The Stroop-Interference-NoGo-Test (STING): A Fast Screening Tool for the Global Assessment of Neuropsychological Impairments

Authors
Bernhard Fehlmann, Hennric Jokeit

Affiliation
Klinik Lengg AG, Institut für Neuropsychologische Diagnostik und Bildgebung

Key words
screening, attention, semantic memory, inhibition, mental speed

Bibliography
DOI http://dx.doi.org/10.1055/s-0043-103267
Neurology International Open 2017; 1: E98–E106
© Georg Thieme Verlag KG Stuttgart · New York
ISSN 2511-1795

Correspondence
Bernhard Fehlmann
Klinik Lengg AG
Institut für Neuropsychologische Diagnostik und Bildgebung
Bleulerstrasse 60
8008 Zürich
Schweiz
bernhard.fehlmann@unibas.ch

ABSTRACT

Background With the Stroop-Interference-NoGo-Test (STING), we introduce an efficient and sensitive screening tool for the assessment of mild to moderate cognitive impairment. Its development was motivated by the ongoing economization of diagnostics and therapy in clinics as well as by the increased recognition of the effects of cognitive impairments on quality of life and professional reintegration. Established screenings such as the MoCA, MMSE and CAMCOG are either more time-consuming or lack sensitivity with regard to mild to moderate impairments in relevant domains.

Methods STING is based on the idea of an omnibus test. It integrates attentional, lexical-semantic, speed- and inhibitory components. In this way, a basic sensorimotor component is separated from a higher-order cognitive/executive component, which allows for differentiation between cognitive and generalised or merely sensorimotor impairments. The norms are based on data from 907 participants (386 M, 521 F). Its discriminative power was investigated in 64 patients (32 M, 32 F) with heterogeneous, but predominantly mild to moderate neuropsychological impairments.

Results The split-half reliability is essentially r = 0.82–0.95. For the parallel-test reliability, the index is r = 0.82–0.91, whereas the test-retest stability is estimated somewhat lower (r = 0.48–0.81). Practice effects are moderate (7–12 %). STING is correlated with many familiar tests, but sets itself apart from mere intelligence testing. Within the age category of 12–34 years, the number of correct items in the more complex second half of the test was predictive for clinical caseness, with a sensitivity of 83 % and a specificity of 47 %. Between the ages of 35 and 64, the classification was improved by the combination with the ratio of both halves, which represents set-shifting costs. Here the sensitivity of 71 % goes hand in hand with a specificity of 70 %.

Discussion STING provides a measure that can be considered sufficiently sensitive for use in the global assessment of cognitive impairment. A positive result does not replace a neuropsychological assessment, but indicates the need for one. The test offers an opportunity to neurologists, psychologists and psychiatrists to objectify mild to moderate, transient, or chronic functional impairments and to evaluate their course over time.

Introduction

Neurologists, psychologists and psychiatrists now have a barely manageable range of cognitive test methods at their disposal. Testing all functional areas in detail would take several hours. Thus, the indication for a comprehensive neuropsychological examination is very limited, especially since the question of costs for care providers and patients is not satisfactorily clarified in any of the German-speaking countries.

Neuropsychological screening is therefore an attempt to test the integrity of higher cognitive functions efficiently and under low

stress conditions for patients. The aim is to evaluate as many cognitive functions as possible while maintaining efficiency.

Even established screening tools are not able to completely meet this challenge:

For instance, De Guise et al. [1] note that the Montreal Cognitive Assessment (MoCA, [2]) mainly assesses memory and speech functions, which remain intact with some neurological diseases. Executive and visuo-perceptual functions, which under certain circumstances are more sensitive to neurological lesions [3, 4], are less well represented. Mitchell [5] criticizes the Mini-Mental State Examination (MMSE, [6]), since it hardly includes attentional and executive functions and is therefore unsuitable for the screening of disorders associated with these functions such as Lewy body dementia and Parkinson's disease dementia.

Especially in the area of low-level cognitive disorders, the MMSE has a low sensitivity, which can motivate the use of alternatives such as the DemTect [7, 8]. The Cambridge Cognitive Examination (CAMCOG; [9]) has similar sensitivity problems [10]. In a comparative study of cognitive dementia screenings, most short screenings such as the clock test [11], the cognitive part of the Alzheimer's Disease Assessment Scale (ADAScog; [12]) and the Boston Naming Test [13] showed unsatisfactory detection rates compared to two elaborate test procedures [14].

The aim of the newly developed STING test is to provide a time-efficient global assessment of cognitive performance, which forms the basis for potentially indicating the need for a more comprehensive neuropsychological examination [15]. It tests the particularly vulnerable bottleneck functions attention [16, 17], lexical-semantic processing [18–20], processing efficiency as an aspect of mental speed [21, 22], and inhibition as part of the executive functions [23, 24].

Methods

Material and execution

In addition to the collection of demographic data, STING consists of two test parts including instruction texts. The first consists of 160 color words showing the four values RED, BLUE, GREEN and YELLOW, of which "all words except YELLOW" are to be crossed out line by line for 40 s. The second part consists of the same color words, which are printed in color. They differ in the two dimensions (color) word and color, with the four values RED, BLUE, GREEN and YELLOW. The task is to "cross out all the color words for which the word and its color do not match" for 80 s. The test material contains a detailed test manual, the sheets of the parallel versions A and B and the required four evaluation templates. The paper-and-pencil test can be carried out individually or as a group test. Based on the pure processing time of 2 min, a total of 4–5 min has to be calculated for the execution in a clinical setting.

Evaluation

The evaluation is supported by templates and evaluation sheets and can be varied in scope according to the test objective. The standard evaluation includes the values bR, R and AQ. The number of correctly solved items bR gives an insight into the simple sensorimotor speed with simple stimulus detection and provides a baseline for the psychometric assessment of cognitive performance. The number of correctly solved items R represents the efficiency of cognitive processing under difficult decision and reaction conditions. This task requires more attentional resources, the multidimensional extraction and coordinated linking of information, the inhibition of an ongoing action and the efficient connection to semantic concepts. The ratio of the 2 test parts AQ (= R/bR) expresses the relative deceleration due to increasing complexity and represents the cognitive costs of task switching. In addition, sensorimotor deficits and different attitudes towards the test can be identified and isolated.

In terms of additional evaluation, deletion (F_D) and omission errors (F_O) can be analyzed in the second part. Based on the number of these errors and the total number of processed items in the second test part, the values L (= G − (3 * (F_D + F_O)), S (= L/G) and K (= S * L) can be calculated. These provide information on the subject's diligence and continuity in executing the test: L corresponds to the number of items that a subject has processed in a concentrated manner, and thus the cognitive effort involved. Given 25 % wrong color-word combinations, random behavior would lead to a performance index of 0 on average due to triple weighting of the errors. S relates the performance index to the number of processed items so that it rewards those who have achieved the same performance while processing fewer items and thus have made fewer errors. K ensures that a subject with one correctly solved item does not receive an equal score as subjects with 100 correctly solved items - as would be the case with the diligence index - by including the total number of items processed in a concentrated manner [25]. These values can provide complementary information, especially in longitudinal studies.

Standardization

Norm data for the STING were based on 907 subjects (386 M, 521 F). Subjects had to have very good knowledge of German and be at least 12 years old to ensure basic literacy skills. People with neurological or neuropsychiatric impairments (depression, multiple sclerosis, ADHD, uncompensated dyslexia) as well as sensorimotor impairments (vision deficiency, hemiparesis, n = 11) were excluded. Processing strategies that were incorrect (n = 64) or strongly different from the norm (n = 94; < Q1 – 1.5 * IQR respectively > Q3 + 1.5 * IQR; [26]) were not considered for the analyses.

The sample included 43 % male and 57 % female subjects. 91 % of the subjects in the sample are right-handed, 9 % are left-handed. Also 91 % stated that their mother tongue was German or that they grew up bilingually with German as one of their languages. The age of the subjects was between 12 and 89 years (M = 42.21, SD = 19.74) with < 1 % without completing school, and 1 % with primary school and 16 % lower secondary school certificates, 23 % with apprenticeship diploma, 13 % higher secondary, 12 % higher vocational and 34 % university or equivalent degrees.

Clinical collective

STING procedure was completed by patients of the Institute for Neuropsychological Diagnostics and Imaging (INDB), the Center for Outpatient Rehabilitation (ZAR) of the Klinik Lengg and the Department of Neuropsychology of the University Hospital Zurich (USZ) in the context of clinically indicated neuropsychological standard diagnostics. Data for patients without diagnosis (n = 11)

or without a diagnosis with neuropsychological relevance (n = 14) or lack of evidence of neuropsychological impairment in their current medical history were disregarded. In light of the overall structure of STING, which was clearly defined at the outset of the study, the following conditions were also excluded: simple, purely executive functional impairments (n = 3), isolated losses in episodic memory (n = 1) or nonverbal learning (n = 1), isolated expressive speech disturbances without impairment of language comprehension (n = 2), compensated reading and spelling impairments (n = 2) mild cranial cerebral traumata without recognizable neuropsychological manifestation (n = 2), as well as patients suffering from epilepsy without recognizable neuropsychological impairments (n = 8). Combinations of these categories were possible (n = 3). The results of two other patients were not included because of their significantly higher age (77 and 84 years, respectively).

The majority of the 64 patients (32 M, 32 F) in the sample were included due to partial mild to moderate attentional and executive functional impairments (F07.8 according to ICD-10; [27]) (40 %). Other more frequent conditions included diagnosed ADHD (13 %), but also suspected ADHD (11 %), mild cranial brain traumata (6 %), newly diagnosed reading and spelling impairments (4 %) and suspected (4 %) or manifest (3 %) depressive episodes. Less frequent diagnoses included word-finding disorders, minor weaknesses in working memory and/or episodic memory, mild intelligence impairment, mild to moderately diminished motivation, anxiety disorder, obsessive-compulsive disorder, postencephalitic syndrome, HIV encephalitis, multiple sclerosis, ischemic infarction, intraventricular hemorrhage (all < 1 %).

Of the 50 % male and female patients, 94 % stated to be right-handed, 6 % were left-handed. 88 % grew up with German as a mother tongue or bilingually with German. Patients were between 14 and 65 years old (M = 36.22, SD = 14.33) with 6 % primary school, 9 % lower secondary, 31 % apprenticeship diploma, 31 % higher secondary, 5 % higher vocational and 17 % university or equivalent degrees.

Ethics

The ethics evaluation was carried out according to the requirements of the Ethics Committee (for psychological and related research) at the Institute of Psychology of the Faculty of Philosophy of the University of Zurich: http://www.phil.uzh.ch/en/forschung/ethik.html#10. The data from the clinical trials have been retrospectively collected, stem from a completely anonymized data pool and comply with the legal requirements of the Swiss Human Research Regulation ("Humanforschungsverordnung").

Quality assessment

Demographic influences

A presumed age effect was first examined visually by means of scatterplots and motivated the formation of the three age categories (see results). Within these categories, systematic demographic effects were evaluated by means of multiple regression analyses and, if necessary, isolated by the calculation of standardized residuals. They are marked with the prefix «ZRes». After the reverse transformation into their non-standardized form, the test values including the corresponding percentile ranks and p values were entered into the norm tables.

Objectivity

The objectivity of application, evaluation and interpretation was assessed seperately and according to established guidelines [28].

Reliability

Within the framework of the quality criteria, reliability coefficients were calculated according to the split-half, the parallel-test and the test-retest method. For the estimation of the split-half coefficients (split-half reliability), the task was divided into 2 halves by time fractionation in a heterogeneous partial sampling of 35 test participants (12 M, 23 F) between 18 and 82 years (M = 49.57, SD = 19.46). Both task halves were correlatively compared and their relationship was evaluated with the Spearman-Brown formula. The parallel-test reliability was estimated by letting a parallel heterogeneous partial sample of 46 participants (18 M, 28 F) between 14 and 82 years (M = 42.13, SD = 23.42) solve both parallel forms A and B of STING in immediate succession. Practice and transference effects were controlled by a cross-over design. The correlation coefficient of both tests represents an estimation of reliability. The test-retest stability was recorded to analyze the time constancy of the entire test procedure. A partial sample of 66 high school students (30 M, 36 F) between 14 and 19 years (M = 16.39, SD = 1.20) was tested (T0; n = 62) and then retested after one week (T1; n = 46) and after one month (T2, relative to the first evaluation; n = 35). The comparisons are based on subjects who have completed both tests (T0 vs. T1: n = 42, T1 vs. T2: n = 34, T0 vs. T2: n = 31). The chosen approach allows a rough estimation of the temporal feature stability. The correlation coefficients of the measurement points can be directly interpreted [29]. In addition to the relative stability of the test values, practice effects of the tested group have to be examined and taken into account. In this regard, the change of the test values over the three survey points was quantified by One-Factor Analysis of Variance (ANOVAs) with measurement repetition and the within-subject factor "measurement time". Significant main effects were Bonferroni-corrected in the post-hoc comparisons.

Validity

The discriminant and convergent validity was assessed by the correlative comparison of the central values bR, R and AQ with the performance of closely and loosely related test methods. In this regard, data points of 128 persons (59 M, 69 F) between 14 and 63 years (M = 34.34, SD = 13.18) were available. In order to assess the criterion validity, the test performances of the clinical group were compared with the norm. The specificity and sensitivity of four different threshold values were compared through receiver-operating-characteristic (ROC) curves, with the area under the curve (AUC) as the statistical measure of discriminatory capacity. Also, positive (LR +) and negative (LR) likelihood ratios are shown. The former is an indicator of how much the likelihood of a clinical abnormality increases when a test result falls below the diagnostic decision threshold. The latter shows how much the probability of clinical abnormality decreases, if a test result is above the cut-off value. These data can be of great value for clinical diagnostics if there is already a hypothesis concerning the probability of neuropsychological impairment prior to screening. This probability is updated using the likelihood ratios (for details and examples of applications see [30]).

Results

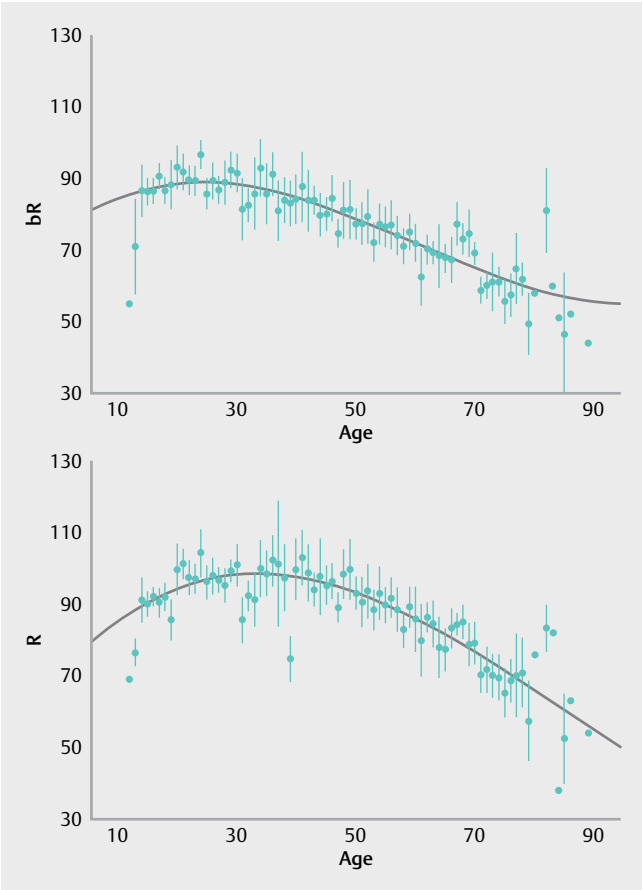
Evaluation

The standard evaluation takes 5–10 min depending on the experience of the evaluator. The additional evaluation takes another 4–8 min.

Demographic influences

The visual analysis of the scatterplots for age effects is shown in ▶ **Fig. 1** using the examples of values bR and R.

The figure suggests a cubic development pattern, with a plateau around age 22–34. The test values L, S, K and AQ show a similar pattern. In order to convert these patterns into adequately and easily applicable norm tables, the norm sample was roughly divided into 3 age sections: 12–34 years (n = 380), 35–64 years (n = 370) and 65–89 years (n = 157). The age influence within categories is taken into account by the age correction of the norm tables, which corresponds to the non-standardized beta coefficient of the regression models. The influence of education is also corrected by the norm tables. Small and partial effects of sex are not considered [31].



▶ **Fig. 1** Distribution of the number of correctly solved items in the norm sample by age. Annotation: n = 907. Error bar: + − 1.5 * Standard error. Above: Correct in the first test part (averaged per age). $y = 74.853359 + (1.276201 * x) - (0.033102 * x^2) + (0.000184 * x^3)$. Explained variance = 29.8%. Below: Correct in the second test section (averaged per age). $y = 70.3792 + (1.8378 * x) - (0.0342 * x^2) + (0.0001 * x^3)$. Explained variance = 24.3%.

Reliability

The estimated reliability values are shown in ▶ **Table 1**.

After three-time testing within one month, practice effects were detected for individual test values, which are listed in ▶ **Table 2**.

The practice benefits by three-time testing within one month develop roughly parallel for bR (7 %) and R (9 %). As a result, the quotient AQ does not react to the test repetition. Among the other values L shows a relatively large transfer effect (12 %), while S and K remain completely stable.

Validity

The results for construct validity are given in ▶ **Table 3**.

Regarding the criterion validity, the R value (AUC = 0.702) of the 12–34-year-olds has proved particularly suitable for differentiating between patients with mild to moderate cognitive impairments and the norm sample. The sensitivity corresponds to 66 % while specificity is 66 % (LR + = 1.90, LR − = 0.53). If the sensitivity is increased to 83 %, specificity decreases to 47 % (LR + = 1.56, LR − = 0.37). Among the 35–64 year olds, the clinical sample scores show lower values in all areas except for the first test part. For 34–65-year-olds, both R (AUC = 0.698) and AQ (AUC = 0.709) are suitable classifiers, which is why cut-off criteria have to be found for both values. In this case, the linking of their decision thresholds of the third criterion needs to be taken into consideration. If test values, which are below both thresholds are classified as conspicuous, a sensitivity of 71 % and a specificity of 70 % are achieved (LR + = 2.40, LR- = 0.41). There are no thresholds for the age category over 65 years. ▶ **Fig. 2** shows the distributions of the most suitable test values for both age categories after they have been standardized and corrected for interferences by regression models.

Discussion

Cost efficiency

Due to its short processing and evaluation time, STING is highly time-efficient, which is a key criterion for the use of diagnostic instruments in times of efficiency pressure [39]. Every neurologist, psychologist and psychiatrist should be able to administer STING efficiently and within minimal time, without having to fear a negative impact on the patient's acceptance and willingness to cooperate.

Objectivity

Objectivity was ensured by means of the uniform test materials, the unequivocal survey of demographic data, the precise instructions for the test application and the additional written presentation of the most important processing guidelines [28]. In addition, an example line and a practice line in both test parts help to understand the exercise. It can be assumed that these efforts ensure objectivity of application [28]. The evaluation follows clear rules, which also cover special cases, and is supported by the enclosed templates. The transparent calculation steps allow a simple and objective evaluation by hand or on the computer. Characterization of values and indices promotes interpretational objectivity [28]. The norm tables were divided according to age and level of education. Percentile ranks and confidence intervals allow the comparison of each subject with a suitable reference norm. By specifying

▶ **Table 1** Reliability estimates as a result of test halving, parallel testing and test repetition for different partial samples.

Test values	bR	R	AQ	L	S	K
Split-half reliability ρ_{tt} = (n = 35)	0.950 * * *	0.921 * * *	0.818 * * *	0.899 * * *	0.654 * *	0.849 * * *
Parallel-test reliability ρ_{tt} = (n = 46)	0.905 * * *	0.897 * * *	0.820 * * *	0.836 * * *	0.405 * *	0.738 * * *
Test-retest stability (n = 42); T_0 vs. T_1 ρ_{tt} =	0.731 * * *	0.754 * * *	0.478 * * *	0.696 * * *	0.630 * * *	0.653 * * *
(n = 34); T_1 vs. T_2 ρ_{tt} =	0.810 * * *	0.832 * * *	0.523 * *	0.782 * * *	0.429 *	0.691 * * *
(n = 31); T_0 vs. T_2 ρ_{tt} =	0.703 * * *	0.776 * * *	0.767 * * *	0.769 * * *	0.522 * *	0.698 * * *

T_0 vs. T_1 corresponds to a time interval of one week, T_1 vs. T_2 corresponds to three weeks and T_0 vs. T_2 corresponds to a one-month interval. Degrees of freedom are based on n–2. The correlations are significant at the level of 0.05 (*), 0.01 (**) and 0.001 (***), respectively.

▶ **Table 2** Practice effects after three-time testing within one month with a partial sample of 66 high school students.

Test values	bR	R	AQ	L	S	K
ANOVA F = (df = 2, 58)	7.340 * *	6.622 * *	0.008	4.128 *	2.725	2.307
Post-hoc comparisons T_0 vs. T_1 M_0, SD_0 = (n = 42) M_1, SD_1 = Average difference =	 93.90, 2.60 96.77, 2.79 2.87 *					
T_0 vs. T_2 M_0, SD_0 = (n = 31) M_2, SD_2 = Average difference =	93.90, 2.60 100.93, 2.42 7.03 * *	88.80, 2.55 96.80, 2.83 8.00 * *		75.80, 3.91 84.70, 3.34 8.90 *		

T_0 vs. T_1 corresponds to a time interval of one week, T_1 vs. T_2 corresponds to three weeks and T_0 vs. T_2 corresponds to a one-month interval. Post-hoc comparisons were carried out in pairs for significant ANOVAs. If identified, the mean differences are significant at the level of 0.05 (*), 0.01 (**) and 0.001 (***), respectively. The p-values of significant results have been Bonferroni-corrected.

practice effects, it is possible at least for 14–19-year-olds to formulate expectations on a repeated test. For other age groups, a cautionary approach is recommended regarding expectations of practice effects at this point. Except for this limitation, the requirement for interpretation objectivity is fulfilled.

Reliability

The high split-half reliability for the key values bR and R is indicative of the feature constancy within the test. Also for L and AQ split-half reliability takes a satisfactory value [25]. It is significantly lower for the diligence index. However, since the number of errors is already contained in bR and R, the diligence index does not represent a performance measure in the narrower sense, but serves to identify improper test processing. The same holds true for the continuity index K that has comparatively high reliability. The parallel test method also indicates good measurement accuracy and conditional constancy: the level of the values bR, R, AQ and L withstand the comparison with related test methods (e.g., FAIR-2: r = 0.76–0.83; [25]). For K and S reliability is slightly and significantly lower, respectively, which, being part of the more qualitative additional evaluation, does not imply any far-reaching limitations.

Validity

The degree to which STING reflects relevant real-life behavior of a subject [29] was assessed in order to evaluate content validity: The combination of (color and word) information must be checked for relevance, using the fast and consistent application of pre-defined criteria as a guiding basis for decision-making and action. At the end of the process, a yes/no decision is made, which must be implemented in a motorically precise manner. The pressure of making a fast and correct decision when faced with perceptively similar stimuli is presumably not unlike many real decision-making situations so that STING has content validity.

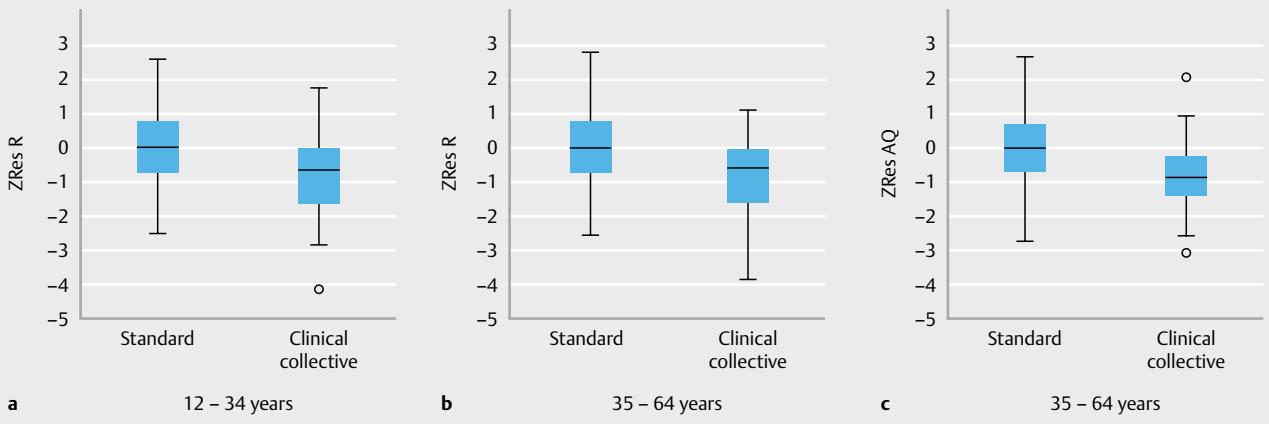
An indicator of a successful delimitation from the construct of intelligence is the weak correlation with the full-scale IQ of the WAIS. The inclusion of the attentional component is ensured through the correlations with the subtests of the TAP 2.0 and the RUFF 2 and 7-test. Among the associations with TMT, the correla-

▶ **Table 3** Correlative comparison of STING with other test procedures.

Test parameter	bR	R	AQ
WAIS full-scale IQ	0.232 *	–	
Repeating numbers	0.375 * *	0.335 * * *	
Picture completion	0.298 * * *	0.316 * *	
Number-symbol test	0.665 * * *	0.527 * * *	
Mosaic test	0.469 * * *	0.299 * *	
Symbol search	0.513 * * *	0.474 * *	
Finding commonalities	–	–	
General knowledge	–	–	
General comprehension	–	–	
TAP 2.0[2]			
Alertness			
Without warning sound	0.383 * * *	0.433 * * *	
No warning sound	–	0.308 * *	
Go/NoGo			
Inhibition error	–	−0.500 * * *	
Reaction speed	–	–	−0.363 *
Omission	–	–	–
Split attention			
Visual	0.350 * *	0.380 * * *	
Auditory	0.242 *	0.278 *	
2 and 7[3]			
Simple visual search	0.548 * * *	0.470 * * *	
Complex visual search	0.596 * * *	0.490 * * *	
TMT[4]			
Part A	0.700 * * *	0.529 * * *	
PArt B	0.419 * * *	0.373 * *	
TMT quotient			0.298 *
FWIT[5]			
2nd test part	0.445 * * *	0.404 * * *	
3rd test part	0.472 * * *	0.474 * * *	
Grooved Pegboard[6]	0.469 *	–	–

1. Wechsler Intelligence test for adults [32], German adaptation of the WAIS-III by David Wechsler [33]. 2: Test battery for attention testing [34]. 3: RUFF 2 and 7 test [35]. 4: Trail-making test (see [36]) 5: Color word interference test according to JR Stroop (Victoria version [37]) 6: Grooved Pegboard [38] The correlations shown are at the level of 0.05 (*), 0.01 (**), and 0.001 (***), respectively. Non-significant correlations are marked with –. In case of empty fields, based on conceptual considerations, no comparisons were calculated.

tion of AQ with the TMT quotient, both of which cover the successful change of attention, is particularly noteworthy. An indicator for the successful integration of the lexical-semantic component is the relationship between STING and the second part of the adapted FWIT, which primarily tests the reading speed. The weak correlations with intelligence can be associated with the speed component that depends on intelligence [40]. At subtest level of the WAIS, correlations with the visuomotor coordination ("number symbol test") and the mental processing speed ("symbol search") become evident. Finally, it is assumed that the inhibition performance is detected with the new development of STING, which is evidenced by the correlation between inhibitory errors in the Go/NoGo subtest of the TAP and the values R and AQ. Furthermore, the more complex part of STING as well as the last third of the color word interference test (FWIT, "Stroop-Test") require suppression of a semantically related category. Motoric confounding (grooved pegboard), which is also detected, can be completely isolated by the formation of the attention quotient, which shows the meaningfulness of this value, especially for older subjects. In summary, it is clear that STING presents multimodal challenges, which are correlated as expected with established test methods. Thus, it is assumed that the heterogeneous construct is valid.



▶ **Fig. 2** Box and whiskers plots of the most suitable test parameters for distinguishing between the standard and the clinical sample. Annotation: The z-standardized residues of the regression models are given. They represent the raw values adjusted for demographic interferences. The length of the box corresponds to the interquartile distance (IQR), the whiskers mark the value which is still within the limit of <Q1 – 1.5 * IQR respectively > Q3 + 1.5 * IQR. Right: Norm: n = 380; Clinical group: n = 29; Middle and Left: Norm: n = 370; Clinical group: n = 35.

Clinical diagnostics focuses on criterion validity, the assessment of which is based on specificity and sensitivity. At the age of 12–34 years, the number of correctly solved items in the more complex test part proved to be the best classifier. It operationalizes the bulk of the overall construction. In the 35–64 age group, the AQ quotient also clearly separates the clinical population from the norm. This is consistent with the motivation of its assessment. On the one hand, with higher age, sensory and motor factors increasingly influence the general processing speed. In order not to unjustifiably punish this slowing down and to classify healthy but somewhat slow working persons as conspicuous, the inclusion of the first test part makes sense and leads to a more reliable prediction. On the other hand, the ability to change tasks and mental flexibility are tested with the quotient. Since both are decreasing in age [41] and with the interaction with neuropsychological impairments, the examination of the costs of task switching is a valid means of diagnosing. STING makes it is possible to apply diagnostic cut-offs to the values R, AQ and their combination, the lower deviation of which is interpreted as an indication of a clinical abnormality. The weighting of sensitivity and specificity must always be context-based. STING proposes four alternative thresholds for clinical use [42]. They are listed in the test manual together with the associated sensitivity and specificity information and the likelihood ratios.

At first, the classification rates achieved in this way do not appear to indicate a particularly high discrimination capacity. Looking at related screening methods, however, it becomes clear that the clean detection of mild impairment forms is generally difficult: Scheurich and Brokate [43] report the highest possible sensitivity of 63 % and specificity of 62 % for the prediction of alcohol dependence. With regard to ADHD, a sensitivity of the TMT-B of 23 % is obtained with the 16th percentile, determined by means of a comprehensive test battery. However, at this point only 4 % of the persons without ADHD were misassigned [44]. In a meta-analysis, frontal lobe patients are not reported to have worse results in test part B compared to the norm. On the other hand, the test part A – which is interestingly correlated more strongly with STING – showed significant differences [45]. In a further study [46], detection of norm deviation of patients with severe disorders such as Alzheimer's disease, Parkinson's disease, Korsakoff syndrome, Huntington's disease, cranial brain trauma and schizophrenia was attempted based on elements from WAIS-III. For processing speed, a sensitivity of 73 % and a specificity of 84 % resulted, while in the case of working memory, the ratio of the correct positives decreased to 58 % with unchanged specificity. The list is by no means exhaustive. However, it is intended to show how strongly the detection performance of a test depends on the severity of the impairment of the patient and the prevalence of the examined feature.

The group of patients for clinical validation is deliberately heterogeneous. The considered diagnoses represent mild to moderate forms of neuropsychological impairment within the spectrum of neurological-psychiatric disorders. The detection of their partly subtle effects is challenging, but all the more important because there are few suitable procedures in this area. Considering these circumstances, the classification rates are appealing in the sense of diagnostic validity. In direct comparison with the TMT, STING also closes an important gap in the operationalization of executive functions (see [45]). In both age categories, the second test section is the better classifier for executive weaknesses. This is an indication that the more complex tasks lead to an increased involvement of the frontal brain areas and the conceptual delineation of the first test section has been successful. STING therefore provides a valuable enhancement particularly in the diagnosis of mild executive impairments. With its deliberately chosen low selectivity in the sense of an omnibus test, STING does not replace a neuropsychological examination but indicates its necessity. This makes it possible for neurologists who are clinically active to objectify a suspected function loss of cognitive performance, to monitor it over time, and, if necessary, to specify it by means of subsequent tests. The combination with more selective test methods is explicitly desired and promises an increase in diagnostic quality.

Limitations and outlook

An extension of the norm is desirable for every test procedure. With regard to STING, it is particularly important for higher age groups, where the sample groups are relatively small (minimum n = 77). A stronger inclusion of educationally deprived strata and a broadening of the scope of application to other languages would be instructive. Furthermore, more experience with parallel test B needs to be made and put into relation with the test form A. Evaluation of STING for a larger and more heterogeneous sample is needed for process monitoring and the characterization of the test performance over time. The current estimate of feature stability is rather crude and is not entirely satisfactory due to its restricted generalizability. In the same way, tests over a larger time span are needed to assess to degree to which existing information on practice effects represents an upper limit. In the clinical context, a widening of the sample to the age range of over 64 years is desirable, so that the previously missing recommendations for threshold values can be completed. The elaborate testing of patients with other disorders and other degrees of disorders can provide valuable new insights and possibly lead to an improvement of the discriminatory performance of STING.

Conclusion

In summary it can be stated that the development, standardization and validation of STING test has been successful. The result is an easy-to-use procedure, which can be applied within a very short time by physicians, psychologists as well as assistants in medical practices without test-specific prior knowledge. Despite the close guidance of the test user during execution and evaluation, STING remains highly flexible. Depending on the test objective, the threshold values can be adapted. The disadvantages of doing so can easily be weighed against the advantages. The method detects mild to moderate impairments sufficiently reliably. Precisely in this area the diagnostic validity of many similar tests is rare, which emphasizes the value of STING. Overall, it has thus proved to be a sufficiently sensitive screening method for the global assessment of cognitive impairments. The overview provided here gives reason to be optimistic that the test procedure will be applied in practice and that its validity can be extended to a wider range of diagnoses. The clinical data collected in this way will be the basis to further establish the value of STING.

Distribution

The test material, consisting of the test manual, both test sheets of the parallel versions A and B and 4 evaluation sheets, can be requested at the following address: sting@gmx.ch.

Conflict of interest

The authors state that there is no conflict of interest. All rights, in particular the right to duplication and dissemination, to reprinting and translation of the test material, belong to the authors. The test material can be ordered against a nominal charge (PDF-file: 25 €, paper form including 10 test sheets and including shipping costs: 50 €).

References

[1] De Guise E, LeBlanc J, Champoux MC et al. The mini-mental state examination and the montreal cognitive assessment after traumatic brain injury: An early predictive study. Brain Inj 2013; 27: 1428–1434

[2] Nasreddine ZS, Phillips NA, Bédirian V et al. The montreal cognitive assessment, MoCA: A brief screening tool for mild cognitive impairment. J Am Geriatr Soc 2005; 53: 695–699

[3] Jaillard A, Naegele B, Trabucco-Miguel S et al. Hidden dysfunctioning in subacute stroke. Stroke 2009; 40: 2473–2479

[4] Rabadi MH, Rabadi FM, Edelstein L et al. Cognitively impaired stroke patients do benefit from admission to an acute rehabilitation unit. Arch Phys Med Rehabil 2008; 89: 441–448

[5] Mitchell J. The mini-mental state examination (MMSE): An update on its diagnostic validity for cognitive disorders. In: Larner AJ ed. Cognitive screening instruments. A practical approach. London: Springer; 2013: 15–46

[6] Folstein MF, Folstein SE, McHugh PR. "Mini-mental state": A practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 1975; 12: 189–198

[7] Kalbe E, Kessler J, Calabrese P et al. DemTect: a new, sensitive cognitive screening test to support the diagnosis of mild cognitive impairment and early dementia. Int J Geriatr Psychiatry 2004; 19: 136–143

[8] Kalbe E, Brand M, Kessler J et al. Der DemTect in der klinischen Anwendung: Sensitivität und Spezifität eines kognitiven Screeninginstruments. Z Gerontopsychol -Psychiatrie 2005; 18: 121–130

[9] Huppert FA, Brayne C, Gill C et al. CAMCOG – A concise neuropsychological test to assist dementia diagnosis: Socio-demographic determinants in an elderly population sample. Br J Clin Psychol 1995; 34: 529–541

[10] De Jager CA, Milwain E, Budge M. Early detection of isolated memory deficits in the elderly: The need for more sensitive neuropsychological tests. Psychol Med 2002; 32: 483–491

[11] Goodglass H, Kaplan EF. The assessment of aphasia and related disorders. Philadelphia: Lea and Febiger 1983

[12] Rosen WG, Mohs RC, Davis KL. A new rating scale for Alzheimer's disease. Am J Psychiatry 1984; 141: 1356–1364

[13] Kaplan EF, Goodglass H, Weintraub S. The boston naming test. Philadelphia: Lea and Febiger; 1983

[14] Uttner I, Wittig S, Von Arnim CAF et al. Kurz und einfach ist nicht immer besser: Grenzen kognitiver Demenzscreenings. Fortschr Neurol Psychiatrie 2013; 81: 188–194

[15] Fehlmann B. Stroop-Interference-NoGo-Test – Entwicklung. Normierung und Validierung eines Screeningverfahrens zur globalen Erfassung neuropsychologischer Beeinträchtigung [Unpublizierte Masterarbeit]. Zürich: Universität Zürich; 2016

[16] Müller HJ, Krummenacher J, Schubert T. Aufmerksamkeitsnetzwerke im Gehirn. In: Müller HJ, Krummenacher J, Schubert T, (Hrsg.). Aufmerksamkeit und Handlungssteuerung. Berlin: Springer Berlin Heidelberg; 2015: 103–121

[17] Brickenkamp R, Schmidt-Atzert L, Liepmann D. Test d2–Revision. Aufmerksamkeits-und konzentrationstest. Manual. Göttingen: Hogrefe; 2010

[18] Rupp S. Semantisch-lexikalische Entwicklungsstörung. In: Thiel MM, Frauer C, Weber S, (Hrsg.). Semantisch-lexikalische Störungen bei Kindern. Sprachentwicklung: Blickrichtung Wortschatz. Berlin: Springer Berlin Heidelberg; 2013;73–106

[19] Jefferies E, Ralph MAL. Semantic impairment in stroke aphasia versus semantic dementia: A case-series comparison. Brain 2006; 129: 2132–2147

[20] Rogers TT, Patterson K, Jefferies E et al. Disorders of representation and control in semantic cognition: Effects of familiarity, typicality, and specificity. Neuropsychologia 2015; 76: 220–239

[21] Jokeit H, Grunwald T. Epilepsie und Gedächtnisbeeinträchtigungen. Z Epileptol 2003; 16: 137–143

[22] Prins ND, Van Dijk EJ, Den Heijer T et al. Cerebral small-vessel disease and decline in information processing speed, executive function and memory. Brain 2005; 128: 2034–2041

[23] Kaiser S, Mundt C, Weisbrod M. Exekutive Kontrollfunktionen und Neuropsychiatrische Erkrankungen-Perspektiven für Forschung und Klinik. Fortschr Neurol Psychiatrie 2005; 73: 438–450

[24] Heise KF, Zimerman M, Hoppe J et al. The aging motor system as a model for plastic changes of GABA-mediated intracortical inhibition and their behavioral relevance. J Neurosci 2013; 33: 9039–9049

[25] Moosbrugger H, Oehlschlägel J. FAIR-2. Frankfurter Aufmerksamkeits-Inventar 2. 2. überarb. ergänzte u. normenaktual. AuflBern: Huber; 2011

[26] Tukey JW. Exploratory data analysis. Reading: Addison-Wesley; 1977

[27] Dilling H, Mombour W, Schmidt MH. Internationale Klassifikation psychischer Störungen: ICD-10 Kapitel V (F) Klinisch-diagnostische Leitlinien. Bern: Huber; 2013

[28] Lienert GA, Raatz U. Testaufbau und Testanalyse. 6. Aufl Weinheim: Beltz; 1998

[29] Bühner M. Einführung in die Test-und Fragebogenkonstruktion. 3. aktual. AuflMünchen: Pearson; 2011

[30] Jaeschke R, Guyatt GH, Sackett DL. 1994; Users' Guides to the Medical Literature: III. How to use an article about a diagnostic test B. What are the results and will they help me in caring for my patients? JAMA 1994; 27: 703–707

[31] Goldhammer F, Hartig J. Testwertinterpretation. In: Moosbrugger H, Kelava A, (Hrsg.). Test- und Fragebogenkonstruktion. Berlin: Springer; 2007: 165–192

[32] Von Aster M, Neubauer A, Horn R. Wechsler Intelligenztest für Erwachsene (WIE). Deutschsprachige Bearbeitung und Adaptation des WAIS-III von David Wechsler. Frankfurt: Harcourt Test Services; 2006

[33] Wechsler D. WAIS-III: Wechsler Adult Intelligence Scale. Administration and Scoring Manual. San Antonio: Psychological Corporation; 1997

[34] Zimmermann P, Fimm B. TAP – Testbatterie zur Aufmerksamkeitsprüfung. Version 2.0. Herzogenrath: PSYTEST; 2002

[35] Ruff RM, Evans RW, Light RH. Automatic detection vs. controlled search: paper-and-pencil approach. Percept Mot Skills 1986; 62: 407–416

[36] Army Individual Test Battery. Manual of directions and scoring. Washington: War Department, Adjutant General's Office; 1944

[37] Spreen O, Strauss E. A compendium of neuropsychological tests. 2nd ed. New York: Oxford University Press; 1998

[38] Matthews CG, Klove H. Instruction manual for the Adult Neuropsychology Test Battery. Madison: University of Wisconsin Medical School; 1964

[39] Köchert R. Auswirkungen der Ökonomisierung auf die Versorgungsqualität in der Neurologie und Psychiatrie. In: Manzei A, Schmiede R, (Hrsg.). 20 Jahre Wettbewerb im Gesundheitswesen. Wiesbaden: Springer Fachmedien Wiesbaden; 2014: 299–317

[40] Neubauer AC, Fink A. Intelligence and neural efficiency. Neurosci Biobehav Rev 2009; 33: 1004–1023

[41] Ashendorf L, McCaffrey RJ. Exploring age-related decline on the Wisconsin Card Sorting Test. Clin Neuropsychol 2008; 22: 262–272

[42] Marx P, Lenhard W. Diagnostische Merkmale von Screeningverfahren. In: Hasselhorn M, Schneider W, (Hrsg.). Frühprognose schulischer Kompetenzen. Göttingen: Hogrefe; 2010

[43] Scheurich A, Brokate B. Neuropsychologie der Alkoholabhängigkeit. Göttingen: Hogrefe; 2009

[44] Lovejoy DW, Ball JD, Keats M et al. Neuropsychological performance of adults with attention deficit hyperactivity disorder (ADHD): Diagnostic classification estimates for measures of frontal lobe/executive functioning. J Int Neuropsychol Soc 1999; 5: 222–233

[45] Demakis GJ. Frontal lobe damage and tests of executive processing: A meta-analysis of the category test, stroop test, and trail-making test. J Clin Exp Neuropsychol 2004; 26: 441–450

[46] Taylor MJ, Heaton RK. Sensitivity and specificity of WAIS–III/WMS–III demographically corrected factor scores in neuropsychological assessment. J Int Neuropsychol Soc 2001; 7: 867–874