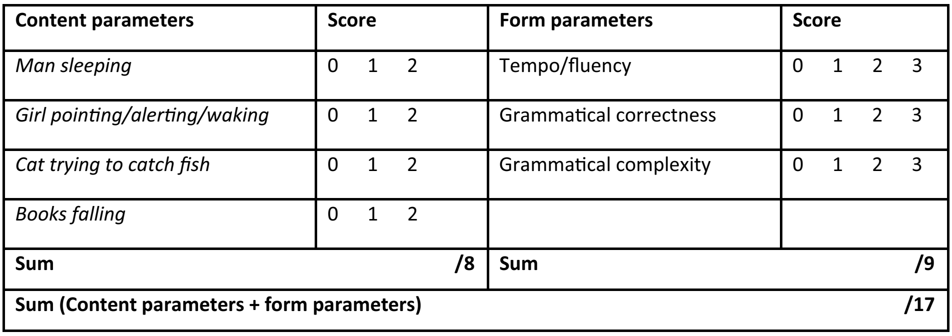
Innovations in the CAT-N

While we took care to follow all criteria and guidelines that were agreed upon across languages in the "Aphasia Assessment and Outcomes" working group, we also made certain innovations. One concerned the scoring of the picture descriptions (subtests 19 and 27). Another was to rearrange the order of some of the items in five of the subtests (subtests 7-10 and 17). The third innovation concerned the replacement of the Disability Questionnaire (DQ), which was part of the original CAT from 2004, with the Aphasia Impact Questionnaire (AIQ).

Picture descriptions

The CAT includes two narratives elicited from the same picture, one oral and one written. The picture shows a man sleeping in his chair while a cat on a shelf above tries to catch goldfish from a fishbowl, and at the same time pushing down a row of books, falling towards the man's head. A child playing on the floor tries to awaken and alert the man. The working group agreed that the original scoring system was too complex, in particular for clinical purposes, but did not decide on a common scoring system.

For the CAT-N, we wanted to score both grammatical skills (form) and how well the participant could describe what happened in the picture (content). We took as our point of departure the scoring system from the Dutch adaptation (CAT-NL; Visch-Brink et al., 2014) and decided on three parameters for form (tempo/fluency (relevant only for oral description); grammatical complexity; grammatical correctness), and four content units (man sleeps; girl points/awakens/alerts; cat chases fish; books fall). Each of the form parameters was scored on a scale from 3 (good) via 2 (medium) and 1 (weak) to 0 (missing), and the content units on a scale from 2 (complete and precise) via 1 (present, but not complete and/or precise) to 0 (missing). The scores were logged on a separate scoring sheet, an English translation of which is shown for the oral subtest in Figure 1. (See sections below for more details on scoring and on inter-rater reliability for the Picture descriptions.)



Content parameters	Score	Form parameters	Score
Man sleeping	0 1 2	Tempo/fluency	0 1 2 3
Girl pointing/alerting/waking	0 1 2	Grammatical correctness	0 1 2 3
Cat trying to catch fish	0 1 2	Grammatical complexity	0 1 2 3
Books falling	0 1 2		
Sum	**/8**	**Sum**	**/9**
Sum (Content parameters + form parameters)			**/17**

Figure 1. The Scoring Sheet for Oral Picture Description, Translated From Norwegian. *Note*. The scoring sheet for the Written picture description is identical, apart from the fluency item being excluded.

Reordering of items in five subtests

For five of the subtests there is an abortion rule (subtests 7-10 and 17); if the participant fails to obtain a positive score for a certain number of consecutive items, that subtest should be aborted. In the norming study, the items in these tests were given in an arbitrary order, and the test administrators were asked not to use the abortion rule if possible, in order for us to be able to investigate whether some items were more difficult than others. For the published test, the items were reordered and put in ascending order of difficulty (see below for details).

Aphasia Impact Questionnaire

The Disability Questionnaire (DQ) was not included in the adaptation of the CAT across languages. However, as reported by Swinburn et al. (2019), it was decided to replace the original DQ in the CAT with a new patient reported outcome measure: the Aphasia Impact Questionnaire (AIQ). (This is to be published in the new, second edition of the English CAT, expected in 2023. The CAT-N is thus the first CAT with AIQ included.) The AIQ is constructed to explore and evaluate the consequences of living with aphasia, and assesses communication, participation, and emotional well-being through a picture-based scale with five response alternatives. For the Norwegian version, we carried out a focus group interview with a group of five persons with aphasia, following an interview guide primarily focusing on reading and writing, and whether the Norwegian AIQ should include digital communication. This resulted in the addition of one question in the AIQ of CAT-N concerning how aphasia affects daily functioning in a digital world, increasing the number of questions in the AIQ from 21 to 22. The final version was then tested on 21 persons with aphasia, not to establish norms, but to see how the AIQ functions in actual use. The CAT-N is the first test for Norwegian which addresses both impairment-based and consequence-focused issues.

Method

Methods of statistical analysis

All calculations and statistical analyses were carried out using R, a software environment for statistical computing (R Core Team, 2020).

Following common convention, we chose α=0.05 as level of significance. To control for familywise error rate (FWER), significance levels were adjusted using Šidák correction when relevant (Abdi, 2007; Šidák, 1967).

Normality of distributions was tested using the Shapiro-Wilk test of normality (Royston, 1995; Shapiro & Wilk, 1965). Skewness was measured as the dimensionless version of the third moment about the mean (γ_1) (Crawley, 2007, pp. 285-287).

Comparisons of two independent samples were made using the rank-based Wilcoxon test (Wilcoxon, 1945; Zimmerman & Zumbo, 1990). Since many of the scores from the test are normally distributed, we could sometimes alternatively have used the parametric Welch' *t*-test, but chose to use the same two-sample test for all analyses in order to simplify comparison of results. Correspondingly, all two-sample effect sizes are given as the point-biserial correlation coefficient r_{pb} (LeBlanc & Cox, 2017). Cohen (1992, p. 99) gives guidelines for interpreting values of r_{pb}:

$r_{pb} \approx 0.1$	small effect
$r_{pb} \approx 0.3$	medium effect
$r_{pb} \approx 0.5$	large effect

We do not report values of Cohen's *d*, since many of the distributions are skewed; in particular, the limited dispersion of the control group values would artificially amplify values of *d*.

Correlations were tested using Spearman's rank-based correlation and the effects are given as r_s (see e.g. Levshina, 2015, pp. 130-134). 95 % confidence intervals (CI) for r_s were calculated using the function spearman.ci in the R package RVAideMemoire (Hervé, 2021).

Internal consistency of subtests and domains was assessed using Cronbach's alpha (Cronbach, 1951), even though using this as the sole instrument has been criticised lately (Cronbach & Shavelson, 2004; Sijtsma & Pfadt, 2021). We used the function cronbach in the R package psy (Falisarrd, 2012). Ivanova and Hallowell (2013, p. 907) refers to a coefficient value of 0.7 as representing "sufficient" internal consistency and Cohen et al. (2011, p. 640) give further guidelines for interpreting the coefficient values:

$\alpha > 0.90$	very highly reliable
$\alpha > 0.80$	highly reliable
$\alpha > 0.70$	reliable
$\alpha < 0.60$	unacceptably low reliability (not reliable)

We evaluated inter-rater reliability with Krippendorff's alpha (Krippendorff, 2004), using the function kripp.alpha in the R package irr (Gamer et al., 2019). Krippendorff suggests that $\alpha > 0.800$ could be regarded as acceptable, $\alpha > 0.667$ only for "tentative conclusions" (p. 241).

Factor analysis with varimax rotation[1] was carried out using the factanal function in the R package stats (R Core Team, 2020). We defined the minimal adequate model as the

Table 1. Overview of the 27 Subtests and 9 Domains of the CAT-N.

No	Subtest	Domain
1	Line bisection	
2	Semantic association	Memory
4	Recognition memory	
5	Gesture object use	
6	Arithmetic	
7	Comprehension of spoken words	Auditory comprehension
9	Comprehension of spoken sentences	
11	Comprehension of spoken stories	
8	Comprehension of written words	Reading comprehension
10	Comprehension of written sentences	
12	Repetition of words	Repetition
13	Repetition of complex words	
14	Repetition of nonwords	
15	Repetition of digit strings	
16	Repetition of sentences	
17	Naming objects	Naming
18	Naming actions	
3	Word fluency	
19	Spoken picture description	Picture description: spoken
20	Reading words	Reading
21	Reading complex words	
22	Reading function words	
23	Reading nonwords	
24	Writing: copying	Writing
25	Writing picture names	
26	Writing to dictation	
27	Written picture description	Picture description: written

smallest model which was not significantly different from a perfect fit, using the built-in X^2 statistic of the factanal function.

Overall design

The entire CAT-N consists of 6 cognitive and 21 linguistic subtests,[2] in addition to the AIQ. Of the 27 subtests, 24 contribute to a total of 9 domains[3], of which 8 are linguistic domains. Two of the linguistic domains consist of just one subtest each, whereas three of the cognitive subtests do not contribute to any domain. The overview of subtests and domains is shown in Table 1 and corresponds to the original CAT.

Respondents and procedure

When the adaptation was completed, we recruited SLTs across the whole country to collect data for the norming study. Following Ivanova and Hallowell's (2013, p. 906) strong recommendation, our initial aim was a sample of 100 PWA and 100 persons without aphasia as a control group (CG). When the Covid-19 pandemic started, testing became difficult, so the ultimate number of participants was 85 PWA and a CG of 84. Inclusion criteria for PWA were 1) known diagnosis of aphasia as a result of stroke, 2) aphasia in all phases from acute to chronic, 3) having capacity to consent. The PWA had already been assessed according to the general procedure of their SLT; for about 50 %, test scores from the *Norwegian Basic Aphasia Assessment* (Reinvang & Engvik, 1980) were supplied. Post-

Table 2. Some Characteristics of the Test Group of People With Aphasia and the Control Group.

Property	Control	Aphasia
N	84	85
Female	49 (58 %)	25 (29 %)
Male	35 (42 %)	60 (71 %)
Age: minimum – maximum	21 – 85	25 – 86
Age: mean (standard deviation), median	56.9 (17.0), 59	61.8 (13.9), 64
Primary and secondary school	6 (7 %)	20 (24 %)
Tertiary education	11 (13 %)	31 (36 %)
≤ 3 years higher education	21 (25 %)	19 (22 %)
> 3 years higher education	46 (55 %)	15 (18 %)
Post-onset (days): mean (sd), median		815 (1220), 371
Post-onset (days): 20-90, 91-365, 365-5078		21, 20, 43

Note. One PWA became unavailable for testing after subtest 12. *N* = Sample size.

onset times (shown in Table 2) varied from 20 days to almost 14 years. We have no further information on neurological background or aphasia severity.

In terms of age, the two groups are quite similar, but the proportion of women is greater in the control group than among the PWA, as is the proportion of people with more than 3 years of higher education. As shown in Table 5, *N* varies somewhat between the subtests. The reason for this is that individual subtests occasionally could not be scored, typically because the test person accidentally was given faulty instructions. Also, one PWA became unavailable after subtest 12.

SLTs collected test data for the PWA by scoring the various subtests following the instructions on the scoring sheets. The two Picture descriptions were not scored by the SLTs, however; recordings of the Oral descriptions were transcribed and both the transcriptions and the Written descriptions were scored by members of the project group, three SLTs and one linguist.

Analysis

Basic scoring of items and subtests

Many subtest items are given raw scores 1 for correct answer and 0 for erroneous or lacking answer. In some subtests, there is a range of obtainable scores for each item, often 0–2. In a few subtests, the scoring scheme is a bit more complicated.

Raw scores of most subtests are obtained by simple addition of the scores of the individual test items. In subtest 1, Line bisection, a lacking answer for one or more items results in no score (NA) for that subtest.

Raw scores of domains are obtained by simple addition of the raw scores of the subtests in that domain. This procedure has the consequence that some subtests contribute considerably more to the domain score than others.

All subtests and domains thus have a theoretical lower and upper bound of their scores. We call these the pessimum value and the optimum value, respectively. These theoretical scores are not achieved in all instances, however, and the smallest and largest actual score of a subtest or a domain are denoted the minimum score and the maximum score, respectively. The minimum and maximum will typically vary between PWA and the CG.

T-scores

T-scores are scores which are standardised to mean (m) 50 and standard deviation (sd) 10 (Clark-Carter, 2005, p. 2067; Frick et al., 2010, pp. 28-29). Raw scores both from subtests and from domains were transformed to T-scores using the rank-based Rankit algorithm (Bliss et al., 1956), which is designed to transform any distribution into one which is approximately normal and with known mean and sd. Solomon and Sawilowsky (2009) compare the four most commonly used transformation algorithms and conclude that the Rankit algorithm has the best performance in general and particularly for heavily skewed raw scores like the ones present in most of the subtests. The outline of the algorithm is as follows:

(1) Sort all raw scores in ascending order of size.
(2) Assign a rank value to each raw score, so that the lowest raw score becomes 1, the second lowest 2, etc. Ties are to be given the mean of their rank values.
(3) Subtract 0.5 from each rank value.
(4) Divide each resulting value by the total number of values.
(5) Apply the inverse cumulative normal distribution function to these values, with the parameters m=50, sd=10.[4]
(6) Round to nearest integer. Rounding is not strictly necessary, but presenting decimals gives a false impression of higher accuracy than is actually present.

The above algorithm was applied to the set of raw scores of the PWA for each subtest and each domain. Since all possible test scores were not present as raw scores in the norm set, conversion keys had to be calculated for the missing values; any values below the minimum value in the norm set were given the T-score given to the minimum raw score and correspondingly for values above the maximum value in the norm set. Hence, obtaining a T-score outside the range present in the normative sample is not possible. Any missing values within the range of actual scores were calculated as linear interpolations of the nearest non-missing (unrounded) T-scores, followed by rounding.

The advantage of T-scores over for example z-scores is that the resulting numbers are more manageable, ranging typically between 20 and 80. The actual boundaries depend on the number of observations in the normative set. Hence, the range of T-scores may vary between different versions of the CAT, depending on the size of the normative set. With *N*=85, all T-scores lie between 25 and 75. Given a normal distribution of T-scores, the values may be used to estimate percentiles (see Table 3).

For raw score distributions which are symmetrical and without prominent mode values near the boundaries, the transformation algorithm works well in transforming the original distribution into a normal distribution. However, most of the subtests have heavily left-

Table 3. T-scores and Corresponding Percentiles in a Normal Distribution.

T-score	Percentile
30	2
40	16
50	50
60	84
70	98

Table 4. Statistics Indicating the Amount of Deviation From the Normal Distribution for the Domain T-scores.

Domain	W	p	γ_1
Memory	0.887	0.000	−0.556
Auditory comprehension	0.988	0.630	−0.072
Reading comprehension	0.994	0.976	−0.066
Repetition	0.994	0.963	−0.006
Naming	0.998	1.000	0.006
Spoken picture description	0.979	0.200	−0.089
Reading	0.978	0.165	−0.145
Writing	0.988	0.634	−0.020
Written picture description	0.899	0.000	0.396

Note. W and p from Shapiro-Wilk's test, γ_1 index of skew.

skewed distributions, many with the optimum as the mode value. Transforming such distributions into truly normal distributions is mathematically impossible. Consequently, T-scores for most subtests are not truly normal, but remain left-skewed and with prominent extreme mode values. Also, the maximum T-scores will not be 75 in such left-skewed distributions, but below 75. Hence, the T-score-to-percentile conversion table above will not be accurate for subtests with heavily skewed distributions. Also, since the skew varies between the subtests, T-scores are not always directly comparable between subtests. Similarly, they are not directly comparable between different language versions of the CAT.

For the domain variables, on the other hand, the resulting distributions of T-scores are largely normal. Written picture description and Memory are the most obvious exceptions. In the latter, 30 out of 85 PWA score the optimum, whereas in the former, 27 out of 84 PWA score the pessimum, i.e. zero (and 5 score the optimum). For Memory, this is a natural consequence of the capacity of memory not necessarily being affected by aphasia (Papathanasiou et al., 2017, p. 4). For Written picture description, the reason is likely to be a near complete loss of writing ability in many of the PWA in the normative sample. We assume these to be natural properties of any test result from a normative sample of PWA.

Normality scores for the domain T-scores are given in Table 4; left and right skews are represented by negative and positive values of γ_1, respectively. As can be seen, Spoken picture description and Reading are also somewhat skewed, although the p-values from the Shapiro-Wilk normality test are above the conventional 0.05 level. The skew in both these domains is caused by a clustering of values near the optimum. In Spoken picture description, 36 of 82 PWA (44 %) score between 14 and the optimum raw score of 17. In Reading, 31 of 84 PWA (37 %) score between 66 and the optimum raw score of 70. In these two domains, it is reasonable to surmise that the subtests were slightly too easy for the normative sample and that a more desirable distribution could have been obtained with slightly increased difficulty. As explained above, the low difficulty of the subtest results in lower optimum T-scores when the test is administered, in the case of Reading causing the optimum T-score to be only 66 (see Figure 2), noticeably lower than the possible optimum of 75 for N=84. For Memory and Written picture description, the deviation from normality is too substantial for parametric methods of analysis to be applied when these domains are involved.

Table 5. Raw Scores for Subtests, Control Group and PWA.

			Control							Aphasia						
Subtest	Pess	Opt	*N*	Med	*M*	SD	Min	Max	C	*N*	Med	*M*	SD	Min	Max	*n*<C
Line bisection	6	0	83	1	0.93	0.75	3.5	0	2	82	1.25	1.43	1.07	5	0	14
Semantic association	0	10	84	10	9.86	0.35	9	10	9	85	10	9.13	1.47	2	10	18
Recognition memory	0	10	84	10	9.80	0.46	8	10	9	85	10	8.73	2.11	0	10	81
Gesture object use	0	12	84	12	11.61	0.69	9	12	10	85	11	10.11	2.49	0	12	22
Arithmetic	0	6	84	6	5.45	0.67	4	6	4	85	5	4.16	1.72	0	6	22
Auditory comprehension: words	0	30	84	30	29.81	0.50	27	30	29	85	29	27.67	3.08	15	30	25
Auditory comprehension: sentences	0	32	84	32	31.37	0.92	28	32	30	85	24	22.95	6.31	6	32	38
Auditory comprehension: stories	0	4	84	4	3.87	0.46	1	4	3	85	3	3.02	1.12	0	4	52
Reading comprehension: words	0	30	84	30	29.93	0.30	28	30	29	85	28	24.98	5.99	4	30	68
Reading comprehension: sentences	0	32	84	32	31.12	1.23	26	32	29	85	22	20.18	7.98	2	32	74
Repetition: words	0	32	84	32	31.83	0.79	27	32	32	85	30	26.21	7.56	2	32	22
Repetition: complex words	0	6	84	6	6	0	6	6	6	84	6	4.70	2.03	0	6	54
Repetition: nonwords	0	10	84	10	9.54	1.24	2	10	8	84	7	6.46	3.13	0	10	31
Repetition: digit lists	0	14	84	12	12.40	1.78	8	14	8	84	8	8.60	3.24	0	14	44
Repetition: sentences	0	12	84	12	11.98	0.22	10	12	12	84	8	8.05	4.51	0	12	22
Naming: objects	0	48	84	48	46.42	2.45	37	48	41	84	34	30.76	12.07	0	48	49
Naming: actions	0	10	84	10	9.85	0.65	6	10	8	84	8	7.46	2.70	0	10	66
Word fluency	0	33*	84	41.5	41.05	9.63	21	60	25	85	10	11.02	7.56	0	33	32
Picture description: spoken	0	17	84	17	16.75	0.53	15	17	16	82	13	11.66	3.91	0	17	71
Reading: words	0	48	84	48	47.98	0.15	47	48	48	84	43	35.96	14.90	0	48	62
Reading: complex words	0	6	84	6	5.98	0.22	4	6	6	84	4	3.50	2.43	0	6	54
Reading: function words	0	6	84	6	6	0	6	6	6	84	6	4.68	2.22	0	6	27
Reading: nonwords	0	10	84	10	9.93	0.37	8	10	10	84	6.5	5.79	3.81	0	10	61
Copying	0	27	84	27	26.82	1.53	13	27	27	81	26	21.16	8.51	0	27	41
Written naming	0	21	84	21	20.54	0.78	17	21	19	84	16	12.63	7.60	0	21	58
Dictation	0	28	84	28	27.89	0.31	27	28	27	83	20	16.10	10.50	0	28	67
Picture description: written	0	14	84	14	13.62	0.92	9	14	12	84	5	5.39	4.97	0	14	73

Note. Pess = pessimum; Opt = optimum; *N* = number of persons; Med = median; *M* = mean; SD = standard deviation; Min = minimum; Max = maximum; C = 95 % cutoff; *n*<C = number of PWA below cutoff.
* There is no theoretical optimum for Word fluency; the value is set to the maximum obtained among PWA.

Scoring of Picture descriptions: oral and written

The two subtests for picture description, oral and written, comprise items related to form and items related to content. In both subtests, there are 4 items related to content, each with an optimum score of 2 points, yielding a total optimum score of 8 for content. The optimum score for each formal item is 3, however, and the number of formal items differs between the subtests: 3 for the oral descriptions and 2 for the written, yielding total

Table 6. Raw Scores for Domains, Control Group and PWA.

			Control							Aphasia						
Domain	Pess	Opt	*N*	Med	*M*	SD	Min	Max	C	*N*	Med	*M*	SD	Min	Max	n<C
Memory	0	20	84	20	19.65	0.65	17	20	19	85	19	17.86	2.85	4	20	38
Auditory comprehension	0	66	84	66	65.05	1.27	61	66	63	85	55	53.65	9.32	30	66	69
Reading comprehension	0	62	84	62	61.05	1.32	55	62	58	85	50	45.15	13.20	6	61	73
Repetition	0	74	84	72	71.75	2.86	58	74	68	84	61.5	54.26	17.66	2	74	62
Naming	0	91*	84	99	97.31	10.61	75	118	79	84	52	49.14	19.73	0	91	81
Spoken picture description	0	17	84	17	16.75	0.53	15	17	16	82	13	11.66	3.91	0	17	71
Reading	0	70	84	70	69.88	0.45	68	70	69	84	60	49.93	21.93	1	70	71
Writing	0	76	84	76	75.25	1.81	61	76	74	80	57	49.26	23.59	0	76	69
Written picture description	0	14	84	14	13.62	0.92	9	14	12	84	5	5.39	4.97	0	14	73

Note. Pess = pessimum; Opt = optimum; *N* = number of persons; Med = median; *M* = mean; SD = standard deviation; Min = minimum; Max = maximum; C = 95 % cutoff; *n*<C = number of PWA below cutoff.

* There is no theoretical optimum for Naming, since the domain contains the subtest for Word fluency; the value is set to the maximum obtained among PWA.

Table 7. T-scores for Subtests, Control Group and PWA.

	Control							Aphasia						
Subtest	*N*	Med	*M*	SD	Min	Max	C	*N*	Med	*M*	SD	Min	Max	n<C
Line bisection	83	53	54.63	8.09	35	66	43	82	50.5	50.05	9.63	25	66	14
Semantic association	84	56	54.43	3.87	45	56	45	85	56	49.60	8.25	25	56	18
Recognition memory	84	56	54.27	3.79	42	56	47	85	56	49.26	8.11	25	56	81
Gesture object use	84	60	56.90	5.13	42	60	46	85	51	49.54	8.80	29	60	22
Arithmetic	84	61	56.81	4.92	47	61	47	85	53	49.71	9.07	30	61	22
Auditory comprehension: words	84	57	55.85	2.75	46	57	50	85	50	49.38	8.52	25	57	25
Auditory comprehension: sentences	84	68	65.37	3.50	56	68	60	85	50	49.95	9.81	25	68	38
Auditory comprehension: stories	84	58	56.83	3.84	36	58	48	85	48	49.86	8.58	27	58	52
Reading comprehension: words	84	60	59.61	1.60	51	60	54	85	51	49.56	9.09	25	60	68
Reading comprehension: sentences	84	75	71.21	5.03	56	75	62	85	50	49.98	9.92	25	75	74
Repetition: words	84	59	58.51	2.22	47	59	59	85	51	49.61	8.96	25	59	22
Repetition: complex words	84	55	55	0	55	55	55	84	55	49.51	7.77	33	55	54
Repetition: nonwords	84	62	60.07	4.51	38	62	53	84	50	49.89	9.20	32	62	31
Repetition: digit lists	84	60	61.58	5.92	47	67	47	84	47	49.88	9.34	29	67	44
Repetition: sentences	84	58	57.92	0.76	51	58	58	84	47	49.56	8.17	37	58	22
Naming: objects	84	73	69.36	5.20	53	73	58	84	50	50.04	9.89	27	73	49
Naming: actions	84	60	59.24	3.10	43	60	49	84	49	49.58	9.17	27	60	66
Word fluency	84	75	73.85	2.95	62	75	67	85	50	50.00	9.94	29	75	32
Picture description: spoken	84	75	72.64	4.90	58	75	65	82	51	50.11	9.89	25	75	71
Reading: words	84	61	60.88	0.77	56	61	61	84	50	49.64	9.20	25	61	62
Reading: complex words	84	59	58.89	0.98	50	59	59	84	50	49.73	8.34	38	59	54
Reading: function words	84	54	54	0	54	54	54	84	54	49.29	7.24	35	54	27
Reading: nonwords	84	61	60.71	1.49	53	61	61	84	49.5	50.00	8.74	36	61	61
Copying	84	57	56.71	1.94	41	57	57	81	49	49.51	8.51	30	57	41
Written naming	84	67	64.23	4.25	52	67	56	84	50	50.12	9.21	37	67	58
Dictation	84	67	66.36	1.87	61	67	61	83	50	49.93	9.44	35	67	67
Picture description: written	84	69	67.82	2.86	54	69	63	84	50	50.20	8.94	40	69	73

Note. *N* = number of persons; Med = median; *M* = mean; SD = standard deviation; Min = minimum; Max = maximum; C = 95 % cutoff; *n*<C = number of PWA below cutoff.

optimum scores for form of 9 and 6, respectively. Figure 1 displays an English translation of the scoring sheet for Oral picture description.

Fluency is scored subjectively based on speed, pausing, hesitation, self-correction and whether contributions by the SLT are needed. Further elaboration and example scoring sheets for both Written descriptions and transcribed Oral descriptions are provided in the manual (Swinburn et al., 2021).

Originally, we wanted to weight content and form scores to make each contribute 50 % of the subtest score. Two mathematical issues turned out to give rise to some unwanted properties in such weighted scores. First, raw scores are converted to T-scores using a rank-based algorithm (see above), which, like all ranks, will magnify smaller differences and diminish larger differences. Since the optimum values for content and form are of similar size, weighted sums scaled up to for example 100 formed clusters of *almost* identical values, separated by substantial value gaps. Converting these clustered values into ranks gave undue importance to the small within-cluster differences and thus altered the characteristic properties of the distribution. Second, weighted sums *without* the scaling up resulted in values which were almost but not quite identical to the unweighted sums; the divergences were small particularly for the written descriptions. Thus, the resulting weighted scores could be confusing to the person administering the CAT-N, and the effect of the weighting would be minute.

For these reasons, both picture descriptions are scored by simple addition of the items, like the majority of the subtests. Hence, the optimum scores for the oral and written picture descriptions are 17 and 14, respectively. This has the consequence that the relative contribution of the formal aspects is somewhat greater in the oral descriptions than in the written.

Reordering of test items in subtests with abortion rule

The administering of subtests 7-10 and 17 involves an abortion rule. If a test person fails to obtain a positive score for 4 consecutive items (8 items for subtest 17), the subtest should be aborted in order to spare the person the unnecessary strain of being confronted with test items the person is not capable of answering. For this procedure to function in an optimal fashion, the test items need to be sorted in ascending order of difficulty. In the norming procedure, therefore, the PWA were given all test items for these subtests, and the results were then used to sort the items for each subtest according to the actual scores. Subsequently, the norming scores were calculated as if the abortion rule had been in effect, i.e. disregarding for each PWA any scores given following a 4-item stretch of scores of zero. In practice, the effect on the final scores of reordering the subtests was minimal for most of the subtests, but for subtest 10, the procedure resulted in reduced scores for 9 % of the PWA, possibly indicating that the resulting scores of these 8 individuals did not fully reflect their competence, which hence could also be the case for some individuals when the test is used clinically. In the published CAT-N, the test items are ordered according to difficulty.

Comparing Word fluency and Object naming

Persons with executive difficulties may experience difficulties in word fluency, even if they do not have aphasia (Amunts et al., 2020, 2021; Swinburn et al., 2021, p. 10). Hence,

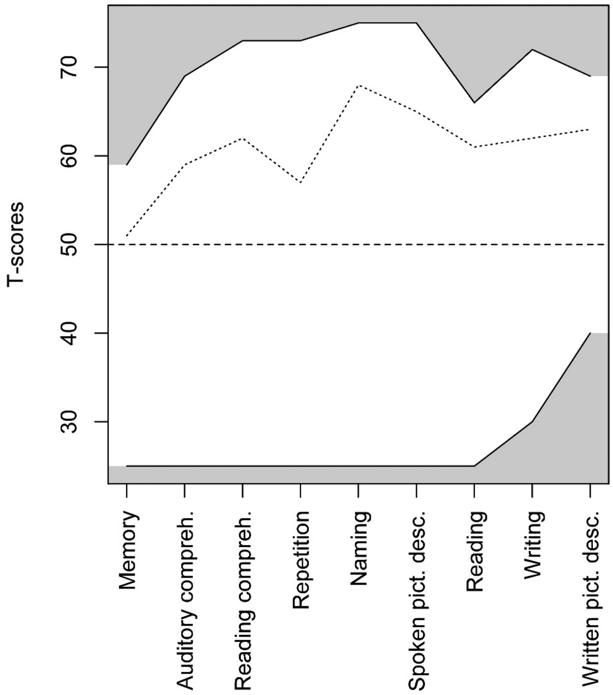


Figure 2. Profile Diagram for the 9 Domains. *Note*. The profile diagram indicates optimum and pessimum T-scores for the domains (solid lines), the 95 % cutoff values (dotted line) and the theoretical mean and median of 50 (dashed line). The idea and design of the diagram is taken from Swinburn et al. (2004, p. 102).

considerably lower scores for Word fluency than for Object naming may indicate a need for further evaluation of executive functions.

The T-scores of both these subtests extend over almost the entire theoretical range from 25 to 75 (Word fluency: 29-75, Object naming: 27-73), and they are both close to normally distributed (W≈0.992, p≈0.88, γ_1≈0.036; W≈0.994, p≈0.96, γ_1≈−0.021). We therefore calculated the divergence between the subtests by simple subtraction of the T-scores of Object naming from the T-scores of Word fluency, and subsequently transformed the resulting difference values into T-scores. The difference values were close to normally distributed (W≈0.982, p≈0.29, γ_1≈−0.233) and yielded T-scores very close to normality (W≈0.995, p≈0.99, γ_1≈−0.017).

Cutoff scores

We used the control group to establish a cutoff score for normal language performance (Ivanova & Hallowell, 2013, p. 906), using the same procedure for both subtests and domains. A cutoff score must be a compromise between sensitivity and specificity, i.e. a balance between false negatives and false positives. In line with the choice of Swinburn et al. (2004, p. 101), we defined the cutoff score as the highest score which includes at least 95 % of the control group of people without aphasia. Since 80 of 84 is 95.2 %, a cutoff score thus defined will yield 4 (or fewer) false positives from the control group sample for the CAT-N. The number of false negatives will vary between subtests and between domains (see Table 13 and Figure 2).

Domain means

We calculated an overall mean T-score for the PWA as the mean of the T-scores for the 8 language domains. Swinburn et al. (2004, p. 46) caution that such mean values may not be reliable if more than 2 missing values are involved; of the 84 PWA in the normative sample who completed the test set in the CAT-N, 4 had missing values for 1 domain, and 1 had missing values for 2 domains. These were included in the calculation of the overall mean T-scores. None had more than 2 missing values.

The domain means are decimal numbers and determining a specific cutoff value for these is hence less straightforward than for the integer values of the individual domains. Any value between the 4th and the 5th individual in the control group would include “at least 95 %” of the control group. We decided to prioritise sensitivity over specificity and chose as the cutoff the value of the person of rank 5 in the CG, which also follows literally the definition of the highest value including at least 95 % of the CG.

Results

Descriptive statistics for subtests and domains

Statistics for raw scores and T-scores for subtests and domains are shown in Tables 5-8.

There is a slight negative correlation between age of the PWA and the domain means, $r_s \approx -0.24$, 95 % CI [−0.04, −0.44]. There is no difference in domain means between genders, although a possible gender effect might be masked by the interaction between gender and education level. There is, however, no effect of education level on domain means, and hence there is no reason to believe that the differences between the PWA and the control

Table 8. T-scores for Domains, Control Group and PWA.

	Control							Aphasia						
Domain	*N*	Med	*M*	SD	Min	Max	C	*N*	Med	*M*	SD	Min	Max	*n*<C
Memory	84	59	56.52	4.25	43	59	51	85	51	49.47	8.87	25	59	38
Auditory comprehension	84	69	65.44	4.01	57	69	59	85	50	50.04	9.79	25	69	69
Reading comprehension	84	73	70.85	3.72	56	73	62	85	50	50.08	9.82	25	73	73
Repetition	84	66	66.81	6.08	49	73	57	84	50.5	50.04	10.07	25	73	62
Naming	84	75	73.76	2.56	66	75	68	84	50	50.02	10.00	25	75	81
Spoken picture description	84	75	72.64	4.90	58	75	65	82	51	50.11	9.89	25	75	71
Reading	84	66	65.50	1.85	58	66	61	84	50	49.77	9.63	25	66	71
Writing	84	72	68.86	4.30	52	72	62	80	50	50.06	9.88	30	72	69
Written picture description	84	69	67.82	2.86	54	69	63	84	50	50.20	8.94	40	69	73
Mean of 8 linguistic domains	84	69.25	68.96	1.90	62.125	68.96	65.75	84	49.94	50.10	7.84	31.875	65.14	84

Note. N = number of persons; Med = median; *M* = mean; SD = standard deviation; Min = minimum; Max = maximum; C = 95 % cutoff; *n*<C = number of PWA below cutoff.

Table 9. Correlations Between Domains.

Domains	Mem	AC	RC	Rep	Nam	Pd:sp	Read	Writ	Pd:wr
Memory	1	0.30	0.28	0.08	0.30	0.24	0.14	0.29	0.34
Auditory comprehension	0.30	1	0.80	0.48	0.59	0.54	0.66	0.71	0.49
Reading comprehension	0.28	0.80	1	0.42	0.69	0.60	0.76	0.67	0.55
Repetition	0.08	0.48	0.42	1	0.55	0.54	0.53	0.55	0.43
Naming	0.30	0.59	0.69	0.55	1	0.75	0.74	0.69	0.65
Picture description: Spoken	0.24	0.54	0.60	0.54	0.75	1	0.64	0.67	0.67
Reading	0.14	0.66	0.76	0.53	0.74	0.64	1	0.68	0.54
Writing	0.29	0.71	0.67	0.55	0.69	0.67	0.68	1	0.81
Picture description: Written	0.34	0.49	0.55	0.43	0.65	0.67	0.54	0.81	1

Note. Mem = memory; AC = auditory comprehension; RC = reading comprehension; Rep = repetition; Nam = naming; Pd:sp = picture description: spoken; Read = reading; Writ = writing; Pd:wr = picture description: written. Values are Spearman's correlation coefficient r_s.

group in distribution of gender and education levels shown in Table 2 have affected the results, e.g. in terms of cutoff values. 5 PWA have an L1 other than Norwegian; they do not deviate systematically from the native speakers. The participants were grouped into 4 dialect areas; there were no effects of dialect. There is no effect of time from onset on the domain means.

Relationships between domains

Table 9 shows r_s between all the domains. Correlations are medium to strong between all the linguistic domains, the lowest value of r_s being 0.42 (between Reading comprehension and Repetition). The majority of the pairwise differences between these coefficient values are not relevant, however, as most of the dispersion between them lies within 95 % confidence intervals (not shown here).

Note the high correlations between Written picture description and Writing (0.81), between Spoken picture description and Naming (0.75), between the two Comprehension domains (0.80) and between the two Reading domains (0.76). These high correlation coefficients correspond to the results from the factor analysis shown below.

Also interesting is the fairly strong positive correlations between *all* of the 8 linguistic domains. Cronbach's alpha for the 8 linguistic domains is 0.92, indicating a strong relationship between the different domains in terms of loss of abilities.

Finally, we calculated the mean correlation coefficient r_s between each linguistic subtest and the other linguistic subtests and found that the mean correlation with the rest of the subtests was weakest for Auditory comprehension of stories (0.28), Repetition of complex words (0.35) and Repetition of nonwords (0.35). However, all these subtests consist of a small number of items (4, 3 and 5, respectively) and their scores will thus vary arbitrarily to a greater degree than other subtests. Of the subtests with a more reliable number of items, Copying has the lowest mean correlation with the other subtests (0.40); as shown below, Copying also is the subtest with the lowest internal consistency.

Table 10. Internal Consistency of the 9 Domains and the 22 Subtests in the Language Battery.

Domain	# subtests	α
Memory	2	0.41
Auditory comprehension	3	0.79
Reading comprehension	2	0.82
Repetition	5	0.90
Naming	3	0.84
Reading	4	0.92
Writing	3	0.85
Subtest	**# items**	**α**
Comprehension of spoken words	15	0.71
Comprehension of spoken sentences	15	0.83
Comprehension of spoken stories	4	0.70
Comprehension of written words	15	0.89
Comprehension of written sentences	15	0.91
Repetition of words	16	0.91
Repetition of complex words	3	0.80
Repetition of nonwords	5	0.70
Repetition of digit strings	6	0.76
Repetition of sentences	4	0.86
Naming objects	24	0.87
Naming actions	5	0.71
Word fluency	2	0.75
Spoken picture description	7	0.88
Reading words	24	0.97
Reading complex words	3	0.82
Reading function words	3	0.90
Reading nonwords	5	0.85
Writing: copying	3	0.59
Writing picture names	5	0.86
Writing to dictation	5	0.86
Written picture description	6	0.92

Note. α = Cronbach's alpha; # subtests = number of subtests in domain; # items = number of test items in subtest. Domain values are based on T-scores from subtests; subtest values are based on raw scores of test items. Spoken picture description and Written picture description are domains as well as subtests.

Reliability

Reliability concerns the consistency, stability and accuracy of a test (Ivanova & Hallowell, 2013, p. 907). This section discusses the internal reliability, test-retest reliability, inter-rater reliability and intra-rater reliability of the CAT-N.

Internal consistency

Table 10 indicates the internal consistency of the subtests in the language battery, showing values of Cronbach's alpha for each subtest, based on the raw scores of each test item. The only subtest showing "unacceptably" low internal consistency is Copying (α≈0.59). All the other subtests have "sufficiently" high values, α≈0.70 or above; most of them with values interpreted by Cohen as "highly" or "very highly" reliable. In general, one expects lower consistency values in subtests comprising few items (Ivanova & Hallowell, 2013, p. 907) and Table 10 demonstrates that several of the subtests with α<0.80 have 5 or fewer items.

Table 10 further indicates the internal consistency of the domains, showing values of Cronbach's alpha for each domain, based on the T-scores of the subtests.

As shown, the Memory domain displays an "unacceptably low" degree of internal consistency, indicating that the Memory domain does not measure one consistent entity, but rather two quite different capacities. The correlation between the two subtests is only r_s≈0.24, 95 % CI [0.024, 0.43], indicating only quite a weak correlation between them, if any. Swinburn et al. (2004, p. 115) found that these two subtests in the CAT-EN clustered "closely" in a hierarchical cluster analysis.

All the other domains display coefficients in the vicinity of the range interpreted as "highly reliable" by Cohen (1992). This is an indication that the correlation between the subtests within each domain is substantial enough for the accumulation of the scores into domain scores to be meaningful. At the same time, the fact that the internal correlation is not perfect indicates that the various subtests do tap into slightly different capacities, as they should, although differences may also be caused by arbitrary variation, especially for subtests with few test items. Among the subtest pairs with lowest correlation are Naming of objects and actions, r_s≈0.51, Repetition of nonwords and sentences, r_s≈0.47, and Auditory comprehension of words and stories, r_s≈0.46. On the other hand, some individual subtests correlate really strongly with their domain, indicating that these subtests alone could function as simplified domains; examples are Auditory comprehension of sentences, r_s≈0.97, Reading comprehension of sentences, r_s≈0.96, Reading of words, r_s ≈0.96, and Word fluency, r_s≈0.93. A large number of items and high dispersion are two (partially related) causes of such strong correlations between a single subtest and its domain; also, a small number of subtests in the domain will tend to cause strong correlation with (at least one of) the subtests.

Test-retest reliability

Collecting data for test-retest reliability analysis proved too demanding within the project, and we have not carried out any such analysis for the CAT-N. Most of the battery has very similar characteristics to the CAT for English, and we refer the reader to Swinburn et al. (2004. pp. 108-109) and Howard et al. (2010, pp. 66-67).

Inter-rater reliability

Swinburn et al. (2004) tested inter-rater reliability (IRR) for the CAT-EN and found "excellent agreement for almost all of the tests" (p. 111). We decided to carry out IRR analyses for the two Picture description subtests, since these have a design for scoring not previously employed (see above) and perhaps rely on a more subjective scoring than the other subtests.

Table 11. Inter-rater Reliability for Picture Descriptions.

Group	Spoken	Written
Control	0.93	0.93
Aphasia	0.81	0.91

Note. Values of Krippendorff's alpha.

Table 12. Intra-rater Reliability for Picture Descriptions.

Group	Spoken	Written
Control	0.97	0.97
Aphasia	0.84	0.91

Note. Values are mean ratios of correspondence for three raters.

All entries were scored by an SLT, and three other persons (two SLTs and one linguist) scored 20 entries each from each group. Values of Krippendorff's alpha (Table 11) show that inter-rater reliability is acceptable (Krippendorff, 2004, p. 241) for both groups and both subtests, although weaker for the spoken descriptions by PWA than for the other three group/subtest combinations.

Investigating the individual test items, alpha values exceed 0.86 for all written test items for both PWA and control group, whereas values for the oral test items display more dispersion, between 0.49 (The child alerts) and 0.96 (The books fall) for PWA, and between 0.60 (The child alerts) and 0.94 (The books fall) for the control group. Additionally, alpha is less than 0.667 for Fluency (PWA) and for Grammatical complexity (CG), indicating less than acceptable reliability for these items (see above). It is worth noting that even though these individual items seem intrinsically difficult to score, the scores of the entire subtests appear to be reliable.

For the rest of the test battery, we refer to Swinburn et al. (2004, pp. 108-109).

Intra-rater reliability

We evaluated intra-rater reliability for the two picture descriptions, both for PWA and for the CG. Three raters rated different samples of n=20 twice with an interval of four months in between. The reliability was calculated as the mean ratios of correspondence for the three. According to Mackenzie et al. (2007) and Nicholas and Brookshire (1995), a ratio of correspondence above 0.8 is considered acceptable. Table 12 shows the mean ratios of correspondence.

Like for inter-rater reliability, the correspondence is weaker for PWA than for the control group, especially for the spoken subtest, again indicating that these items are more difficult to score consistently. Looking at the individual test items (not shown here), all items have mean ratios of 0.8 or higher, indicating acceptable correspondence.

Validity

The validity of a test concerns the extent to which it measures what it is intended to measure and hence whether the test should be used as a foundation on which to base conclusions (Ivanova & Hallowell, 2013, p. 908). This section focuses on the sensitivity and specificity of the CAT-N, its concurrent validity and its construct validity.

Table 13. Comparison of T-scores Between the Control Group and PWA.

Subtest/domain	N_c	N_a	W	p(W)	z	r_{pb}	ppn<C
Line bisection	83	82	4330	0.0021	3.071	0.239	0.17
Semantic association	84	85	4680	0.0000	4.405	0.339	0.21
Recognition memory	84	85	4791.5	0.0000	4.608	0.354	0.26
Memory	**84**	**85**	**5257.5**	**0.0000**	**5.811**	**0.447**	**0.45**
Gesture object use	84	85	5300	0.0000	5.885	0.453	0.26
Arithmetic	84	85	5209	0.0000	5.427	0.417	0.29
Auditory comprehension: words	84	85	5134	0.0000	5.813	0.447	0.45
Auditory comprehension: sentences	84	85	6601.5	0.0000	9.739	0.749	0.80
Auditory comprehension: stories	84	85	5247	0.0000	6.363	0.489	0.26
Auditory comprehension	**84**	**85**	**6590.5**	**0.0000**	**9.626**	**0.740**	**0.81**
Reading comprehension: words	84	85	5955.5	0.0000	8.607	0.662	0.61
Reading comprehension: sentences	84	85	6914	0.0000	10.640	0.818	0.87
Reading comprehension	**84**	**85**	**6947.5**	**0.0000**	**10.901**	**0.839**	**0.86**
Repetition: words	84	85	5709.5	0.0000	7.946	0.611	0.64
Repetition: complex words	84	84	4830	0.0000	6.105	0.471	0.37
Repetition: nonwords	84	84	5801.5	0.0000	7.808	0.602	0.52
Repetition: digit lists	84	84	5962.5	0.0000	7.911	0.610	0.26
Repetition: sentences	84	84	5566	0.0000	8.016	0.618	0.58
Repetition	**84**	**84**	**6482.5**	**0.0000**	**9.447**	**0.729**	**0.74**
Naming: objects	84	84	6724.5	0.0000	10.273	0.793	0.79
Naming: actions	84	84	5701	0.0000	8.014	0.618	0.38
Word fluency	84	85	7051	0.0000	11.338	0.872	0.95
Naming	**84**	**84**	**6977**	**0.0000**	**11.272**	**0.870**	**0.96**
Picture description: spoken	**84**	**82**	**6702.5**	**0.0000**	**10.932**	**0.848**	**0.87**
Reading: words	84	84	6106	0.0000	9.363	0.722	0.74
Reading: complex words	84	84	5768	0.0000	8.536	0.659	0.64
Reading: function words	84	84	4662	0.0000	5.626	0.434	0.32
Reading: nonwords	84	84	6024	0.0000	9.074	0.700	0.73
Reading	**84**	**84**	**6615**	**0.0000**	**10.555**	**0.814**	**0.85**
Copying	84	81	5054.5	0.0000	6.977	0.543	0.51
Written naming	84	84	6361	0.0000	9.289	0.717	0.69
Dictation	84	83	6606	0.0000	10.645	0.824	0.81
Writing	**84**	**80**	**6429.5**	**0.0000**	**10.290**	**0.804**	**0.86**
Picture description: written	**84**	**84**	**6735**	**0.0000**	**10.646**	**0.821**	**0.87**

Note. N_c = number of controls; N_a = number of PWA; W = statistic from Wilcoxon test; p = p-value from Wilcoxon test; z = z-value from Wilcoxon test; r_{pb} = point-biserial correlation coefficient; ppn<C = proportion of PWA below the 95 % cutoff of the control group. Domains are in bold.

Sensitivity and specificity

We have investigated how well the CAT-N discriminates between PWA and persons without known aphasia through comparison of PWA and CG, domain means and domain cutoffs.

Comparison of PWA and control group

We compared subtest T-scores and domain T-scores between the PWA and the control group. Table 13 shows the results of the comparisons and the effect sizes. Since multiple significance tests have been carried out, the p-values should be compared to α adjusted for FWER: α_{27}=0.0019 for the subtests and α_{9}=0.0068 for the domains, corresponding to α_{1} =0.05 without correction.

Table 13 shows that the p-value for Line bisection is just over the FWER-adjusted α_{27}, and only 17 % of PWA fall below the 95 % cutoff for this subtest. All other subtests and all domains distinguish well between PWA and the CG. Moreover, all linguistic domains demonstrate very strong effects: r_{pb} between 0.73 and 0.87. In terms of sensitivity, the cutoff for each linguistic domain correctly points out at least 74 % of PWA; 6 of them point out 85 % or more. The Word fluency subtest alone has a sensitivity of 95 % (with specificity set at 95 %). This suggests that the Word fluency subtest would function as a quick rough indicator of aphasia on its own.

Domain means

The cutoff value for the domain means (65.75) yields the maximum sensitivity (100 % of the PWA in the sample) and an acceptable specificity (95.2 % of the control group sample, per definition).

Domain cutoff values

Swinburn et al. (2004, p. 120) suggest using the cutoff values of the individual linguistic domains as discrimination criterion; more than 1 of 8 domain T-scores below the cutoff value of the domain indicates aphasia. In the CAT-N, this procedure correctly identifies 95.2 % (80 of 84) of persons without aphasia and 98.8 % (83 of 84) of PWA. In addition, this simpler procedure obviates the need for mean values or other calculations which may lead to mistakes. Interestingly, the two procedures do not yield the same false positives from the CG, indicating that accuracy could potentially be improved by combining procedures. However, with no available method for validation, such combinations of methods may result in overfitting and prove without merit when applied clinically or more generally to a new sample.

Concurrent validity

A subsample of n=34 of the PWA in the normative sample were also tested with *Norwegian Basic Aphasia Assessment* (NBAA; Reinvang & Engvik, 1980; see also Sundet & Engvik, 1985) and we calculated the correlation between the relevant domains of the CAT-N and subtests of NBAA, and between the domain mean of the CAT-N and the overall aphasia coefficient of the NBAA (Table 14). Not all PWA completed all subtests of the NBAA, resulting in variation in sample size (n).

We compared the CAT scores for the subsample of 34 PWA with the remaining 50 in the full sample and found no differences. The subsample can therefore be considered representative of the full sample.

Table 14. Correlations Between the CAT-N Domains and Subtests of NBAA.

Domain	n	r_s	95 % CI
Auditory comprehension	32	0.41	0.02-0.69
Reading comprehension	24	0.73	0.47-0.88
Repetition	30	0.87	0.69-0.95
Naming	29	0.66	0.33-0.88
Reading	24	0.79	0.53-0.91
Writing	24	0.57	0.11-0.91
Domain mean/Aphasia coefficient	22	0.71	0.42-0.87

Note. n = number of persons; r_s = Spearman's correlation coefficient; CI = confidence interval. Comparisons were made on a subsample of n = 34 of the CAT-N normative sample.

Table 15. Factor Loadings Resulting From Exploratory Factor Analysis.

Subtest	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Auditory comprehension: words	0.177	0.223	0.130	**0.654**	0.195	0.029
Auditory comprehension: sentences	0.347	0.174	0.242	**0.707**	0.147	0.210
Auditory comprehension: stories	0.174	0.024	0.229	**0.575**	−0.038	0.020
Reading comprehension: words	**0.524**	0.049	0.223	**0.514**	0.291	−0.130
Reading comprehension: sentences	**0.486**	0.067	0.203	**0.624**	0.288	0.109
Repetition: words	0.099	**0.912**	0.073	0.153	0.206	0.078
Repetition: complex words	0.181	**0.800**	0.171	−0.055	0.006	0.259
Repetition: nonwords	0.051	**0.747**	0.123	0.180	0.049	−0.010
Repetition: digit lists	0.148	**0.519**	0.128	0.366	0.081	**0.557**
Repetition: sentences	0.225	**0.609**	0.193	0.021	0.159	**0.605**
Naming: objects	0.410	0.355	0.279	0.229	**0.668**	0.058
Naming: actions	0.212	**0.603**	0.238	0.102	**0.487**	−0.065
Word fluency	0.306	0.088	0.470	0.258	**0.521**	0.103
Picture description: spoken	0.323	0.231	0.439	0.187	**0.492**	0.250
Reading: words	**0.728**	0.262	0.140	0.335	0.317	0.043
Reading: complex words	**0.679**	0.194	0.213	0.289	0.283	0.096
Reading: function words	**0.795**	0.085	0.204	0.295	0.107	0.093
Reading: nonwords	**0.809**	0.163	0.279	0.158	0.048	0.110
Copying	0.321	−0.009	**0.542**	0.379	0.165	−0.235
Written naming	0.312	0.271	**0.712**	0.258	0.197	0.080
Dictation	0.170	0.310	**0.772**	0.351	−0.014	0.238
Picture description: written	0.199	0.160	**0.778**	0.144	0.298	0.055

Note. Factor loadings > 0.475 are in bold.

All correlation coefficients are positive, as expected, the weakest correlation being for auditory comprehension (0.41). The two tests differ in number and types of test items and methods, especially for auditory comprehension, which may explain the fairly weak correlation for this domain. All the other domains have strong correlations, but the wide confidence intervals demonstrate the level of uncertainty; for Writing, the lower end of the confidence interval is as low as 0.11.

Construct validity

The construct validity of a test concerns the extent to which it is consistent with the underlying theoretical understanding of its object of study. An important aspect of the construct validity of the CAT-N is the construction of its domains, which we attempted to validate by performing an exploratory factor analysis on the T-scores from the 22 linguistic subtests. The minimal adequate model explains 74 % of the variance and has 6 factors as opposed to the 8 pre-defined domains of the test. Table 15 shows factor loadings for the 22 variables on the 6 factors.

Stevens (2009, p. 333) recommends considering only variables with factor loadings about 0.4 or greater for interpretation purposes. We chose a somewhat higher threshold in order to minimise the number of subtests contributing to more than one factor; only 3 subtest/factor combinations were affected by this. The salient subtests (> 0.475) are highlighted in the table, showing that 5 of the subtests belong to 2 domains, but that the great majority of subtests uniquely contribute to one domain only, given the chosen threshold.

With the exception of Action naming clustering with Repetition, all factors in the analysis are conceptually coherent, even though they do not correspond fully to the pre-

Table 16. Internal Consistency in the 5 Linguistic Domains Resulting From the Factor Analysis.

Factor no	Factor label	Cronbach's α
1	Reading (decoding and comprehension)	0.92
2	(Spoken) repetition	0.89
3	Written production	0.89
4	Comprehension (auditory and reading)	0.87
5	Spoken production	0.88

defined domains. Factor 1 collects Reading (decoding) and Reading comprehension into one common Reading domain. Factor 2 corresponds to the existing Repetition domain, with the exception of Action naming, the inclusion of which we are not able to explain. Factor 3 collects Writing and Written picture description into one common domain of Written production. Factor 4 collects Auditory and Reading comprehension into one common domain of Comprehension. Factor 5 collects (spoken) Naming and Spoken picture description into one common domain of Spoken production. All these 5 factors constitute conceptually coherent entities. Last, and least important in terms of the contribution of the factor to the explanation of the total variation, Factor 6 consists of those two tasks of Repetition involving sequences. We see no conceptually substantial reason to distinguish between different types of Repetition, and as these two tasks are also included in Factor 2, we disregard Factor 6 in the following, even though it is part of the minimal adequate model.

Table 16 shows values of Cronbach's alpha for these 5 alternative domains, of which none correspond fully to the pre-defined domains. The table demonstrates that the values of Cronbach's α are all 0.87 or above and generally a bit higher than for the pre-defined domains (shown in Table 10); this is hardly surprising, given that the aim of a factor analysis is to find the "best" clustering of variables into factors. Merging the two picture descriptions into Spoken and Written production, respectively, seems conceptually sensible and reduces the number of domains by 2. The fact that Reading comprehension clusters with both (general) Comprehension and Reading distinguishes it from Auditory comprehension and indirectly also from the subtests of Reading related to decoding.

It is important to realise the limitations of a factor analysis. The factor analysis is set to discover a certain number of factors and rotates the multidimensional space with the aim of letting the factors emerge. This is a very powerful tool which tends to exaggerate the salience of the factors and understate the relationships between the variables which are not found to be part of the same factor. It is important to realise that *all* the linguistic subtests correlate positively in the normative sample, although a few of the correlation coefficients are close to zero and their confidence intervals contain zero. Also, some of the subtests which are *not* found to be part of the same factor correlate strongly – in some cases more strongly than some of the correlations within a factor.

Concluding remarks

We have presented the Norwegian version of the CAT, with emphasis on its psychometric properties. The CAT-N is a general aphasia test which examines linguistic skills broadly, within the cognitive neuropsychological tradition. We have shown that it distinguishes well between PWA and persons without aphasia and we have documented issues of reliability and validity.

As mentioned in the introduction, the development of CAT-N is part of an effort to develop standardised and comparable aphasia tests for different languages. Numerous factors which are difficult to control for may affect test attributes, test results and test statistics, such as demographic differences in populations, morphological and orthographic differences between languages, and arbitrary differences in sampling of test persons or in the difficulty of subtests or individual test items. As CAT is now adapted for several languages, cross-linguistic analysis of test properties becomes feasible and will be a natural next step within this co-operative initiative. So far, a comparison of CAT-N and the Croatian version of CAT has been carried out (Matić Škorić et al., submitted), finding good general correspondence, but also some differences in individual subtests and one domain. We welcome more contrastive studies and believe transparency and thorough documentation to be key issues in that respect.

We have two observations concerning the current CAT construct. First, the Semantic association subtest (called "Semantic memory" in the CAT-EN) "assesses access to semantic memory" (Swinburn et al., 2004, p. 15), whereas Recognition memory assesses non-verbal, visual recognition memory. In the original CAT-EN, the two correlate (Swinburn et al., 2004, p. 46) and are combined into a common Memory domain. In the CAT-N, this correlation is weak and the internal consistency of the Memory domain is "unacceptably" low. Hence, the analyses from the CAT-N do not confirm a common construct of memory and indicate that Semantic association and Recognition memory are two fairly unrelated capacities.

Second, the factor analysis we carried out to investigate the construct validity of the CAT-N indicates a simpler construct than the 8-domain model of the CAT. It is worth noting that the simpler model comprising only 5 domains would seem to match the traditional view of fundamental linguistic modalities (listening, speaking, reading, writing) better than the original 8-domain model. On the other hand, the inclusion of some of the subtests in more than one domain complicates the model and thereby the construct. Our analyses demonstrate somewhat improved internal consistency, and it would be interesting to assess it in terms of sensitivity and specificity. The factorial model for CAT-N has more factors than the factorial models for both the English and the Hungarian versions of the CAT (Swinburn et al., 2004, p. 117; Zakariás & Lukács, 2022, p. 1139), both of which have three. Comparing factorial models of different data is inherently difficult, however, especially as the method used by Zakariás and Lukács deviates from ours in several aspects and delimiting criteria are not documented by Swinburn et al.

Finally, an important aim in writing this article has been to make explicit and unambiguous some of the design decisions we have made in constructing the CAT-N, such as the scoring scheme for Picture descriptions, the reordering of items in the subtests with abortion rules, the transformation of raw scores into T-scores, the calculation of divergence in T-scores between Word fluency and Object naming, and the determination of cutoff value for the domain means. It is our hope that this documentation will be helpful in future adaptations of the CAT.

Endnotes

1 Even though uncorrelated underlying factors could not be expected in this case, the orthogonal varimax rotation was chosen in order to accommodate comparison with the English version of the CAT (Swinburn et al., 2004) and simplify interpretation of the model (e.g. Stevens, 2009. p. 331).

2 The Word fluency subtest is included among the cognitive subtests, but contributes to the calculations of the linguistic domains.

3 We use the term *domain* where Swinburn et al. (2004) use *modality* in order to avoid confusion with what is commonly known as the fundamental linguistic modalities (listening, speaking, reading, writing).

4 In R, T-scores may be calculated according to this algorithm using the following command:

```
qnorm((rank(x, na.last="keep")-0.5)/length(na.omit(x)), m=50, sd=10)
```

where x is a vector variable holding all raw scores for a subtest or domain, i.e. the 85 PWA scores in this case. In SPSS, there is a ready-made menu option for Rankit transformation.

Geolocation information

Norway

Acknowledgements

The authors thank the Collaboration of Aphasia Trialists funded by COST and the Tavistock Trust for Aphasia for their support in setting research priorities and promoting international/multidisciplinary collaboration in aphasia research. We furthermore thank the other members of the Norwegian adaptation team, Nina Høeg and Ingvild Røste.

Disclosure statement

This work was partly supported by the Research Council of Norway through its Centres of Excellence funding scheme [223265] and by Statped.

The authors report no conflict of interest.

Funding

This work was supported by the Norges Forskningsråd and by Statped [223265].

ORCID

Bård Uri Jensen http://orcid.org/0009-0000-9106-9051
Monica I. Norvik http://orcid.org/0000-0002-5240-6825
Hanne Gram Simonsen http://orcid.org/0000-0001-6762-5505

References

Abdi, H. (2007). The Bonferroni and Šidák corrections for multiple comparisons. In N. Salkind (Eds.), *Encyclopedia of measurement and statistics*. Sage.

Amunts, J., Camilleri, J. A., Eickhoff, S. B., Heim, S., & Weis, S. (2020). Executive functions predict verbal fluency scores in healthy participants. *Scientific Reports*, *10*(1), 1–11.

Amunts, J., Camilleri, J. A., Eickhoff, S. B., Patil, K. R., Heim, S., von Polier, G. G., & Weis, S. (2021). Comprehensive verbal fluency features predict executive function performance. *Scientific Reports*, *11*(1), 1–14.

Bates, E., Wulfeck, B., & MacWhinney, B. (1991). Cross-linguistic research in aphasia: An overview. *Brain and Language*, *41*(2), 123–148. https://doi.org/10.1016/0093-934X(91)90149-U

Bliss, C. I., Greenwood, M. L. & White, E. S. (1956). A rankit analysis of paired comparisons for measuring the effect of sprays on flavor. *Biometrics*, *12*(4), 381–403. https://doi.org/10.2307/3001679

Clark-Carter, D. (2005). T-scores. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 4, p. 2067). Wiley.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, *1*(3), 98–101. https://doi.org/10.1111/1467-8721.ep10768783

Cohen, L., Manion, L., & Morrison, K. (2011). *Research methods in education* (7th ed.). Routledge.

Crawley, M. J. (2007). *The R book*. John Wiley.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cronbach, L. J. & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*(3), 391–418. https://doi.org/10.1177/0013164404266386

Falissard, B. (2012). *psy: Various procedures used in psychometry*. R package version 1.1. https://CRAN.R-project.org/package=psy

Frick, P. J., Barry, C. T. & Kamphaus, R. W. (2010). *Clinical assessment of child and adolescent personality and behavior* (3rd ed.). Springer.

Fyndanis, V., Lind, M., Varlokosta, S., Kambanaros, M., Soroli, E., Ceder, K., Grohmann, K. K., Rofes, A., Simonsen, H. G., Bjekić, J., Gavarró, A., Kraljević, J. K., Martínez-Ferreiro, S., Munarriz, A., Pourquie, M., Vuksanović, J., Zakariás, L. & Howard, D. (2017). Cross-linguistic adaptations of The Comprehensive Aphasia Test: Challenges and solutions. *Clinical Linguistics & Phonetics*, *31*(7–9), 697–710. https://doi.org/10.1080/02699206.2017.1310299

Gamer, M., Lemon, J. & Singh, I. F. P. (2019). *irr: Various coefficients of interrater reliability and agreement*. R package version 0.84.1. https://CRAN.R-project.org/package=irr

Gitterman, M. R., Goral, M. & Obler, L. K. (Eds.). (2012). *Aspects of multilingual aphasia* (Vol. 8). Multilingual Matters.

Guevara, E. R. (2010). NoWaC: A large web-based corpus for Norwegian. *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop* (pp. 1–7). Association for Computational Linguistics.

Hervé, M. (2021). *RVAideMemoire: Testing and plotting procedures for biostatistics*. R package (Version 0.9-80). https://CRAN.R-project.org/package=RVAideMemoire

Howard, D., Swinburn, K. & Porter, G. (2010). Putting the CAT out: What the Comprehensive Aphasia Test has to offer. *Aphasiology*, *24*(1), 56–74. https://doi.org/10.1080/02687030802453202

Ivanova, M. V. & Hallowell, B. (2013). A tutorial on aphasia test development in any language: Key substantive and psychometric considerations. *Aphasiology*, *27*(8), 891–920. https://doi.org/10.1080/02687038.2013.805728

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Sage.

Kuvač Kraljević, J., Matić, A. & Lice, K. (2020). Putting the CAT-HR out: Key properties and specificities. *Aphasiology*, *34*(7), 820–839. https://doi.org/10.1080/02687038.2019.1650160

LeBlanc, V. & Cox, M. A. A. (2017). Interpretation of the point-biserial correlation coefficient in the context of a school examination. *The Quantitative Methods for Psychology*, *13*(1), 46–56. https://doi.org/10.20982/tqmp.13.1.p046

Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis*. John Benjamins.

Lind, M., Simonsen, H. G., Hansen, P., Holm, E. & Mevik, B. H. (2015). Norwegian Words: A lexical database for clinicians and researchers. *Clinical Linguistics & Phonetics*, *29*(4), 276–290. https://doi.org/10.3109/02699206.2014.999952

Mackenzie, C., Brady, M., Norrie, J. & Poedjianto, N. (2007). Picture description in neurologically normal adults: Concepts and topic coherence. *Aphasiology*, *21*(3–4), 340–354. https://doi.org/10.1080/02687030600911419

Matić Škorić, A. M., Norvik, M. I., Kuvač Kraljević, J., Røste, I. & Simonsen, H. G. (submitted). Comprehensive Aphasia Test (CAT): What do Croatian and Norwegian data reveal?

Maviş, İ., Tunçer, A. M., Selvi-Balo, S., Tokaç, S. D. & Özdemir, Ş. (2021). The adaptation process of the Comprehensive Aphasia Test into CAT-Turkish: Psycholinguistic and clinical considerations. *Aphasiology*, *36*(4), 493–512. https://doi.org/10.1080/02687038.2021.1923947

Nicholas, L. E. & Brookshire, R. H. (1995). Presence, completeness, and accuracy of main concepts in the connected speech of non-brain-damaged adults and adults with aphasia. *Journal of Speech, Language, and Hearing Research*, *38*(1), 145–156. https://doi.org/10.1044/jshr.3801.145

Paivio, A., Yuille, J. C. & Madigan, S. A. (1968). Concreteness, imagery and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, *76*, 1–25. https://doi.org/10.1037/h0025327

Papathanasiou, I., Coppens, P. & Davidson, B. (2017). Aphasia and related neurogenic communication disorders: Basic concepts, management, and efficacy. In I. Papathanasiou & P. Coppens (Eds.), *Aphasia and related neurogenic communication disorders* (pp. 3–14). Jones & Bartlett Learning.

Paradis, M., & Libben, G. (1987). *The Assessment of Bilingual Aphasia*. Erlbaum.

Pastizzo, M. J. & Carbone, R. F. (2007). Spoken word frequency counts based on 1.6 million words in American English. *Behavior Research Methods*, *39*(4), 1025–1028. https://doi.org/10.3758/BF03193000

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/.

Reinvang, I. & Engvik, H. (1980). *Norsk grunntest for afasi (NGA)*. Universitetsforlaget.

Royston, P. (1995). Remark AS R94: A remark on Algorithm AS 181: The W-test for normality. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *44*(4), 547–551. https://doi.org/10.2307/2986146

Shapiro, S. S. & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3 and 4), 591–611. https://doi.org/10.2307/2333709

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, *62*(318), 626–633. https://doi.org/10.1080/01621459.1967.10482935

Sijtsma, K., & Pfadt, J. M. (2021). Part II: On the use, the misuse, and the very limited usefulness of Cronbach's alpha: Discussing lower bounds and correlated errors. *Psychometrica*. https://doi.org/10.1007/s11336-021-09789-8

Simonsen, H. G., Lind, M., Hansen, P., Holm, E. & Mevik, B. H. (2013). Imageability of Norwegian nouns, verbs and adjectives in a cross-linguistic perspective. *Clinical Linguistics & Phonetics*, *27*(6–7), 435–446. https://doi.org/10.3109/02699206.2012.752527

Solomon, S. R. & Sawilowsky, S. S. (2009). Impact of rank-based normalizing transformations on the accuracy of test scores. *Journal of Modern Applied Statistical Methods*, *8*(2), 448–462. http://digitalcommons.wayne.edu/coe_tbf/5

Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). Erlbaum.

Sundet, K. & Engvik, H. (1985). The validity of aphasic subtypes. *Scandinavian Journal of Psychology*, *26*(1), 219–226. https://doi.org/10.1111/j.1467-9450.1985.tb01159.x

Swinburn, K., Best, W., Beeke, S., Cruice, M., Smith, L., Pearce Willis, E., Ledingham, K., Sweeney, J. & McVicker, S. J. (2019). A concise patient reported outcome measure for people with aphasia: The Aphasia Impact Questionnaire 21. *Aphasiology*, *33*(9), 1035–1060. https://doi.org/10.1080/02687038.2018.1517406

Swinburn, K., Porter, G. & Howard, D. (2004). *Comprehensive Aphasia Test*. Psychology Press.

Swinburn, K., Porter, G., Howard, D., Høeg, N., Norvik, M. I., Røste, I. & Simonsen, H. G. (2021). *CAT-N – Comprehensive Aphasia Test* [User Manual]. Novus forlag.

Swinburn, K., Porter, G., Howard, D., Kuvač Kraljević, J., Lice, K. & Matić, A. (2020). *Sveobuhvatni test za procjenu afazije CAT-HR*. [Comprehensive Aphasia Test – Croatian]. Naklada Slap.

Visch-Brink, E., Vandenborre, D., de Smet, H. J. & Mariën, P. (2014). *Comprehensive Aphasia Test – Nederlandse bewerking – Handleiding*. Pearson.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83. https://doi.org/10.2307/3001968

Zakariás, L., & Lukács, Á. (2022). The Comprehensive Aphasia Test – Hungarian: Adaptation and psychometric properties. *Aphasiology*, *36*(9), 1127–1145. https://doi.org/10.1080/02687038.2021.1937921

Zimmerman, D. W. & Zumbo, B. D. (1990). The relative power of the Wilcoxon-Mann-Whitney test and Student t test under simple bounded transformations. *Journal of General Psychology*, *117*(4), 425–436. https://doi.org/10.1080/00221309.1990.9921148