Validity, Reliability and Normative Data of The Stroop Test Çapa Version

Derya Durusu EMEK SAVAŞ[1], Deniz YERLİKAYA[2], Görsev G. YENER[3],
Öget ÖKTEM TANÖR[4]

SUMMARY

Objective: The Stroop test Çapa version does not have normative data, despite its extensive use in clinical and research settings to assess executive functions. The aim of the present study was to test the validity and reliability of the Stroop test Çapa version and to establish stratified normative data in individuals aged between 18-83 years.

Method: The norm determination phase of the study included 541 healthy participants, stratified by age, education, and gender. The relative contributions of the demographic variables on the completion times of Stroop subtests were assessed with multiple linear regression analysis. The main effects of age, education and gender variables and of interactions between these on the completion times of subtests were investigated with 6x3x2 ANOVA design. In addition, the concurrent validity, test-retest reliability and internal consistency of the test were examined.

Results: Multiple linear regression models that included age and education accounted for 23-42% of the completion time variances of all subtests. In the factorial ANOVA, main effects, as well as interaction effects of age and education were found on all subtests. For all Stroop subtests, the completion times were the shortest for the individuals in the 18-29 age group with the highest education level and longest for the individuals in the 70-83 age group with the lowest education level. The test demonstrated acceptable internal consistency and high test-retest reliability.

Conclusion: Normative data of the Stroop Test Çapa Version were provided for the assessment of executive functions in young and middle-aged adults and elderly population.

Keywords: Stroop test, Çapa version, normative data, validity, reliability, neuropsychological test

INTRODUCTION

The Stroop test is a widely used executive function test to evaluate selective attention, speed of information processing, response inhibition and cognitive flexibility. The original test was developed by John Ridley Stroop in 1935, but there are several versions of the test currently in use. While the number and type of stimuli and the task sequence vary in different Stroop versions (Strauss et al. 2006); they all reveal a phenomenon called the Stroop effect. In general, the Stroop effect occurs when individuals are presented with incongruent colour-word stimuli (e.g., the task requires to read the word “red” printed in blue ink or to name the colour of the ink instead of reading the word). Successful performance requires to inhibit an automatic behaviour (i.e., reading) in favour of a less practiced one (i.e., naming the colour of the ink). The attempt to inhibit automatic behaviour causes interference and results in longer reaction times, which is known as the Stroop effect (Stroop 1935).

Functional neuroimaging studies showed increased activation in the anterior cingulate cortex, which is strongly

Received: 31.07.2018, **Accepted:** 08.02.2019, **Available Online Date**: 08.08.2019

[1]Assist. Prof., Dokuz Eylül University, Departments of Experimental Psychology and Neurosciences, [2]Psychologist, Dokuz Eylül University, Department of Neurosciences, [3]Prof., Dokuz Eylül University Medical School, Department of Neurology, İzmir, [4]Prof., İstanbul University Medical School, Department of Neurology, Laboratory of Neuropsychology, İstanbul, Turkey.

e-mail: *derya.emek@deu.edu.tr*

https://doi.org/10.5080/u23549

implicated in selective attention, as well as in the middle frontal gyrus, motor areas and temporal lobe regions during Stroop performance (Alvarez and Emory 2006). While the Stroop test is known to be a sensitive tool for frontal lobe damage, several studies indicated that the brain activation during complex tasks such as the Stroop is not restricted to the frontal lobes but distributed to a larger neural network (Cohen et al. 1990, Alvarez and Emory 2006). Bilateral superior medial frontal lesions were found to be associated with increased number of errors and delayed reaction times on the Stroop incongruent condition; hence, error analysis was suggested to be considered in the assessment of populations with neurological disorders (Stuss et al. 2001). In addition to frontal lobe lesions, Stroop performance was reported to be impaired in neurological and psychiatric disorders (Kang et al. 2013), including Alzheimer's disease (Bondi et al. 2002), Parkinson's disease and frontotemporal dementia (Hsieh et al. 2008), attention deficit hyperactivity disorder (Balint et al. 2009, Rapport et al. 2001), major depression (Kravariti et al. 2009), bipolar disorder (Kravariti et al. 2009, Torrent et al. 2006) and schizophrenia (Camozzato and Chaves 2002).

The Stroop test is one of the most frequently used neuropsychological tests in Turkey and worldwide, as it is quickly administrable to various age groups and suitable for bedside examination. The most commonly used versions of Stroop include the Golden (1978), Victoria (Regard 1981), Dodrill (1978) and Comalli/Kaplan (Comalli et al. 1962). The Kaplan version uses the same subtests developed by Commali et al. (1962), but the order of administration is different (Strauss et al. 2006, Mitrushina et al. 2005). Different administration methods of Stroop require separate normative data for each version.

The Stroop test TBAG version, which is a combination of the original test (Stroop 1935) and the Victoria Stroop Test (VST) (Spreen and Strauss 1991), was developed and standardized for Turkish by Karakaş (2004), along with other neuropsychological tests within the BILNOT battery. While the VST has three conditions, the TBAG version has five. The participant is required to read colour names, name colours of dots and neutral words and lastly, name colours of words printed in incongruent colours (Karakaş 2004).

The Stroop test *Çapa* version used in the current study is an adaptation of a Stroop version developed by Weintraub (2000), by the Laboratory of Neuropsychology in the Behavioural Neurology and Movement Disorders Unit at Istanbul University Faculty of Medicine (*Çapa*). The administration of the Çapa version is similar to the Kaplan version. The participants are firstly asked to name the colour of rectangles, then to read the colour names and finally to name the colour of the colour words printed in incongruent colours, respectively. The number of errors and spontaneous corrections made is also counted. While the original Kaplan version has three conditions and 100 items in each condition, the Çapa version consists of three conditions and 60 items in each condition. Previous studies showed that long administration times may cause fatigue in elderly and patient populations, which result in poor test performance; thus, short test forms are recommended (Klein et al. 1997, Troyer et al. 2006, Kang et al. 2013). Kang et al. (2013) evaluated the validity of an abbreviated form of the Kaplan version by comparing the completion times of the first- and second-halves of the test. The authors concluded that the Stroop effect was evident in both halves and therefore, the abbreviated version including 50 items for each subtest should be sufficient to assess the speed of information processing and response inhibition in the elderly (Kang et al. 2013).

There are some differences between the Çapa and TBAG versions of the Stroop test that are currently used in Turkey. Compared to TBAG, the Çapa version (1) takes into account the number of errors and spontaneous corrections; (2) has fewer conditions and therefore, the administration time is shorter; (3) detects any presence of colour blindness or colour naming deficits in participants in the first condition; (4) would allow for a more accurate representation of the elderly population, as the research design includes three consecutive age groups for individuals over 50 years of age; and (5) is more accessible and widely-used with the test materials being available free of charge.

The Stroop test Çapa version has been used extensively in various hospitals, clinics and research settings without available normative data. However, standardized and well-normed measures are needed to avoid subjective and erroneous interpretation of test results. In an unpublished master's thesis, Tumaç (1997) provided normative data for the Çapa version on a sample of 180 adults across three age groups (15-28, 32-45 and 50-75 years) and three levels of education (low, moderate and high). Stroop performance was reported to be significantly better in the 15-28 age group compared to the 50-75 age group and in the high education group compared to the low education one in all conditions (Tumaç 1997).

Previous studies investigating the effects of age, education and gender on Stroop performance have consistently reported that aging negatively affects the test performance (Graf et al. 1995, Ivnik et al. 1996, Klein et al. 1997, Lucas et al. 2005, Van der Elst et al. 2006, Moering et al. 2004, Karakaş 2004). Studies also reported a significant effect of education on Stroop performance, indicating that individuals with higher levels of education had shorter completion times (Ivnik et al. 1996, Moering et al. 2004, Van der Elst et al. 2006, Karakaş 2004).

On the other hand, gender differences in Stroop performance were found to be relatively small. Several studies did not find any significant differences between male and female participants (Ivnik et al. 1996, Lucas et al. 2005, Troyer et al. 2006, Zalonis et al. 2009, Bayard et al. 2011, Bezdicek et al. 2015, Karakaş 2004), yet few studies reported that females significantly outperformed males on Stroop test (Moering et al. 2004, Van der Elst et al. 2006, Seo et al. 2008).

The present study aimed to test the concurrent validity, test-retest reliability and internal consistency of the Stroop test *Çapa* version, which has been widely used in Turkey without available norms, and to provide normative data stratified according to demographic variables for healthy adults aged 18-83 years. In line with this, the effects of age, education and gender on Stroop performance were examined.

METHOD

Participants

The study included 549 healthy individuals aged between 18-83 years. In line with previous studies (Hankee et al. 2016, Kang et al. 2013, Zalonis et al. 2009), the total sample was first divided into six subgroups (18-29, 30-39, 40-49, 50-59, 60-69 and 70-83 years) by decade of age and then into three strata of educational levels as low (5-8 years), moderate (9-11 years) and high (12 years and above) education. Data collection phase also accounted for gender along with age and education and each subgroup in the 6 *x* 3 *x* 2 ANOVA design included at least 10 participants.

Participants aged 18-49 years were recruited from various community sources including announcements in university billboards. Following a detailed personal and medical history, participants were administered the Stroop test Çapa version and the original Trail Making Test (TMT) (Reitan 1955).

The study population over 50 years of age consisted of healthy volunteers who participated in previous research studies conducted between 2011-2018 in the Department of Neurosciences at Dokuz Eylul University. The original screening procedures included a detailed neurological examination, laboratory tests, structural magnetic resonance imaging (MRI) and neuropsychological assessment regarding attention, memory, executive functions, visuospatial abilities and language. These individuals were also questioned for and did not report any subjective cognitive complaints. The Stroop test Çapa version was administered with the following tests as part of the routine neuropsychological assessment.

The **Mini-Mental State Examination** (Folstein et al. 1975) is a brief screening test to assess orientation, memory, attention, calculation, and language abilities. Güngen et al. (2002) suggested the cut-off score of 23/24 to detect mild dementia in the elderly Turkish population. The **Oktem Verbal Memory Processes Test** (OVMPT) (Öktem 1992) is a test of short- and long-term verbal episodic memory, standardized and validated for the Turkish population (Öktem 2011). The **Visual Reproduction Subtest of the Wechsler Memory Scale-Revised (WMS-R)** (Wechsler 1987) is a test of short-term and long-term visual memory. The validity and reliability of the test were established by Karakaş et al. (1996) and the normative data were provided by Mollahasanoğlu (2002) for the Turkish population. The **WMS-R Digit Span Test** (Wechsler 1987) consists of two parts: Digit Span Forward and Digit Span Backward. The validity and reliability of the test were established by Karakaş et al. (1996) and the normative data were provided by Mollahasanoğlu (2002) for the Turkish population. The **Verbal Fluency Tests** are used to assess semantic (i.e., animal) and phonemic (i.e., K, A, S) fluency. The normative data for the Turkish population were provided by Tumaç (1997). The **Clock Drawing Test (CDT)** is used to assess planning, abstract thinking and visuospatial/constructive abilities. Individuals were presented with a pre-drawn circle 10 cm in diameter and asked to put in the numbers of the clock and then to set the time to '10 past 11'. The CDT was scored according to the Manos and Wu (1994) scoring method, which was standardized and normed for the Turkish population over 50 years of age by Emek-Savaş et al. (2018). The **Boston Naming Test** (Kaplan et al. 1983) is a tool to assess confrontation naming ability. The present study used the 15-item short form (Mack et al. 1992), for which the normative data have not yet been formally published for the Turkish population. The **Geriatric Depression Scale** (GDS) (Yesavage et al. 1983) consists of 30 "Yes/No" questions to detect symptoms of depression in the elderly population. The GDS was validated for the Turkish population by Ertan et al. (1997) and the optimal cut-off score was 14.

Neuropsychological assessment was performed by neuropsychologists. The presence of depressive symptoms was evaluated by clinical interview and the GDS. The exclusion criteria for all participants consisted of any neurological or psychiatric disease, use of medication known to interfere with cognition, uncontrolled medical illness, history of alcohol or substance abuse, and/or head trauma with unconsciousness. Furthermore, additional exclusion criteria were applied for individuals over age 50 which included: (1) test scores below the age- and education-adjusted norms; (2) MMSE scores below 27; (3) GDS scores 14 and above; and (4) presence of severe vascular lesions (i.e., Fazekas score 2 and 3) and/or atrophy on structural brain MRI.

As the Stroop effect requires automatic and fluent reading ability, the present study included participants who had a minimum of five years of formal education,

Table 1. Demographic and Clinical Characteristics of Study Participants

		Age Groups					
Levels of Education		18-29	30-39	40-49	50-59	60-69	70-83
Low Education	N	23	22	22	26	36	28
	Age	22.26 ± 3.99	34.27 ± 3.64	44.18 ± 3.38	54.38 ± 2.64	64.31 ± 2.85	74.0 ± 3.50
	Education	7.74 ± 0.86	5.95 ± 1.43	5.55 ± 1.10	5.92 ± 1.35	6.22 ± 1.38	6.00 ± 1.41
	Gender (M/F)	11/12	10/12	10/12	11/15	18/18	13/15
	MMSE	-	-	-	28.50 ± 1.29	28.73 ± 1.08	28.71 ± 0.83
Moderate Education	N	22	23	24	29	24	22
	Age	23.05 ± 2.19	35.00 ± 3.00	44.17 ± 3.23	54.21 ± 2.64	64.71 ± 3.09	75.86 ± 3.73
	Education	10.86 ± 0.46	10.74 ± 0.54	11.00 ± 0.00	10.86 ± 0.44	10.92 ± 0.41	10.95 ± 0.21
	Gender (M/F)	11/11	12/11	12/12	11/18	10/14	11/11
	MMSE	-	-	-	29.15 ± 0.90	28.86 ± 0.95	29.08 ± 0.90
High Education	N	62	30	26	43	51	28
	Age	23.56 ± 2.86	33.13 ± 2.93	45.85 ± 2.95	55.77 ± 2.70	64.61 ± 3.21	74.71 ± 4.26
	Education	14.92 ± 2.66	16.63 ± 2.37	15.77 ± 2.67	15.21 ± 1.77	15.39 ± 2.36	15.89 ± 2.06
	Gender (M/F)	30/32	18/12	12/14	14/29	28/23	15/13
	MMSE	-	-	-	29.37 ± 0.88	29.29 ± 0.90	29.45 ± 0.80

MMSE: Mini-Mental State Examination Test; M: Male; F: Female; Low education: 5-8 years; Moderate education: 9-11 years; High education: 12 years and above.

which corresponds to primary education in Turkey. The demographic characteristics of the participants are presented in Table 1. The Supplementary Table 1 shows the detailed neuropsychological test scores of the study participants over age 50 stratified by age and level of education. The study protocol was approved by the ethics committee of the Dokuz Eylul University (13.07.2017; approval ID: 20170/18-02).

Materials

Stroop Test Çapa Version

The Stroop test Çapa version is an adaptation of a Stroop version developed by Weintraub (2000), by the Laboratory of Neuropsychology at Istanbul University Faculty of Medicine (*Çapa*). The test consists of two cards, each containing 60 items, presented in six rows of 10 items. On the first card, there are small rectangles (0.5 x 1 cm) printed in red, green and blue. The second card contains names of the three colours (i.e., red, green and blue) printed in incongruent colours (e.g., the word ‘red’ is printed in blue ink).

The Stroop test Çapa version consists of three parts: Stroop A, Stroop B and Stroop C.

Stroop A. The first part of the test requires individuals to name the colours (i.e., red, green and blue) of the small rectangles as quickly as possible, following a sequence from left to right. If any sign of colour blindness or colour naming deficits is present, the test is discontinued.

Stroop B. The second part of the test requires individuals to read as fast as possible the colour names (i.e., red, green and blue) that are printed in incongruent ink colours.

Stroop C. The last part requires individuals to name the colour of the ink the words are printed in (i.e., red, green and blue), instead of reading the words. The phenomenon known as the Stroop effect is revealed in this part because the inhibition of an automatic behaviour to perform an unusual one causes “interference”.

On each part, an individual’s performance is timed in seconds with a chronometer and recorded. On Stroop C, the number of errors and number of spontaneous corrections is also recorded. The “resistance to interference” (Stroop D), often used in clinical assessment is calculated as the reaction time difference between Stroop C and Stroop B.

Trail Making Test (TMT)

The TMT is a commonly used neuropsychological test to assess attention, response inhibition and cognitive flexibility. The original TMT was used in the current study (Reitan 1955). The TMT consists of two parts. The first part (TMT A) requires individuals to draw a line connecting circles with numbers in them (1-25) in numerical order, without lifting the pen from the paper. The second part (TMT B) consists of letters from A to L and numbers from 1 to 13, and individuals should draw a line alternating between numbers and letters in sequential order (i.e., 1-A, 2-B, 3-C…). Before each part, a practice trial is administered. The response times and the number of errors for each part are recorded.

A Turkish version of the TMT has been standardized using the Turkish alphabet for TMT B (Cangoz et al. 2009). The normative data for healthy adults aged 18-49 years were provided by Türkeş et al. (2015) and for individuals over age 50 by Cangoz et al. (2009).

Statistical Analysis

Statistical analyses were performed using SPSS v. 24.0. The following measures were included in the analyses: (1) the time to name the colours of rectangles (Stroop A); (2) the time to read the colour names (Stroop B); (3) the time to name the colour of the ink (Stroop C); (4) the number of spontaneous corrections in Stroop C; (5) the number of errors in Stroop C; and (6) the calculated time of resistance to interference (Stroop D).

The relative contributions of age (in years), education (in years) and gender variables on Stroop A, Stroop B, Stroop C and Stroop D performances were examined using stepwise multiple linear regression analysis. The variables were included in regression analyses with the order of age, education and gender. Age and education were entered as continuous variables and gender was coded as 0 or 1 for males and females, respectively.

The main effects of age (18-29, 30-39, 40-49, 50-59, 60-69, and 70-83 years), education (5-8, 9-11, and ≥12 years) and gender (male, female) variables and of interactions between them on test scores were investigated with a series of 6 *x* 3 *x* 2 ANOVA. Bonferroni correction was employed for post-hoc analysis. The number of spontaneous corrections and the number of errors on Stroop C were compared between age and education subgroups using the Chi-square test.

The concurrent validity was determined with Pearson correlation analysis by examining correlations between the Stroop C and TMT A, TMT B and TMT B-A scores, which are known to tap similar cognitive functions. The test-retest reliability was assessed with Pearson correlation analysis and the internal consistency was tested with Cronbach's alpha.

A value of $p<0.05$ was considered statistically significant for all analyses.

RESULTS

The assumptions of multiple linear regression and ANOVA were tested and met. Outliers were identified (a z-score value of +/- 3) for all test scores and excluded from further analysis. Six participants had Stroop D scores three standard deviations (SD) below the group mean. These individuals had longer response times on Stroop B compared to Stroop A, indicating poor reading skills; which resulted in extremely shorter Stroop D measures. As previous studies showed that automatic reading ability is necessary to reveal reliable Stroop interference (MacLeod and Dunbar, 1988; Ivnik et al., 1996), these participants were excluded from the study. Moreover, two participants whose TMT B-A scores exceeding the group mean ± 3 SD were excluded from the study.

Prior to multiple linear regression analysis, multivariate outliers were determined by calculating the Mahalanobis distance for all test scores. The threshold value for Mahalanobis distance was determined as df:3 and 16.266 for $p<0.001$. No multivariate outliers were detected. Thereafter, other assumptions of multiple linear regression analysis, linearity and multicollinearity, were tested respectively. Demographic variables included in the model showed linear relationships with test scores, and multicollinearity was not detected. The assumptions of normality and homoscedasticity of errors were tested by examining the scatter plots of standardized residuals and standardized predicted values. The assumption of independence of errors was tested with the Durbin-Watson test. Multiple linear regression analysis requires at least 40 participants for each independent variable in the model (Tabachnick and Fidell 2013). The final sample consisting of 541 participants met this requirement.

Prior to factorial ANOVA, histogram graphs and skewness values of Stroop A, Stroop B, Stroop C and Stroop D scores were examined for each subgroup to test the normal distribution of the data.

Multiple Linear Regression Analysis Results

The relative contributions of the age, education and gender variables on Stroop A, Stroop B, Stroop C and Stroop D scores were examined. Gender did not contribute to the regression model for any subtest score. However, age and education significantly influenced all Stroop subtest scores. The regression model including age and education accounted for 23-42% of total variance for all Stroop subtest scores (Table 2), indicating that older age and a lower level of education are associated with poorer performance on all Stroop subtests.

Factorial ANOVA Results

Stroop A

Main effects of age [$F_{(5,505)}=26.509$, $p<0.001$] and education [$F_{(2,505)}=17.439$, $p<0.001$] and an interaction effect of age *x* education [$F_{(10,505)}=2.458$, $p=0.007$] were found on Stroop A scores. The main effect of gender was not significant [$F_{(1,505)}=0.650$, $p=0.421$]. Post-hoc comparisons did not reveal significant differences between the 18-29 and 30-39 age groups and between the 40-49 and 50-59 age groups on Stroop A scores. The 50-59 age group performed better than the 60-69 age group on Stroop A; however, the difference between groups did not reach statistical significance ($p=0.050$). Significant differences were found between all

Table 2. Stepwise Multiple Linear Regression of Age (in years) and Education (in years) on Stroop Scores

		b	Standard error	ß	t	p	R²	ANOVA
Stroop A	(Constant)	35.120	1.321		26.593	<.001	.249	$F_{(2,538)}$=89.220, p<.001
	Age	.195	.018	.411	10.960	<.001		
	Education	-.506	.075	-.253	-6.749	<.001		
Stroop B	(Constant)	30.980	1.240		24.982	<.001	.353	$F_{(2,538)}$=146.850, p<.001
	Age	.212	.017	.441	12.667	<.001		
	Education	-.740	.070	-.366	-10.509	<.001		
Stroop C	(Constant)	61.348	2.533		24.222	<.001	.418	$F_{(2,538)}$=192.944, p<.001
	Age	.570	.034	.551	16.703	<.001		
	Education	-1.29	.144	-.297	-8.989	<.001		
Stroop D	(Constant)	30.439	2.319		13.127	<.001	.225	$F_{(2,538)}$=78.311, p<.001
	Age	.357	.031	.435	11.424	<.001		
	Education	-.552	.132	-.160	-4.194	<.001		

other age groups, indicating better performances in younger age groups (for all, p<0.015).

Pairwise comparisons showed that the high education group (≥12 years) had better Stroop A performances than the moderate and low education groups (p<0.001). The Stroop A scores did not differ between the low and moderate education groups (p=0.058). The interactions of age and education on Stroop A scores are summarized in Table 3a and 3b.

Stroop B

Main effects of age [$F_{(5,505)}$=35.582, p<0.001] and education [$F_{(2,505)}$=48.297, p<0.001] and an interaction effect of age *x* education [$F_{(10,505)}$=3.504, p<0.001] were found on Stroop B scores. The main effect of gender was not significant [$F_{(1,505)}$=0.027, p=0.869]. Post-hoc comparisons did not reveal significant differences between the 18-29 and 30-39 age groups, the 40-49 and 50-59 age groups and the 50-59 and

Table 3a. Comparisons between Age Groups Stratified by Levels of Education

Levels of Education	Age Groups	18-29	30-39	40-49	50-59	60-69	70+
Low Education	18-29		-	B, C	B, C	B, C	A, B, C
	30-39			A, B, C	A, B, C	A, B, C	A, B, C
	40-49				-	-	B
	50-59					-	B
	60-69						B
	70+						
Moderate Education	18-29		-	C, D	A, C, D	A, B, C, D	A, B, C, D
	30-39			-	-	A, B, C, D	A, B, C, D
	40-49				-	A, B, C	A, B, C
	50-59					C, D	A, B, C
	60-69						B
	70+						
High Education	18-29		-	C, D	A, B, C, D	A, B, C, D	A, B, C, D
	30-39			-	C, D	A, B, C, D	A, B, C, D
	40-49				-	-	C, D
	50-59					-	C, D
	60-69						C, D
	70+						

Low education: 5-8 years; Moderate education: 9-11 years; High education: ≥12 years; A: Stroop A; B: Stroop B; C: Stroop C; D: Stroop D.
In multiple comparisons, Stroop subtests that remain significant after Bonferroni correction were reported.
The "-" sign indicates no significant difference between the compared groups (p>.05).

Table 3b. Comparisons etween Education Groups Stratified by Age

Age Groups	Levels of Education	5-8	9-11	+12
18-29	Low Education		-	C, D
	Moderate Education			-
	High Education			
30-39	Low Education		-	-
	Moderate Education			-
	High Education			
40-49	Low Education		B	A, B, C
	Moderate Education			C
	High Education			
50-59	Low Education		B	B, C
	Moderate Education			-
	High Education			
60-69	Low Education		-	B, C
	Moderate Education			A, B, C, D
	High Education			
70+	Low Education		-	A, B
	Moderate Education			A, B
	High Education			

Low education: 5-8 years; Moderate education: 9-11 years; High education: ≥12 years; A: Stroop A; B: Stroop B; C: Stroop C; D: Stroop D.
In multiple comparisons, the Stroop subtests that remain significant after Bonferroni correction were reported.
The "-" sign indicates no significant difference between the compared groups (p>.05).

Table 4. Normative Values for the Stroop A

		Levels of Education		
		Low Education	Moderate Education	High Education
Age Groups				
18-29	n	(23)	(22)	(62)
	Mean ± SD	37.48 ± 9.67	32.36 ± 5.51	33.56 ± 6.37
	5%	24.60	26.00	24.17
	Median	36.00	32.00	33.00
	95%	68.40	49.35	43.85
30-39	n	(22)	(23)	(30)
	Mean ± SD	35.32 ± 7.55	33.48 ± 5.26	33.20 ± 6.95
	5%	25.00	26.40	25.00
	Median	33.50	36.00	31.50
	95%	48.85	44.00	49.50
40-59	n	(48)	(53)	(69)
	Mean ± SD	42.58 ± 10.06	38.81 ± 7.49	37.04 ± 5.75
	5%	26.80	26.00	28.00
	Median	41.00	38.00	37.00
	95%	68.20	52.30	46.00
60-69	n	(36)	(24)	(51)
	Mean ± SD	42.47 ± 8.85	44.67 ± 7.28	39.31 ± 6.61
	5%	31.00	31.25	28.60
	Median	40.00	44.00	40.00
	95%	64.05	59.00	52.00
70+	n	(28)	(22)	(28)
	Mean ± SD	46.54 ± 6.97	49.27 ± 11.46	40.36 ± 5.65
	5%	34.45	34.45	29.90
	Median	47.00	46.00	40.50
	95%	61.10	79.05	50.00

Low education: 5-8 years; Moderate education: 9-11 years; High education: 12 years and above; SD: Standard deviation.

60-69 age groups on Stroop B scores. Significant differences were found between all other age groups (for all, p<0.028).

Moreover, pairwise comparisons revealed significant differences between all educational groups (p<0.001), indicating improved Stroop B performance with higher levels of education. The interactions of age and education on Stroop B scores are summarized in Table 3a and 3b.

Stroop C

Main effects of age [$F_{(5,505)}$=49.626, p<0.001] and education [$F_{(2,505)}$= 31.186, p<0.001] and an interaction effect of age *x* education [$F_{(10,505)}$=1.871, p=0.047] were found on Stroop C scores. The main effect of gender was not significant [$F_{(1,505)}$=0.005, p=0.944]. In post-hoc comparisons, significant differences were found between all age groups (for all, p<0.047), except the 40-49 and 50-59 age groups on Stroop C scores. Pairwise comparisons revealed significant differences between all educational groups (p<0.021). The interactions of age and education on Stroop C scores are summarized in Table 3a and 3b.

No significant differences were observed between male and female participants [$\chi^2(5)$=3.464, p=0.629] or between age groups [$\chi^2(25)$=28.897, p=0.268] on the number of spontaneous corrections in Stroop C. However, the number of spontaneous corrections differed between educational groups [$\chi^2(10)$=66.411, p<0.001]. Pairwise comparisons revealed significant differences between all educational groups (p<0.013), indicating higher numbers of spontaneous corrections in the low education group compared to moderate and high education groups and in the moderate education group compared to the high education group.

No significant differences were observed between male and female participants [$\chi^2(5)$=7.624, p=0.178] or between age groups [$\chi^2(25)$=32.026, p=0.157] on the number of errors in Stroop C. However, the number of errors differed between educational groups [$\chi^2(10)$=30.277, p=0.001]. Pairwise comparisons revealed higher numbers of errors in low and moderate education groups compared to the high education group (p<0.006). No significant difference was found between low and moderate education groups (p=0.242).

Stroop D

Main effects of age [$F_{(5,505)}$=21.981, p<0.001] and education [$F_{(2,505)}$=6.577, p=0.002] and an interaction effect of age *x* education [$F_{(10,505)}$=2.300, p=0.012] were found on Stroop D scores, which is calculated as the reaction time difference between Stroop C and Stroop B. The main effect of gender was not significant [$F_{(1,505)}$=0.005, p=0.945]. Post-hoc comparisons did not reveal significant differences between the 40-49 and 50-59 age groups, the 40-49 and 60-69 age groups and the 50-59 and 60-69 age groups on Stroop D scores. Significant differences were found between all other age groups (for all, p<0.046).

Pairwise comparisons showed that the high education group had better Stroop D performances than the moderate and low education groups (p<0.001). The Stroop D scores did not differ between the low and moderate education groups. The interactions of age and education on Stroop D scores are summarized in Table 3a and 3b.

Normative Data

As there were no main or interaction effects of gender on any subtest score, the normative data were not stratified by gender. Moreover, there were no significant differences between the 40-49 and 50-59 age groups on any subtest score and the Stroop performances of these age groups did not differ according to educational levels (p>0.05; Table 3a). In line with these findings, the normative data were established for the 40-59 age group.

The normative data (means and standard deviations) for the completion time of each Stroop subtest stratified by age and education are presented in Tables 4-7. Tables 6b and

Table 5. Normative Values for the Stroop B

		Levels of Education		
		Low Education	**Moderate Education**	**High Education**
Age Groups				
18-29	n	(23)	(22)	(62)
	Mean ± SD	29.61 ± 7.43	27.95 ± 6.14	26.34 ± 4.77
	5%	20.40	20.30	19.00
	Median	27.00	27.00	26.00
	95%	43.80	47.60	35.85
30-39	n	(22)	(23)	(30)
	Mean ± SD	31.32 ± 7.66	29.52 ± 5.88	26.67 ± 4.02
	5%	20.45	19.60	20.10
	Median	29.50	28.00	27.00
	95%	45.85	41.00	35.80
40-59	n	(48)	(53)	(69)
	Mean ± SD	38.00 ± 7.50	32.47 ± 5.77	30.19 ± 5.28
	5%	25.00	23.00	21.50
	Median	37.00	32.00	30.00
	95%	50.55	42.30	40.00
60-69	n	(36)	(24)	(51)
	Mean ± SD	37.67 ± 8.72	37.21 ± 9.46	31.86 ± 6.55
	5%	26.25	23.50	25.00
	Median	36.00	35.50	30.00
	95%	56.00	58.50	48.20
70+	n	(28)	(22)	(28)
	Mean ± SD	45.54 ± 8.18	44.64 ± 12.97	32.39 ± 5.91
	5%	32.00	26.60	24.00
	Median	46.00	40.00	31.50
	95%	61.00	79.60	46.05

Low education: 5-8 years; Moderate education: 9-11 years; High education: 12 years and above; SD: Standard deviation.

Table 6a. Normative Values for the Stroop C

		Levels of Education		
		Low Education	**Moderate Education**	**High Education**
Age Groups				
18-29	n	(23)	(22)	(62)
	Mean ± SD	68.39 ± 13.93	58.00 ± 10.26	55.02 ± 9.41
	5%	41.20	41.90	38.00
	Median	66.00	55.50	54.50
	95%	95.00	81.85	72.40
30-39	n	(22)	(23)	(30)
	Mean ± SD	68.68 ± 13.66	68.52 ± 15.66	59.30 ± 10.59
	5%	41.35	49.20	40.65
	Median	68.00	67.00	58.00
	95%	89.70	107.80	78.45
40-59	n	(48)	(53)	(69)
	Mean ± SD	84.08 ± 13.31	76.89 ± 15.84	71.22 ±15.59
	5%	65.48	50.00	51.00
	Median	84.00	74.00	67.00
	95%	111.40	105.60	99.00
60-69	n	(36)	(24)	(51)
	Mean ± SD	86.14 ± 17.70	90.67 ± 19.48	73.61 ±12.15
	5%	63.40	58.25	53.00
	Median	80.00	88.00	73.00
	95%	117.45	126.00	95.40
70+	n	(28)	(22)	(28)
	Mean ± SD	92.89 ± 17.78	95.32 ± 18.44	85.54 ± 15.01
	5%	68.35	77.00	62.80
	Median	90.00	88.50	86.50
	95%	120.55	136.05	108.55

Low education: 5-8 years; Moderate education: 9-11 years; High education: 12 years and above; SD: Standard deviation.

Table 6b. Normative Values for the Number Spontaneous Corrections on Stroop C

		Levels of Education		
		Low Education	**Moderate Education**	**High Education**
Age Groups	Number of Spontaneous Corrections			
18-29	Mean (± 1 SD)	1.78 (0.40 – 3.16)	1.18 (0.27 – 2.09)	0.48 (0.00 – 1.24)
	Frequency			
	0	17.4	22.7	66.1
	1	30.4	45.5	21.0
	2	26.1	22.7	11.3
	3	13.0	9.1	1.6
	>3	13.0	0	0
30-39	Mean (± 1 SD)	1.73 (0.38 – 3.08)	1.43 (0.09 – 2.77)	0.53 (0.00 – 1.35)
	Frequency			
	0	22.7	30.4	63.3
	1	22.7	21.7	23.3
	2	22.7	34.8	10.0
	3	27.3	4.3	3.3
	>3	4.5	8.6	0
40-59	Mean (± 1 SD)	1.71 (0.00 – 3.17)	0.85 (0.00 – 1.97)	0.87 (0.00 – 1.98)
	Frequency			
	0	27.1	50.9	53.6
	1	20.8	26.4	18.8
	2	25.0	13.2	15.9
	3	10.4	7.5	10.1
	>3	16.7	1.9	1.4
60-69	Mean (± 1 SD)	1.42 (0.00 – 2.92)	1.54 (0.00 – 3.21)	1.00 (0.00 – 2.31)
	Frequency			
	0	41.7	37.5	47.1
	1	13.9	20.8	29.4
	2	19.4	16.7	9.8
	3	13.9	8.3	7.8
	>3	11.1	16.6	5.9
70+	Mean (± 1 SD)	1.71 (0.00 – 3.47)	1.05 (0.00 – 2.75)	0.46 (0.00 – 1.15)
	Frequency			
	0	42.9	54.5	64.3
	1	10.7	27.3	25.0
	2	3.6	4.5	10.7
	3	21.4	0	0
	>3	21.5	13.6	0

Mean ± 1 standard deviation (SD) and corresponding percentage values were reported for each group.
Low education: 5-8 years; Moderate education: 9-11 years; High education: 12 years and above.

Table 6c. Normative Values for the Number of Errors on Stroop C

		Levels of Education		
		Low Education	**Moderate Education**	**High Education**
Age Groups	Number of Errors			
18-29	Mean (± 1 SD)	0.35 (0.00 – 0.92)	0.18 (0.00 – 0.58)	0.05 (0.00 – 0.27)
	Frequency			
	0	69.6	81.8	95.2
	1	26.1	18.2	4.8
	2	4.3	0	0
	3	0	0	0
	>3	0	0	0
30-39	Mean (± 1 SD)	0.32 (0.00 – 0.97)	0.17 (0.00 – 0.66)	0.07 (0.00– 0.44)
	Frequency			
	0	77.3	87.0	96.7
	1	13.6	8.7	0
	2	9.1	4.3	3.3
	3	0	0	0
	>3	0	0	0
40-59	Mean (± 1 SD)	0.40 (0.00 – 1.38)	0.38 (0.00 – 1.17)	0.30 (0.00 – 1.09)
	Frequency			
	0	79.2	77.4	82.6
	1	12.5	11.3	10.1
	2	2.1	7.5	2.9
	3	4.2	3.8	2.9
	>3	2.1	0	1.4
60-69	Mean (± 1 SD)	0.67 (0.00 – 1.98)	0.46 (0.00 – 1.34)	0.18 (0.00 – 0.74)
	Frequency			
	0	75.0	75.0	88.2
	1	5.6	8.3	7.8
	2	5.6	12.5	2.0
	3	5.6	4.2	2.0
	>3	8.3	0	0
70+	Mean (± 1 SD)	0.75 (0.00 – 1.99)	0.55 (0.00 – 1.35)	0.11 (0.00 – 0.53)
	Frequency			
	0	64.3	63.6	92.9
	1	14.3	18.2	3.6
	2	10.7	18.2	3.6
	3	3.6	0	0
	>3	7.1	0	0

Mean ± 1 standard deviation (SD) and corresponding percentage values were reported for each group.
Low education: 5-8 years; Moderate education: 9-11 years; High education: 12 years and above.

Table 7. Normative Values for the Stroop D

		Levels of Education		
		Low Education	**Moderate Education**	**High Education**
Age Groups				
18-29	n	(23)	(22)	(62)
	Mean ± SD	38.78 ± 9.26	30.05 ± 7.81	28.68 ± 7.40
	5%	18.20	16.45	17.15
	Median	40.00	30.00	29.00
	95%	55.20	48.20	40.85
30-39	n	(22)	(23)	(30)
	Mean ± SD	38.27 ± 11.47	38.74 ± 12.88	32.63 ± 9.79
	5%	15.60	22.20	16.95
	Median	39.50	37.00	31.00
	95%	60.50	66.80	51.00
40-59	n	(48)	(53)	(69)
	Mean ± SD	45.85 ± 12.19	44.42 ± 13.85	41.03 ± 14.63
	5%	24.45	24.70	22.00
	Median	45.00	44.00	38.00
	95%	64.55	73.90	69.00
60-69	n	(36)	(24)	(51)
	Mean ± SD	48.47 ± 18.37	53.46 ± 16.15	41.75 ± 11.31
	5%	22.50	27.25	25.40
	Median	45.00	51.50	41.00
	95%	79.90	78.75	63.60
70+	n	(28)	(22)	(28)
	Mean ± SD	47.36 ± 16.89	50.68 ± 14.93	53.14 ± 15.16
	5%	20.45	25.75	31.35
	Median	44.50	51.00	54.00
	95%	75.30	81.85	77.55

Low education: 5-8 years; Moderate education: 9-11 years; High education: 12 years and above; SD: Standard deviation.

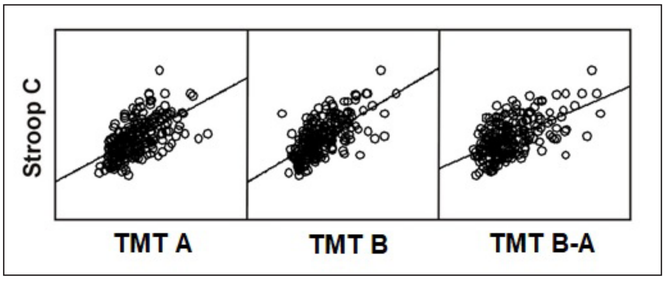

Figure 1. Scatterplots Displaying the Correlations between Stroop C and TMT A, TMT B and TMT B-A Scores.

6c contain normative data on the number of spontaneous corrections and errors, respectively.

Concurrent Validity

Stroop C scores were moderately correlated with TMT A (r=0.60), TMT B (r=0.65), and TMT B-A (r=0.57) scores (for all, p<0.001, Figure 1).

Test-Retest Reliability

The test-retest reliability of the Stroop test Çapa version was examined separately for the study populations aged 18-49 years and over 50 years of age, with randomly selected participants from each group. The Stroop test Çapa version was re-administered to 50 participants aged 18-49 years approximately two weeks after the initial testing. Fifty individuals were randomly selected from the study population over 50 years of age and their test scores at the first-year follow-up assessment (i.e., 12-months after the baseline neuropsychological assessment) were included in the analysis.

The test-retest reliability coefficients for the Stroop A, Stroop B, Stroop C and Stroop D scores were 0.78, 0.67, 0.88 and 0.80 for participants aged 18-49 years, and 0.84, 0.64, 0.81 and 0.77 for individuals over 50 years of age.

Internal Consistency

The Cronbach's alpha coefficient was 0.77 and the standardized alpha coefficient was 0.86 when Stroop A, Stroop B and Stroop C were included in the analysis. The Cronbach's alpha coefficient was decreased to 0.64 when Stroop A was removed. Similarly, the Cronbach's alpha coefficient was decreased to 0.66 when Stroop B was removed. However, the Cronbach's alpha coefficient was increased to 0.85 when Stroop C was removed. The number of spontaneous corrections and errors on Stroop C were not included in the internal consistency analysis, as they are participant-controlled responses and are very rarely observed in healthy individuals.

DISCUSSION

The present study tested the reliability and validity of the Stroop test Çapa version, which has been extensively used in clinical and research settings in Turkey and provided the normative data for individuals aged 18-83 years.

Previous studies reported that Stroop performance was affected by age, education and gender. The current normative study examined the relative contributions of these demographic variables on the Stroop test Çapa version performance. Age and education were accounted for 23-42% of the total variances of subtest scores. The times required for colour naming, word reading, interference and resistance to interference were found to be prolonged with increasing age. On the other hand, all subtest performances were improved with greater educational attainment.

The influence of age on test performance has been previously shown for different Stroop versions (Ivnik et al. 1996, Klein et al. 1997, Moering et al. 2004, Zalonis et al. 2009, Hankee et al. 2016, Karakaş 2004). Prolonged reaction times for colour naming and word reading with increasing age could be explained by the decreased speed of information processing associated with normal aging (Kang et al. 2013). In the present study, reaction times were found to be the shortest in the youngest group and longest in the oldest group for all subtests. There was no influence of age on the number of spontaneous corrections and errors that were recorded in the interference task (i.e., Stroop C).

Based on previous literature (Hankee et al. 2016, Kang et al. 2013, Zalonis et al. 2009), the participants were divided into six age groups by decade, with the exceptions of 18-29 and 70-83 age groups. Our results indicate a significant decline in the Stroop test Çapa version performance with each successive decade. While colour naming and word reading abilities were similar between the 18-29 and 30-39 age groups, the 18-29 age group showed better performances in interference and resistance to interference tasks. These findings suggest that even when motor speed was preserved in young adults, inhibitory control may still be influenced by age. As there was no difference in Stroop performance between the 40-49 and 50-59 age groups, normative values were provided for the 40-59 age group. The influence of age on the Stroop test Çapa version performance became evident again after 60 years of age. The individuals aged 70 years and above had significantly longer reaction times than the 60-69 age group in all subtests.

Several studies have shown that Stroop performance was affected by educational attainment (Anstey et al. 2000, Ivnik et al. 1996, Mitrushina et al. 2005, Troyer et al. 2006, Moering et al. 2004, Hankee et al. 2016, Karakaş 2004). In the present study, the levels of education were stratified as low (5-8 years), moderate (9-11 years) and high (12 years and above) education, in line with the education system in Turkey. All subtest performances were improved as the educational level increased. Moreover, an education effect was observed on the number of spontaneous corrections and errors. The number of spontaneous corrections was found to be different between individuals with low, moderate and high education; indicating lower numbers of spontaneous corrections with higher levels of education. Similarly, individuals with low and moderate education had higher numbers of errors in comparison to participants with high education. In the present study, an effect of education but not age was observed on the number of spontaneous corrections and errors; suggesting that age was more associated with the speed of information processing, while education was more related to inhibitory control. Other studies did not report an education effect on Stroop performance (Anstey et al. 2000, Moering et al. 2004). Therefore, educational stratification for normative data is not required but recommended (Mitrushina et al. 2005). In the present study, due to significant differences in performance between the education groups, normative data were provided for three educational levels for all subtests.

In comparison to age and education, gender had a small effect on Stroop performance. The absence of gender differences on test performance was also reported in other studies with different Stroop versions (Ivnik et al. 1996, Lucas et al. 2005, Troyer et al. 2006, Zalonis et al. 2009, Bayard et al. 2011, Bezdicek et al. 2015), including the Stroop test TBAG version which was developed by Karakaş (2004) and has available normative data for Turkey. The current study included the gender variable in the regression and ANOVA models along with age and education. However, as there was no influence of gender on any test score, the normative data were not stratified by gender.

In an unpublished master's thesis study including 180 healthy individuals, Tumaç (1997) reported the normative data for the Stroop test *Çapa* version for three age groups (15-28 years, 32-45 years, and 50-75 years) and three levels of education (0-5 years, 6-14 years, and 15 years and above). In Tumaç's (1997) study, between-group comparisons were performed separately for age and education using one-way ANOVA, which renders impossible to examine the interaction effects between variables. Other limitations include small sample size and very large age and education ranges. However, in line with the present study, Tumaç (1997) reported the effects of age and education but not gender on all subtest performances. In general, longer reaction times were observed with increasing age and lower levels of education (Tumaç 1997). In the present study, the most prominent change in test performance was observed between the 60-69 and the 70 years and above age groups. As individuals aged 50-75 years were examined as a single age group in Tumaç's (1997) study, the influence of physiological aging-related cognitive changes on test performance could not be assessed.

The present study also examined the concurrent validity, internal consistency and test-retest reliability of the Stroop test *Çapa* version. Results of our validity and reliability analyses showed that the Stroop test has strong psychometric properties. The concurrent validity was established by examining the relationships between the interference task score and the original TMT A, TMT B and TMT B-A scores. As the cognitive functions assessed by TMT were known to change with age, in the Turkish standardization study, the construct validity of the test was established by demonstrating that the subtest scores change as a function of age (Cangoz et al. 2009). The present study used the original TMT, which has unpublished normative data for the Turkish population. Similar to the Turkish version, the subtest scores of the original TMT were observed to change as a function of age, providing support for the construct validity of the test. However, a potential limitation would concern the use of the original TMT in this study, as the normative data were not published. Concerning the concurrent validity of the Stroop test *Çapa* version, moderate correlations were observed between the

Stroop interference task and the TMT subtest scores. As expected, the strongest relationship was observed with the TMT B scores, which assesses executive functions such as attention, planning, set-shifting and response inhibition.

Evidence from studies involving clinical samples indicates that the Stroop interference is associated with factors representing the speed of information processing, rather than executive function measures such as the TMT B (Strauss et al. 2006, Bondi et al. 2002, Boone et al. 1998). Earlier studies reported that the Stroop interference was more strongly associated with the TMT A and concluded that the TMT A was more sensitive to frontal lobe damage than the TMT B (Demakis 2004). In line with previous studies, the weakest relationship was observed between the Stroop interference and the TMT B-A scores (May and Hasher 1998, Strauss et al. 2006).

The Stroop test Çapa version showed moderate internal consistency when all subtests were included in the analysis and high internal consistency when the interference task was removed. These findings reflect overall compatibility among the Stroop subtests; nevertheless, the interference task taps on different cognitive functions than colour naming and word reading.

In the present study, the test-retest reliability was examined separately for the study populations aged 18-49 years and over 50 years of age. The interference task displayed the highest test-retest reliability. The Stroop test Çapa version was re-administered to participants aged 18-49 years after a two-week interval, and for participants over 50 years of age, previous test scores at the 12-month follow-up were examined. The test-retest reliability was found to be high and/or acceptable for both populations. In line with these findings, previous studies using different Stroop versions reported test-retest reliability coefficients between 0.60 and 0.90 (Franzen et al. 1987, Strauss et al. 2005, Lemay et al. 2004).

Except for the Stroop test Çapa version, the TBAG version developed by Karakaş (2004) under the BILNOT battery, is the only Stroop test that has available normative data for the Turkish population. The effects of demographic variables on Stroop performance were found to be similar for both versions. The normative data for the TBAG version were established for two age groups (i.e., 20-54 years and 55 years and older) and two levels of education (i.e., 5-8 years and 9 years and above) (Karakaş 2004). However, individuals aged 55 years and older were represented as a single age group in all the neuropsychological tests standardized within the BILNOT battery (Karakaş 2004). In the present study, individuals over 50 years and above were represented in three age groups, and as significant differences in performance were observed between these groups, the normative data were established for narrow age ranges. Therefore, in comparison to the TBAG version, the Stroop test Çapa version is a more comprehensive and sensitive assessment tool for the elderly population.

The test-retest reliability was assessed with a 12-month interval in individuals aged 20 years and above for the TBAG version and in individuals aged 50 years and above for the Çapa version. In both studies, the lowest test-retest reliability coefficients were obtained for the subtest that requires individuals to read the colour names printed in incongruent ink colours (i.e., 0.26 for TBAG Part 2 and 0.64 for Çapa Stroop B). Overall, the Stroop test Çapa version displayed higher test-retest reliability than the TBAG version, when other subtests involving similar tasks were compared.

In other normative studies from Turkey, healthy participants were selected on the basis of participant information sheets or short screening tests (i.e., MMSE, Montreal Cognitive Assessment, etc.) and functional assessment scales. The greatest strength of the present study is the use of neurological examination and neuroimaging results together with an extensive neuropsychological test battery to ensure the exclusion of common disorders among the elderly population that could affect cognitive functions such as mild cognitive impairment, depression or severe vascular lesions.

One limitation of the current study is that the test-retest reliability was assessed separately in individuals aged 18-49 years and 50 years and above using different retest intervals (i.e., 2 weeks and 12-months). Longer time intervals between test administrations increase the risk of developing cognitive impairment in older individuals. However, the retest administrations for individuals over 50 years of age were performed using a comprehensive neuropsychological test battery including the Stroop test Çapa version, and no changes were observed in cognitive functions at the 12-month follow-up assessment. In this regard, although the retest interval was long, findings from individuals over 50 years of age provide support for the reliability of the test.

Different versions of the Stroop test require separate normative data stratified by demographic variables. The present study established the normative values of the Stroop test Çapa version stratified by age and education for healthy adults aged 18-83 years, which can be used in research and clinical practice settings.

REFERENCES

Alvarez JA, Emory E (2006) Executive function and the frontal lobes: a meta-analytic review. Neuropsychol Rev 16: 17-42.

Anstey KJ, Matters B, Brown A et al (2000) Normative data on neuropsychological tests for very old adults living in retirement villages and hostels. Clin Neuropsychol 14: 309-17.

Bálint S, Czobor P, Komlósi S et al (2009) Attention deficit hyperactivity disorder (ADHD): gender- and age-related differences in neurocognition. Psychol Med 39: 1337-45.

Bayard S, Erkes J, Moroni C (2011) Victoria Stroop Test: normative data in a sample group of older people and the study of their clinical applications in the assessment of inhibition in Alzheimer's disease. Arch Clin Neuropsychol 26: 653-61.

Bezdicek O, Lukavsky J, Stepankova H et al (2015) The Prague Stroop Test: Normative standards in older Czech adults and discriminative validity for mild cognitive impairment in Parkinson's disease. J Clin Exp Neuropsychol 37: 794-807.

Bondi MW, Serody AB, Chan AS et al (2002) Cognitive and neuropathologic correlates of Stroop Color-Word Test performance in Alzheimer's disease. Neuropsychology 16: 335-43.

Boone KB, Pontón MO, Gorsuch RL et al (1998) Factor analysis of four measures of prefrontal lobe functioning. Arch Clin Neuropsychol 13: 585-95.

Camozzato A, Chaves ML (2002) Schizophrenia in males of cognitive performance: discriminative and diagnostic values. Rev Saude Publica 36: 743-8.

Cangoz B, Karakoc E, Selekler K (2009) Trail Making Test: normative data for Turkish elderly population by age, sex and education. J Neurol Sci 283: 73-80.

Cipolotti L, Spanò B, Healy C et al (2016) Inhibition processes are dissociable and lateralized in human prefrontal cortex. Neuropsychologia 93: 1-12.

Cohen JD, Dunbar K, McClelland JL (1990) A parallel distributed processing model of the Stroop effect. Psychol Rev 97: 332–61.

Comalli Pe Jr, Wapner S, Werner H (1962) Interference effects of Stroop color-word test in childhood, adulthood, and aging. J Genet Psychol 100: 47-53.

Demakis GJ (2004) 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 26: 441-50.

Dodrill CB (1978) A neuropsychological battery for epilepsy. Epilepsia 19: 611-23.

Emek-Savaş DD, Yerlikaya D, Yener GG (2018) Validity, Reliability and Turkish Norm Values of the Clock Drawing Test for Two Different Scoring Systems. Turk J Neurol 24: 143-52.

Ertan T, Eker E, Sar V (1997) Geriatrik depresyon ölçeğinin Türk yaşlı nüfusunda geçerlilik ve güvenilirliği. Noro Psikiyatr Ars 34: 62-71.

Folstein MF, Folstein SE, McHugh PR (1975) "Mini-mental state": a practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 12: 189-98.

Franzen MD, Tishelman AC, Sharp BH et al (1987) An investigation of the test-retest reliability of the Stoop color-word test across two intervals. Arch Clin Neuropsychol 2: 265-72.

Golden CJ (1978) Stroop Color and Word Test: A Manual for Clinical and Experimental Uses. Chicago, IL, Stoelting Co.

Graf P, Uttl B, Tuokko H (1995) Color-and picture-word Stroop tests: performance changes in old age. J Clin Exp Neuropsychol 17: 390-415.

Güngen C, Ertan T, Eker E et al (2002) Reliability and validity of the standardized Mini Mental State Examination in the diagnosis of mild dementia in Turkish population. Turk Psikiyatri Derg 13: 273-81.

Hankee LD, Preis SR, Piers RJ et al (2016) Population normative data for the CERAD word list and Victoria Stroop Test in younger-and middle-aged adults: cross-sectional analyses from the framingham heart study. Exp Aging Res 42: 315-28.

Hsieh YH, Chen KJ, Wang CC et al (2008) Cognitive and motor components of response speed in the Stroop test in Parkinson's disease patients. Kaohsiung J Med Sci 24: 197-203.

Ivnik RJ, Malec JF, Smith GE et al (1996) Neuropsychological tests' norms above age 55: COWAT, BNT, MAE token, WRAT-R reading, AMNART, STROOP, TMT, and JLO. Clin Neuropsychol 10: 262-78.

Kang C, Lee GJ, Yi D et al (2013) Normative data for healthy older adults and an abbreviated version of the Stroop test. Clin Neuropsychol 27: 276-289.

Kaplan E, Goodglass H, Weintraub S (1983) The Boston Naming Test. 2. Baskı, Lea & Febiger, Philadelphia.

Karakas S, Kafadar H, Eski R (1996) Test-retest reliability of the Turkish standardization of Wechsler Memory Scale-Revised. Turk Psikol Derg 11: 46-55.

Karakaş S (2004) BİLNOT Bataryası El Kitabı: Nöropsikolojik Testler için Araştırma ve Geliştirme Çalışmaları. 1. Baskı, Ankara, Dizayn Ofset.

Klein M, Ponds RW, Houx PJ et al (1997) Effect of test duration on age-related differences in Stroop interference. J Clin Exp Neuropsychol 19: 77-82.

Kravariti E, Schulze K, Kane F et al (2009) Stroop-test interference in bipolar disorder. Br J Psychiatry 194: 285-6.

Lemay S, Bédard MA, Rouleau I et al (2004) Practice effect and test-retest reliability of attentional and executive tests in middle-aged to elderly subjects. Clin Neuropsychol 18: 284-302.

Lucas JA, Ivnik RJ, Smith GE et al (2005) Mayo's older African Americans normative studies: Norms for boston naming test, controlled oral word association, category fluency, animal naming, token test, wrat-3 reading, trail making test, Stroop test, and judgment of line orientation. Clin Neuropsychol 19: 243-69.

Mack WJ, Freed DM, Williams BW et al (1992) Boston Naming Test: Shortened versions for use in Alzheimer's disease. J Gerontol 47: 154-8.

MacLeod CM, Dunbar K (1988) Training and Stroop-like interference: Evidence for a continuum of automaticity. J Exp Psychol Learn Mem Cogn 14: 126-35.

Manos PJ, Wu R (1994) The ten point clock test: a quick screen and grading method for cognitive impairment in medical and surgical patients. Int J Psych Med 24: 229-44.

May CP, Hasher L (1998) Synchrony effects in inhibitory control over thought and action. J Exp Psychol Hum Percept Perform 24: 363-79.

Mitrushina M, Boone KB, Razani J et al (2005) Handbook of Normative Data for Neuropsychological Assessment. 2. Baskı, New York, NY, US, Oxford University Press.

Moering RG, Schinka JA, Mortimer JA et al (2004) Normative data for elderly African Americans for the Stroop color and word test. Arch Clin Neuropsychol 19: 61-71.

Mollahasanoğlu A (2002) Normal Deneklerde Bir Grup Görsel ve Sözel Bellek Tersleri Performansına Yaş ve Eğitimin Etkisi. İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü, Psikoloji Bölümü, Yayımlanmamış Yüksek Lisans Tezi.

Oktem O (1992) A verbal test of memory processes. Arch Neuropsychiatry 29: 196-206.

Oktem-Tanor O (2011) Öktem Sözel Bellek Süreçleri Testi (ÖKTEM SBST) El Kitabı. Birinci Baskı, Ankara, Türk Psikolog Derneği Yayınları.

Ozdeniz E (2001) Bir Grup Sağ Hemisfer ve Dikkat Testleri Performansına Yaş ve Eğitim Değişkenlerinin Etkisi. İstanbul Üniversitesi, Sosyal Bilimler Enstitüsü, Psikoloji Bölümü, Yayımlanmamış Yüksek Lisans Tezi.

Rapport LJ, Van Voorhis A, Tzelepis A et al (2001) Executive functioning in adult attention-deficit hyperactivity disorder. Clin Neuropsychol 15: 479-91.

Regard M (1981) Cognitive rigidity and flexibility: A neuropsychological study. Unpublished doctoral dissertation. University of Victoria.

Reitan RM (1955) The relation of the trail making test to organic brain damage. J Consult Psychol 19: 393-4.

Seo EH, Lee DY, Choo IH et al (2008) Normative study of the Stroop Color and Word Test in an educationally diverse elderly population. Int J Geriatr Psychiatry 23: 1020-7.

Spreen O, Strauss E (1991) A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary. 1. Baskı, New York, NY, US, Oxford University Press.

Strauss E, Sherman EM, Spreen O (2006) A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary. 3. Baskı, New York, NY, US, Oxford University Press.

Strauss GP, Allen DN, Jorgensen ML et al (2005) Test-retest reliability of standard and emotional Stroop tasks: An investigation of color-word and picture-word versions. Assessment 12: 330-37.

Stroop JR (1935) Studies of interference in serial verbal reaction. J Exp Psychology 18: 643-62.

Stuss DT, Floden D, Alexander MP et al (2001) Stroop performance in focal lesion patients: dissociation of processes and frontal lobe lesion location. Neuropsychologia 39: 771-86.

Tabachnick BG, Fidell LS (2013) Multiple Regression. Using Multivariate Statistics. 6. Baskı, Boston, MA, Pearson, s. 124.

Torrent C, Martínez-Arán A, Daban C et al (2006) Cognitive impairment in bipolar II disorder. Br J Psychiatry 189: 254-9.

Troyer AK, Leach L, Strauss E (2006) Aging and response inhibition: Normative data for the Victoria Stroop Test. Neuropsychol Dev Cogn B Aging Neuropsychol Cogn 13: 20-35.

Tumac A (1997) Normal deneklerde frontal hasarlara duyarlı bazı testlerde performansa yaş ve eğitimin etkisi. İstanbul Üniversitesi Sosyal Bilimler Enstitüsü, Psikoloji Bölümü, Yayımlanmamış Yüksek Lisans Tezi, İstanbul.

Türkeş PN, Can PH, Kurt PM et al (2015) A Study to Determine the Norms for the Trail Making Test for the Age Range of 20-49 in Turkey. Turk Psikiyatri Derg 26: 189-96.

Van der Elst W, Van Boxtel MP, Van Breukelen GJ et al (2006) The Stroop color-word test: influence of age, sex, and education; and normative data for a large sample across the adult age range. Assessment 13: 62-79.

Wechsler D (1987) Wechsler Memory Scale-Revised Manual. San Antonio, TX, Psychological Corporation.

Weintraub S (2000) Neuropsychological assessment of mental state. Principles of Cognitive and Behavioral Neurology, 2. Baskı, MM Mesulam (Ed), New York, NY, Oxford University Press s. 130.

Yesavage JA, Brink TL, Rose TL et al (1983) Development and validation of a geriatric depression screening scale: a preliminary report. J Psychiatr Res 17: 37-49.

Zalonis I, Christidi F, Bonakis A et al (2009) The stroop effect in Greek healthy population: normative data for the Stroop Neuropsychological Screening Test. Arch Clin Neuropsychol 24: 81-8.

Supplementary Table 1. Neuropsychological Test Scores of the Study Participants Over 50 Years of Age

Education	Neuropsychological Tests	Age (in years)		
		50-59	60-69	70+
5-8 years	MMSE	28.50 ± 1.29	28.73 ± 1.08	28.71 ± 0.83
	GDS	6.75 ± 7.23	6.58 ± 4.40	7.00 ± 5.06
	Digit Span Forward	5.33 ± 0.82	4.89 ± 0.63	4.77 ± 0.69
	Digit Span Backward	4.17 ± 0.41	4.18 ± 0.39	4.09 ± 0.29
	OVMPT IM	6.17 ± 1.47	5.36 ± 1.37	5.00 ± 1.31
	OVMPT Total	123.33 ± 13.11	112.75 ± 13.92	110.45 ± 8.69
	OVMPT FR	13.00 ± 1.55	13.04 ± 1.45	12.86 ± 1.25
	OVMPT TR	15.00 ± 0.00	14.82 ± 0.48	15.00 ± 0.00
	Visual STM	12.20 ± 1.30	10.30 ± 2.49	8.73 ± 1.36
	Visual LTM	11.20 ± 1.78	10.20 ± 2.46	8.48 ± 1.04
	CDT	9.80 ± 0.45	9.38 ± 0.64	9.57 ± 0.61
	Semantic Fluency	22.00 ± 3.90	20.89 ± 5.20	19.91 ± 3.37
	Phonemic Fluency	35.50 ± 12.37	35.40 ± 11.14	25.17 ± 2.93
	BNT-15	14.83 ± 0.41	14.80 ± 0.50	14.89 ± 0.32
9-11 years	MMSE	29.15 ± 0.90	28.86 ± 0.95	29.08 ± 0.90
	GDS	7.23 ± 4.68	5.29 ± 3.52	7.93 ± 3.54
	Digit Span Forward	5.53 ± 0.74	5.25 ± 1.16	4.67 ± 0.62
	Digit Span Backward	4.40 ± 0.51	4.65 ± 0.99	4.20 ± 0.56
	OVMPT IR	5.33 ± 1.23	5.05 ± 1.05	5.44 ± 1.50
	OVMPT Total	116.80 ± 12.62	113.30 ± 12.18	107.38 ± 12.33
	OVMPT FR	12.73 ± 1.28	13.20 ± 1.61	12.94 ± 1.24
	OVMPT TR	14.87 ± 0.35	14.95 ± 0.22	14.94 ± 0.25
	Visual STM	11.60 ± 2.10	10.20 ± 2.33	9.88 ± 1.82
	Visual LTM	11.20 ± 2.24	9.20 ± 1.82	9.00 ± 1.97
	CDT	9.67 ± 0.49	9.65 ± 0.61	9.56 ± 0.63
	Semantic Fluency	24.40 ± 5.68	21.35 ± 4.13	19.00 ± 4.83
	Phonemic Fluency	38.67 ± 11.75	33.17 ± 8.46	34.81 ± 11.27
	BNT-15	15.00 ± 0.00	14.47 ± 0.61	14.75 ± 0.45
12 years and above	MMSE	29.37 ± 0.88	29.29 ± 0.90	29.45 ± 0.80
	GDS	5.63 ± 4.39	6.49 ± 4.51	4.52 ± 3.73
	Digit Span Forward	6.03 ± 1.28	6.10 ± 1.03	5.89 ± 0.96
	Digit Span Backward	4.67 ± 0.72	4.88 ± 0.99	4.46 ± 0.58
	OVMPT IR	6.33 ± 0.67	6.25 ± 1.62	5.71 ± 1.46
	OVMPT Total	124.11 ± 11.61	123.18 ± 11.83	113.79 ± 13.53
	OVMPT FR	13.33 ± 1.31	13.27 ± 1.44	12.68 ± 1.39
	OVMPT TR	15.00 ± 0.00	15.00 ± 0.00	14.96 ± 0.19
	Visual STM	12.50 ± 1.84	12.79 ± 1.53	10.73 ± 2.24
	Visual LTM	11.61 ± 2.35	11.94 ± 2.17	10.42 ± 2.37
	CDT	9.80 ± 0.47	9.81 ± 0.39	9.73 ± 0.53
	Semantic Fluency	23.72 ± 4.12	23.90 ± 5.25	22.79 ± 3.89
	Phonemic Fluency	44.44 ± 13.87	43.96 ± 10.87	42.80 ± 10.60
	BNT-15	14.92 ± 0.37	14.84 ± 0.48	14.92 ± 0.27

MMSE: Mini-Mental State Examination Test, GDS: Yesavage Geriatric Depression Scale, OVMPT: Oktem Verbal Memory Processes Test, OVMPT IM: Immediate Recall, OVMPT Total: Total Learning, OVMPT FR: Free Recall, OVMPT TR: Total Recall, Visual STM: The Visual Reproduction Subtest of the WMS-R Short-Term Memory, Visual LTM: The Visual Reproduction Subtest of the WMS-R Long-Term Memory, CDT: Clock Drawing Test, BNT-15: 15-Item Short Form of the Boston Naming Test.

INSTRUCTIONS FOR THE STROOP TEST ÇAPA VERSION

Preparation of the Material

The stimulus cards of the Stroop test Çapa version should be printed in colour with the same aspect ratios. It should be ensured that a printer setting that would affect the stimuli colours is not active. After the cards have been printed, they should be cut on the dashed lines and covered separately by PVC before being used.

Administration

The individual should be seated, and care should be taken to ensure that there are no distracting stimuli in the environment during test administration. The Stroop test Çapa version consists of two stimulus cards. The first card contains red, blue and green coloured rectangles and the second card contains colour names printed in incongruent ink colours. The Stroop test Çapa version consists of three parts: Stroop A, Stroop B and Stroop C.

Stroop A (Naming the Colours of Coloured Rectangles):

The first part of the test requires the individual to name the colours of the small rectangles on the first stimulus card. Position the first stimulus card in front of the individual. Point the stimulus card and give the following instruction, "Now I will ask you to name the colours of these rectangles as fast as possible. Follow a sequence from left to right. When you complete the line, move to the line below it. Start when you are ready". Start the chronometer as the individual starts naming the colours and stop when the individual names the last rectangle. Record the completion time, in seconds, on the scoring sheet. If any sign of colour blindness or colour naming deficits is present, the test should be terminated.

Stroop B (Reading the Colour Names):

The second part of the test requires the individual to read the colour names that are printed in incongruent ink colours on the second stimulus card. Position the second stimulus card in front of the individual. Point the stimulus card and give the following instruction, "I will ask you to read the written words as fast as possible. You will follow a sequence from left to right, as in the previous task. When you complete the line, move to a subline. Start when you are ready". Start the chronometer as the individual starts reading the colour names and stop when the individual completes the task. Record the completion time, in seconds, on the scoring sheet.

Stroop C (Naming the Colour of the Ink):

The last part requires the individual to name the colour of the ink that the words are printed in. Again, point the second stimulus card and give the following instruction, "Now, instead of reading the colour names that you have just read, I will ask you to name the ink colour of these words as fast as possible. If you realize that you have made a mistake, please correct it immediately and then continue naming. Start when you are ready".

Start the chronometer as the individual starts naming the ink colours and stop when the individual completes the task. Record the completion time, in seconds, on the scoring sheet.

Underline the colour names on the scoring sheet when the individual makes an error. When the individual realizes that he/she has made a mistake and correct it immediately, mark this as a spontaneous correction by drawing a circle around the colour name. Record the total number of spontaneous corrections and errors on the scoring sheet.

Stroop D (Resistance to Interference):

Calculate the reaction time difference between Stroop C and Stroop B, and record it on the scoring sheet as Stroop D.