Discriminative validity of the Stroop Test Çapa Version for executive function deficits in bipolar disorder

Ceren Hıdıroğlu Ongun[1], Derya Durusu Emek Savas[2], Deniz Ceylan[3], Ayşegül Özerdem[4]
[1]Assis. Prof., [2]Assoc. Prof., Dokuz Eylul University Faculty of Letters, Department of Psychology, Experimental Psychology Program, Izmir, Turkey https://orcid.org/0000-0002-5754-6042-https://orcid.org/0000-0001-7042-697X
[3]Assoc. Prof., Koc University School of Medicine, Department of Psychiatry, Istanbul, Turkey https://orcid.org/0000-0002-1438-8240
[4]Prof., Mayo Clinic Department of Psychiatry and Psychology, Rochester, Minnesota, USA https://orcid.org/0000-0002-9455-5896

SUMMARY

Objective: Cognitive impairment is a well-recognized feature of bipolar disorder and has been investigated as a potential endophenotypic marker. The Stroop test is a widely used measure of executive functions, particularly response inhibition and cognitive set shifting. In this study, we aimed to evaluate the Stroop Test Çapa Version by assessing its sensitivity and specificity in detecting executive function impairment in individuals with bipolar disorder during euthymia.

Method: In this retrospective study, 156 euthymic individuals with bipolar disorder type I and 125 healthy controls were included. Receiver Operating Characteristic (ROC) analyses were conducted separately for the completion times of the Stroop A, B, and C subtests, as well as Stroop D, calculated as the difference in reaction time between Stroop C and Stroop B. Optimal, diagnostic, and screening cut-off points were identified for each score type, along with their corresponding sensitivity, specificity, and positive and negative predictive values.

Results: Participants with bipolar disorder required significantly more time to complete all Stroop subtests compared to healthy controls ($p < 0.007$ for all comparisons). Among the subtests, Stroop C demonstrated the highest discriminative ability (AUC = 0.671; $p < 0.0001$), followed by Stroop A (AUC = 0.659; $p < 0.0001$), Stroop D (AUC = 0.649; $p < 0.0001$), and Stroop B (AUC = 0.606; $p = 0.0019$).

Discussion: Our findings indicate that the Stroop Test Çapa Version, when used alone, does not yield high sensitivity or specificity in identifying bipolar disorder. Therefore, it should be integrated with other neuropsychological assessments to enhance the clinical and cognitive evaluation of individuals with bipolar disorder.

Key Words: Bipolar disorder, executive functions, Stroop Test Çapa Version, sensitivity, specificity, discriminative validity

INTRODUCTION

Bipolar disorder (BD) is a lifelong disorder characterized by cyclic episodes of mania and depression, separated by periods of remission. It affects mood, cognitive functions, and overall medical condition, leading to impaired functioning in individuals diagnosed with the disorder. The severity of cognitive dysfunction in BD varies widely; while some individuals maintain intact cognitive abilities, others experience impairments ranging from mild to severe (1). A growing body of research supports the existence of cognitive subgroups within BD, differentiated by the level of cognitive functioning (2,3,4). The extent and progression of cognitive deficits in BD have been explored through both cross-sectional (5,6) and longitudinal studies (7,8,9). Cognitive impairment is prominent during manic and depressive episodes but also persists during the euthymic phase and is considered an endophenotype for BD (10,11,12). During euthymia, the most consistently affected domains are verbal memory, attention, processing speed, and response inhibition, typically with moderate to large effect sizes (10,12,13). Among these, impairments in verbal memory and executive functions

DOI: 10.5505/kpd.2025.92603

Cite this article as: Hidiroglu Ongun C, Emek Savas DD, Ceylan D, Ozerdem A. Discriminative validity of the Stroop Test Çapa Version for executive function deficits in bipolar disorder. Turkish J Clin Psych 2025; 28:

The arrival date of article: 23.03.2025, **Acceptance date publication:** 21.07.2025

Turkish J Clinical Psychiatry 2025;28:

appear to be the most prominent.

Cognitive impairment negatively affects the psychosocial and occupational functioning of individuals with BD (14,15,16), and it has even been discussed that measured cognitive symptoms may be a better predictor of functioning than measured emotional symptoms (16). Therefore, understanding the nature of cognitive impairment and its contributing factors in BD is essential for developing strategies to prevent cognitive decline and effective treatments.

Yatham et al. (17) conducted a study to develop a standardized research battery including validated cognitive measures for the assessment of cognitive functioning in individuals with BD. This battery, known as the International Society for Bipolar Disorders-Battery for Assessment of Neurocognition, ISBD-BANC, is considered appropriate for use in research settings or large-scale clinical trials where cognitive screening, repeated assessment of cognitive performance, or evaluation of treatment effects is required. One of the study's objectives was to summarize the cognitive domains most significantly impaired in BD and their respective measurements. The cognitive tasks and tests included in the battery were selected based on a review of existing meta-analytic studies. Furthermore, given the overlap in neuropsychological functions between BD and schizophrenia, the Consensus Cognitive Battery and its components—developed through the Measurement and Treatment Research to Improve Cognition in Schizophrenia (MATRICS) initiative for clinical research in schizophrenia—were assessed for their applicability in BD. As a result of the evaluation, in addition to the Consensus Cognitive Battery, effective tests were identified in the areas of verbal learning (e.g., California Verbal Learning Test) and executive functions (e.g., Stroop Test, Trail Making Test part B, Wisconsin Card Sorting Test). Researchers have considered that executive functions are not uniformly impaired in all patients with BD and encompass a wide range of higher-level cognitive processes. They have addressed the cognitive tasks defined for different components of executive functions based on the associated neural circuits. The Stroop test has been included among the core subtests identified for executive functions due to its short application time, ease of administration, adequate reliability levels, repeatability, and international applicability (18).

Stroop Test Çapa Version

The Stroop test, designed by John Ridley Stroop in 1935, is a neuropsychological test widely used today in both experimental research and clinical practice in Turkey and worldwide (19). This test generally assesses the ability to inhibit cognitive conflict that arises during the simultaneous processing of two different features of a stimulus, and the effort exerted during this process results in a prolonged response time to complete the task (19).

There are different forms of the Stroop test, such as the Golden form (20) and the Victoria form (21), which differ in terms of the number of stimuli, the type of stimuli, and the order in which the tasks are given (22, 23). The Stroop Test Çapa Version (23) and the Stroop Test TBAG version (24) are commonly used in Turkey. The Stroop Test TBAG version was created by combining the original Stroop and Victoria forms, standardized by calculating norm values and shown to evaluate characteristics similar to other Stroop tests (24). The Stroop Test Çapa Version, on the other hand, is an adaptation of the Stroop test form developed by Weintraub (25) at the Neuropsychology Laboratory of Istanbul University Faculty of Medicine (Çapa) (23). The Stroop Test Çapa Version has certain advantages over the TBAG Version, such as the evaluation of spontaneous corrections and error counts, having fewer subtests, being free of charge, shorter administration time, established normative values stratified by demographic variables, and a higher representational strength for the elderly population (23).

Cognitive Functions Measured by the Stroop Test

The Stroop test primarily measures response inhibition (26, 27) and has been reported to assess cognitive functions such as selective attention (28), information processing speed (29, 30), and cognitive flexibility (29) in the literature. Successful performance on the Stroop task requires the significant use of attention functions. Additionally, research findings indicate that individual differences in working memory capacity predict performance on the Stroop test (31, 32).

According to Periáñez et al. (33), the subtests of the Stroop test are related to different cognitive functions. For example, the task of reading color names reflects visual scanning speed, while the task of naming colors reflects both working memory and visual scanning speed. The final subtest, which requires participants to name the colors of color names printed in a different color ink, relates not only to these functions but also to the process of conflict monitoring.

The Discriminative Validity of the Stroop Test in Neuropsychiatric Disorders

The Stroop test is considered valid and reliable across various cultures, and its normative values have been established (e.g., 34, 35). However, there is a lack of studies focusing on the sensitivity and specificity of the test. Sensitivity refers to a test's ability to accurately identify individuals who truly exhibit the trait it aims to measure, while specificity pertains to the test's proficiency in recognizing those who do not possess that trait. In the context of neuropsychological tests, a test that achieves high levels of sensitivity and specificity effectively distinguishes between individuals with and without cognitive impairments. These metrics are essential for evaluating the ability of neuropsychological assessments to reflect individuals' real-world performance and reliability. Reliable tests improve data quality in clinical settings and neuropsychological research. According to Sørensen et al. (36), the Stroop interference score of errors distinguishes children diagnosed with attention deficit hyperactivity disorder (ADHD) from children with different diagnoses and children with typical development and predicts impulse control in their daily lives as reported by their parents. This study is the first to show that the number of errors made during this trial is more specific to ADHD diagnosis than a score based on interference time. Other studies have also reported that the Stroop test has good discriminatory value in distinguishing individuals with an ADHD diagnosis from healthy individuals (37). On the other hand, Homack and Riccio's (38) meta-analysis study found that the Stroop test did not show specificity in distinguishing children and adolescents with ADHD diagnoses from other clinical groups, such as learning disabilities, autism, and Tourette syndrome. Thus, the specificity of the Stroop test for ADHD is unclear, and it is insufficient for ADHD diagnosis.

Lubrini et al. (39) reported that individuals with traumatic brain injury (TBI) and schizophrenia scored lower than healthy controls on all test conditions in the Stroop test, demonstrating the test's ability to distinguish these individuals from healthy individuals. The Serial Color-Word Test has been shown to be successful in classifying severe illnesses such as schizophrenia and BD, but it has not been able to successfully distinguish individuals with schizophrenia from those with BD in all cases (40). The Stroop test was examined for its sensitivity and specificity in cognitive screening among a sample of elderly adults with severe psychiatric disorders, but its effectiveness in this group was found to be lacking (41). The researchers concluded that more studies are necessary to assess whether the Stroop test could be useful in clinical settings, especially considering its short administration time and strong psychometric properties.

As can be seen from the research results presented above, the sensitivity of the Stroop test may be stronger than its specificity (20). Consistent with this view, Stroop test performance has been similarly impaired across different age ranges in many neuropsychiatric disorders associated with functional and structural changes in the brain, particularly in the frontal areas (38). Snyder, Miyake, and Hankin (42) and Zelazo (43) also note that executive function impairments are a transdiagnostic marker of atypical development and various psychiatric disorders, emphasizing the need for better assessment of executive functions in these disorders.

In studies conducted in Turkey for various purposes using different forms of the Stroop test, individuals diagnosed with BD demonstrated poorer performance compared to healthy controls (44,45). On the other hand, some studies have found no statistically significant differences in Stroop performance between individuals with BD and healthy control groups (46,47). No studies have examined the sensitivity and specificity of the Stroop test in BD.

Purpose of the Study

Neuropsychological assessment is considered a crucial component in the diagnostic evaluation and monitoring of mood disorders, such as BD. However, there are currently no established gold-standard tools for assessment (48). Consequently, it is essential to identify valid and reliable tests that can effectively differentiate individuals with BD from those who are healthy. These tests may significantly contribute to the early detection of the disorder, aid in diagnosing cognitive impairment accurately, and assist in cognitive rehabilitation planning. In particular, further research is necessary to evaluate the sensitivity and specificity of neuropsychological tests assessing executive functions, which are known to be significantly affected in BD.

The current study aims to evaluate the sensitivity and specificity of the Stroop Test Çapa Version (23), which is commonly used to assess executive functions in Turkey. It specifically aims to determine how effectively this test identifies executive function impairments in individuals diagnosed with BD during the remission period, and how it differentiates these individuals from healthy participants.

METHOD

Study Design and Sample

This retrospective study has received approval from the Koç University Clinical Research Ethics Committee under protocol number 2024.351.IRB2.150. Data were collected from 156 individuals diagnosed with BD type I and 125 healthy controls. These participants were involved in two large-scale research projects conducted separately in 2007 and 2018 at the Department of Neuroscience, Institute of Health Sciences, Dokuz Eylül University. The results of these comprehensive studies, which utilized this data, have been previously published (49, 50). Participants with BD were recruited from the Department of Psychiatry at Dokuz Eylül University Hospital, while the control group was formed through announcements circulated among students and staff at Dokuz Eylül University.

The inclusion criteria applicable to both studies are: having a diagnosis of bipolar disorder type I according to the Diagnostic and Statistical Manual of Mental Disorders-IV-TR (DSM-IV-TR) diagnostic criteria, being in the remission phase for at least 6 months [(Hamilton Depression Scale (HAM-D-17) and Young Mania Rating Scale (YMRS) scores $\leq$ 7)] and not having received any other psychiatric Axis I diagnosis other than BD. The healthy control sample comprises individuals identified as having no psychiatric disorders, as assessed through the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I) in related projects. Furthermore, the control group's subjective complaints were evaluated, and only those without any concerns regarding cognitive functions were selected. Individuals with hearing and visual impairments that could interfere with neuropsychological assessments, as well as those with mental retardation, degenerative neurological diseases, cerebrovascular diseases, brain surgery, epilepsy, cerebral tumors, or a history of head trauma resulting in loss of consciousness, were excluded from the study. Additionally, individuals who had a history of alcohol or substance abuse within the year prior to the studies were also excluded. All participants in the BD diagnosis group are receiving one or more drug treatments. Detailed demographic characteristics and Stroop Test Çapa Version scores for all participants aged 18 to 65 are presented in Table 1.

Data Collection Tools

The Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I): SCID-I is a semi-structured interview technique created by First et al. (51) to assess first-axis psychiatric disorders based on DSM-IV criteria. A study validating and testing its reliability in Turkish was conducted by Özkürkçügil et al. (52).

Table 1. Demographic Characteristics and Stroop Test Capa Version Scores of Study Participants

	Participants with BD (n=156)	Healthy Controls (n=125)	*p*
Age (Years)	35.63 – 9.92	33.33 – 10.87	0.065
Gender (M/F)	63/93	49/76	0.840
Education (Years)	12.65 – 3.59	13.31 – 3.66	0.144
Stroop A	40.92 – 10.66	35.63 – 6.61	<0.001
Stroop B	30.76 – 7.30	28.58 – 6.93	0.006
Stroop C	76.94 – 23.10	65.80 – 18.66	<0.001
Stroop D	46.40 – 20.02	37.22 – 15.54	<0.001

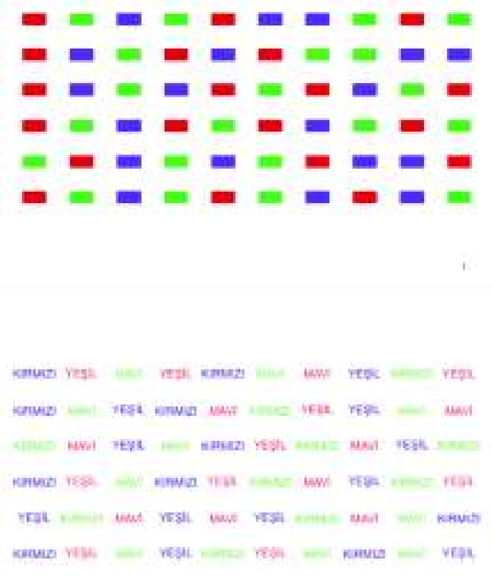
Values are presented as mean – standard deviation. The completion times of Stroop A, B, and C subtests and the calculated Stroop D time are in seconds. Age, education, and completion times were analyzed with independent samples t-tests, while gender was assessed using a Pearson x^2 test. BD: Bipolar Disorder, M: Male, F: Female.

Hamilton Depression Rating Scale-17 (HDRS-17): Max Hamilton (53) developed the 17-item HDRS, which clinicians frequently use to assess the presence or absence of depressive symptoms and their severity at mild, moderate, or severe levels. The scale consists of 17 items and assesses depressive symptoms experienced in the past week. Its total score ranges from 0 to 53. Akdemir et al. (54) conducted the Turkish validity and reliability study.

The Young Mania Rating Scale (YMRS): YMRS was developed by Young et al. (55) to assess the severity of manic symptoms. This scale is based on both interviews and observations and consists of a total of 11 items. Among these, seven items use a five-point Likert scale, while the remaining four use a nine-point Likert scale. Scoring is determined by the clinician's observations during the interview and the patient's self-reported symptoms from the past 48 hours. The total score on the scale ranges from 0 to 60. A Turkish validity and reliability study was conducted by Karadağ et al. (56) in 2001.

Stroop Test Çapa Version: The validity, reliability, and normative study was conducted by Emek Savaş et al. (23) with a large sample size of 541 participants. The test includes two stimulus cards (see Figure 1) arranged in a 6×10 format, totaling 60 items, and an application form for recording responses. The first card features rectangular boxes in red, green, and blue colors; the second card features color names printed in incongruent colors (e.g., the word "green" printed in blue ink). The test is divided into three sequential subtests. Table 2 provides information about these subtests, the tasks for each section, and the types of scores recorded.

In the literature, the Stroop test's "interference" time and "resistance to interference" time are utilized to evaluate challenges in response inhibition, which typically refers to the suppression of behaviors that are incompatible with a given task. The Stroop Test "interference" time is typically expressed as the time spent in trials where the colors of ink printed in incongruent colors are named, while the Stroop test "resistance to interference" time is based on a calculation derived from the difference between the times spent in different parts of the test. The equivalents of these two scores in the Stroop Test Çapa Version are Stroop C and Stroop D scores, respectively. According to Emek Savaş et al. (23), the interference part of the Stroop test, as a distinct function from the other parts, evaluates the suppression of automatic responses and the generation of unfamiliar new responses. Therefore, the Cronbach's alpha coefficient calculated for all subtests was higher when the interference part was not included. In addition, it has been shown that the test-retest reliability is high



Resource: Emek Sava DD, Yerlikaya D, G Yener G, ktem Tan r . Validity, reliability and normative data of the Stroop Test ˙apa Version. Turk Psikiyatri Derg 2020; 31(1):9 -21.

Figure 1. Stroop Test Çapa Version Stimulus Cards.

Table 2. The Stroop Test Capa Version Subtests, Tasks, and Score Types

Stimulus card	Subtests	Tasks	Score Types
1	Stroop A	The individual is asked to name the colors of the small rectangular boxes in the order shown.	Completion time
2	Stroop B	The individual is asked to read color names printed in incongruent colors in the order shown.	Completion time
2	Stroop C	The individual is asked to name the color of the colored words printed in incongruent colors, following the order shown.	Completion time, number of errors and spontaneous corrections
Stimulus card	Subtest	Calculation	Calculated score type
-	Stroop D (resistance to interference time)	The difference between the completion times for Stroop C and Stroop B subtests is calculated (Stroop D=Stroop C-Stroop B).	Time

and/or sufficient for the 18-49 age group and the 50 and older age group, which were examined separately. Test-retest reliability coefficients for Stroop A, B, C, and D subtests ranged from 0.67 to 0.88 for individuals aged 18-49 and from 0.64 to 0.84 for individuals aged 50 and older (23). Furthermore, correlations calculated between Stroop C completion times and the Trail Making Test part A completion time (r = 0.60), part B completion time (r = 0.65), and the B-A time difference (r = 0.57) were found to be moderate for the concurrent validity of the form. These concurrent validity and reliability values indicate that the Stroop Test Çapa Version has strong psychometric properties.

Procedure

In both studies that used the data (49, 50), individuals diagnosed with BD and healthy participants who met the inclusion criteria were provided with comprehensive information about the research and gave written informed consent. Clinical interviews with participants were conducted at the Department of Psychiatry at Dokuz Eylül University Hospital, while neuropsychological testing took place at the Department of Neuroscience within the Institute of Health Sciences at Dokuz Eylül University. Trained psychiatrists conducted all SCID-I interviews, and the diagnoses of individuals with BD and the absence of any other concurrent Axis I psychiatric diagnosis were confirmed through these interviews. Clinical scales were administered, and individuals who scored 7 or below on these scales and had been in remission for at least six months were referred for a neuropsychological assessment after their demographic information was collected. Similarly, healthy controls, determined to have no psychiatric disorders through SCID-I interviews, were also referred for neuropsychological assessment following the collection of their demographic details. The neuropsychological assessment included tests evaluating attention, memory, executive functions, visual-spatial functions, and language skills. Experienced neuropsychologists administered the tests to both groups in a single session, following a standardized sequence. Psychiatrists and neuropsychologists were informed about the purpose of the study. The data from both studies described above were combined for this study, and individuals with complete Stroop Test Çapa Version scores formed the study sample.

Statistical Analyses

The groups' age, education, and Stroop Test Çapa Version subtest completion times were compared using an independent samples t-test, and the gender variable was compared using a Pearson χ^2 test (see Table 1). The study's dependent variable is a binary variable indicating the diagnostic group (bipolar disorder = 1; healthy control = 0). The independent variables are continuous variables, including the completion times for the Stroop A, Stroop B, and Stroop C subtests of the Stroop Test Çapa Version and the calculated Stroop D time. The normality assumption was assessed using histograms, Q-Q plots, and skewness-kurtosis coefficients, revealing a right-skewed distribution in the completion times for all subtests. The Levene test was conducted to examine the equality of variances. In cases where this assumption was violated, the Welch's t-test was applied to the subtests.

To evaluate the discriminative ability of the Stroop Test Çapa Version in BD, receiver operating characteristic (ROC) curves and the area under the

Table 3. Cut-off points for the Stroop Test Capa Version subtests for healthy controls and participants with bipolar disorder

	Optimal Cut-off Point	Diagnostic Cut-off Point	Screening Cut-off Point	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	AUC (C.I.)
Stroop A	>38			51.28	72.80	70.02	54.5	0.659 (0.600-0.714)
		>40		41.03	80.00	71.9	52.1	
			>32	82.05	35.20	61.2	61.1	
Stroop B	>27			66.03	55.20	64.8	56.6	0.606 (0.546-0.663)
		>33		28.21	80.80	64.7	47.4	
			>24	83.33	28.80	59.4	58.1	
Stroop C	>74			48.08	79.20	74.3	55.0	0.671 (0.613-0.726)
		>75		46.79	80.00	74.5	54.6	
			>59	80.77	40.00	62.7	62.5	
Stroop D	>41			53.85	71.20	70.0	55.3	0.649 (0.590-0.705)
		>45		41.03	80.00	71.9	52.1	
			>30	81.41	31.20	59.6	57.4	

PPV: Positive predictive value, NPV: Negative predictive value, AUC: Area under the curve, C.I.: Confidence interval

curve (AUC) with a 95% confidence interval were calculated for each independent variable. For each metric, the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated. The PPV represents the likelihood that individuals identified as having bipolar disorder by the test indeed possess the disorder, whereas the NPV indicates the probability that individuals categorized as negative by the test are genuinely healthy. Furthermore, optimal, diagnostic, and screening cut-off points were established for each metric. The optimal cut-off point was defined as the value at which sensitivity and specificity intersect, utilizing the Youden Index for this determination. The Youden Index (J = sensitivity + specificity − 1) is a measure that ranges from 0 to 1 and is calculated for each possible cut-off point on the ROC curve. A J value of 0 indicates no discriminative power, while a J value of 1 indicates perfect classification. The optimal cut-off point is identified as the point farthest from the main diagonal of the ROC curve, which represents random classification and corresponds to the highest J value. The diagnostic cut-off point is defined as the threshold at which specificity reaches 80% or higher. Conversely, the screening cut-off point is established when sensitivity reaches 80% or higher. Descriptive statistics and group comparisons were conducted using SPSS 29 software, while ROC analyses were carried out using MedCalc 23.1.1 software. A two-tailed p<0.05 was accepted as the significance level for all statistical tests.

RESULTS

In the study, individuals diagnosed with BD and healthy controls exhibited comparable characteristics regarding age (p=0.065), gender (p=0.840), and education (p=0.144). It was found that the BD group took significantly longer to complete all subtests of the Stroop Test Çapa Version compared to the control group (p<0.007 for all; Table 1).

The discriminant validity of the Stroop Test Çapa Version was assessed through ROC analyses. The highest AUC value for differentiating the BD group from healthy controls was observed in the Stroop C subtest (AUC = 0.671; p < 0.0001). This was followed by Stroop A (AUC = 0.659; p < 0.0001), Stroop D (AUC = 0.649; p < 0.0001), and Stroop B (AUC = 0.606; p = 0.0019). The AUC represents the probability that the test duration of an individual randomly selected with a BD diagnosis exceeds that of a randomly selected healthy individual. AUC values are interpreted as follows: an AUC of 0.50 indicates no discriminative power (random chance); 0.60-0.70 reflects weak to moderate discriminative ability; 0.70-0.80 signifies good discrimination; 0.80-0.90 suggests very good discriminative capacity; and an AUC of ≥ 0.90 denotes excellent discriminative ability. For instance, an AUC value of 0.66 implies that the test demonstrates moderate discriminatory power, indicating a 66% probability that an individual with a BD diagnosis will have a more extended test duration than a healthy individual. While this level of discrimination is significantly better than chance (50%), it has not yet reached a clinically excellent standard.

ROC curves for individuals with BD and healthy controls are shown in Figure 2. The optimal, diagnostic, and screening cut-off points determined for each subtest are presented in Table 3 along with sensitivity, specificity, PPV, and NPV values.

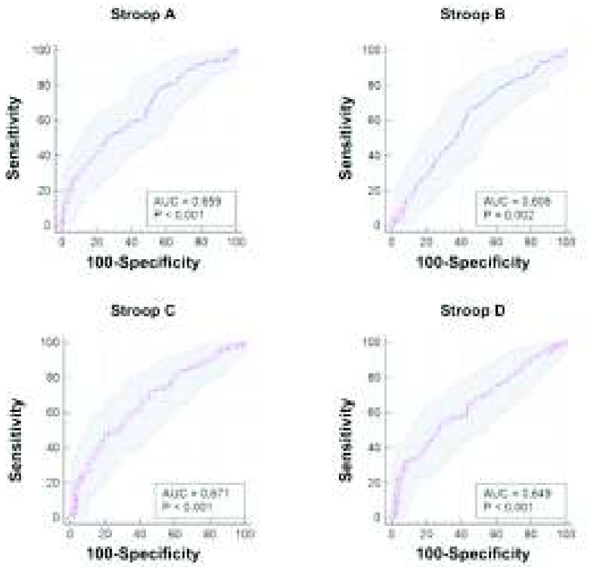


Figure 2. ROC curves comparing participants with bipolar disorder to healthy controls.

The cut-off points presented in Table 3 summarize the diagnostic performance across various clinical scenarios. For instance, the optimal cut-off point for the Stroop C subtest is greater than 74 seconds. At this threshold, the sensitivity is 48%, the specificity is 79%, the PPV is 75%, and the NPV is 55%. This indicates that the test correctly identifies nearly half of the individuals with BD, and three out of four individuals who test positive are confirmed BD cases. In scenarios where the specificity is at least 80%, adjusting the cut-off point to greater than 75 seconds leads to a decrease in sensitivity to 47%. While this adjustment minimizes the risk of misdiagnosing healthy individuals, it unfortunately leads to the underdiagnosis of half of the BD cases. In situations that require high sensitivity for screening, setting the threshold to greater than 59 seconds increases the sensitivity to 81%. However, this raises the risk of reducing specificity to 40%. The Stroop A and Stroop D subtests demonstrate a similar trade-off between sensitivity and specificity, while Stroop B consistently shows lower sensitivity, remaining below 50%. The other panels (Stroop A, B, and D) in Table 3 should be interpreted with similar consideration.

DISCUSSION

The aim of this study was to assess the sensitivity and specificity of the Stroop Test Çapa Version for detecting executive function impairment in euthymic individuals with BD and differentiating them from healthy controls.

The findings show that the BD group completed all subtests of the Stroop Test Çapa Version (Stroop A, B, C, and D) longer than healthy controls. The AUC values obtained from ROC analyses were above 0.6, indicating an acceptable level. This result demonstrates that the performance of individuals diagnosed with BD differs significantly from that of healthy individuals regarding executive functions and that the Stroop Test Çapa Version has a certain discriminative power in distinguishing between BD and healthy individuals. These findings are consistent with existing results indicating that cognitive functions are impaired in bipolar disorder when compared to healthy individuals, and that this cognitive impairment can be observed before the onset of the disease, at the onset of the disease, and during periods of remission throughout the disease (10). Furthermore, most studies conducted during the remission phase indicate that the impact of medication on ongoing cognitive impairment during this phase is minimal (14,57,58).

The highest AUC value was observed for the Stroop C (interference) subtest in the study. The literature indicates that the Stroop C subtest demands greater executive functioning and response inhibition because it involves color names printed in incongruent ink colors (23). Considering that BD is particularly associated with difficulties in emotion regulation and cognitive flexibility, it is consistent that Stroop C is more sensitive in distinguishing the BD-diagnosed group from healthy controls.

The fact that Stroop A and D subtests also have significant AUC values indicates that color naming and resistance to interference times also have partial discriminatory power. However, the low AUC value observed for Stroop B suggests that the task of reading color names requires relatively less cognitive effort. As a result, it may provide lower discriminative power in distinguishing between BD cases and healthy controls.

The results of optimal, diagnostic, and screening cut-off points show that Stroop test times alone cannot achieve high sensitivity and specificity values. This indicates that the test may serve as an auxiliary tool in a general screening or clinical evaluation process rather than as a diagnostic instrument. However, identifying different cut-off points targeting ≥80% specificity or ≥80% sensitivity may guide clinicians in diagnostic processes requiring high specificity or screening purposes requiring high sensitivity. Furthermore, given BD's complex clinical picture and heterogeneous nature, it is beneficial to evaluate the Stroop Test Çapa Version as part of a comprehensive neuropsychological assessment battery rather than as a standalone determinant. These findings are consistent with Golden's (20) view that the Stroop test can be used as part of a larger battery to screen for brain dysfunction.

To our knowledge, there are currently no studies in Turkey that examine the sensitivity and specificity of the Stroop test in individuals diagnosed with BD, which can be used to compare with the findings of the present study. On the other hand, the Stroop test is used in numerous studies examining cognitive impairment as an endophenotypic marker for BD (12). Recent discussions and studies have focused on whether cognitive impairment in BD is progressive, often referred to as cognitive neuroprogression (59). Researchers are also examining subgroups of individuals with cognitive differences in BD, highlighting the heterogeneity of etiological and genetic risk factors associated with the disorder (2,3,60). As a result, investigating tests that effectively measure cognitive functions has become increasingly important. For instance, a recent meta-analysis by Bora (60) identified three cognitive subtypes of BD. Approximately one-third of individuals with BD demonstrate good overall functioning and cognitive performance, exhibiting a slight increase in executive function compared to healthy control groups. The same meta-analysis showed that approximately one-quarter of these individuals experienced impairment comparable to that seen in individuals diagnosed with schizophrenia in six cognitive domains, including executive functions. Based on these results, the use of appropriate and valid tests when assessing executive functions and their subcomponents in BD is critical, both because of the significant impact of these functions on psychosocial functioning and daily life (61) and for the identification and determination of cognitive subgroups.

The findings of this study support the applicability, validity, and reliability of the Stroop Test Çapa Version (23) for neuropsychological assessment in individuals with BD. Key strengths of the study include the careful matching of the BD group to healthy control participants based on age, education, and gender, as well as a substantial sample size. However, several limitations should be noted. The retrospective design of the study, along with the broad time span over which participants were evaluated, may have restricted the collection of detailed information regarding the disease progression. In addition, the use of the fourth edition (DSM-IV) of the Diagnostic and Statistical Manual of Mental Disorders, which was valid at the time the data were collected, rather than the current fifth edition (DSM-5) for diagnostic interviews and inclusion criteria, may be considered a limitation. Furthermore, the lack of comparisons with other executive function tests beyond the Stroop Test Çapa Version somewhat limits the test's specificity. Another limitation of the study is that individuals with BD were not compared with participants having other neuropsychiatric diagnoses, making it challenging to assert that the results are exclusive to BD. In addition, participants with BD were receiving pharmacological treatment and constituted a highly heterogeneous group in terms of medication types and monotherapy versus polytherapy. The neuroprotective or neurotoxic effects of these medications on cognitive functions remain a topic of considerable debate (62,63). The failure to assess medication effects due to the diversity of drug use in the sample group further constitutes a limitation of this study.

Future research should integrate neuropsychological test batteries with neuroimaging techniques to provide a more comprehensive understanding of cognitive processes in BD. Studies focusing on different age groups or illness stages may help establish more specific cut-off points for the Stroop Test Çapa Version.

In conclusion, the Stroop Test Çapa Version demonstrates a notable level of discriminant validity in differentiating individuals diagnosed with BD Type I from healthy controls. Specifically, the Stroop C subtest appears to be more sensitive than the other subtests in reflecting the cognitive differences associated with BD, particularly regarding cognitive interference and response inhibition processes. However, rather than relying on the Stroop test as a standalone diagnostic tool, it is recommended to incorporate it into a comprehensive neuropsychological assessment. For instance, the Stroop test may be employed in longitudinal studies examining cognitive functioning in individuals at high risk for BD, as well as in large-scale research comparing cognitive subgroups within BD in terms of cognitive efficiency. The present study holds particular relevance for neuropsychologists and clinicians in Turkey who use the Stroop Test Çapa Version, as it contributes to a better understanding of the test's sensitivity and specificity. Such knowledge supports the appropriate selection of assessment tools and promotes accurate interpretation of results.

Correspondence address: Assis. Prof., Ceren Hidiroglu Ongun, Dokuz Eylül University Faculty of Letters, Department of Psychology, Experimental Psychology Program, Izmir,Turkey
cerennhi@gmail.com

REFERENCES

1. Green MJ, Girshkin L, Kremerskothen K, Watkeys O, Quidé Y. A systematic review of studies reporting data-driven cognitive subtypes across the psychosis spectrum. Neuropsychol Rev 2020; 30(4):446-460. doi: 10.1007/s11065-019-09422-7.

2. Burdick KE, Russo M, Frangou S, Mahon K, Braga RJ, Shanahan M, Malhotra AK. Empirical evidence for discrete neurocognitive subgroups in bipolar disorder: clinical implications. Psychol Med 2014; 44(14):3083-96. doi: 10.1017/S0033291714000439.

3. Bora E, Hıdıroğlu C, Özerdem A, Kaçar ÖF, Sarısoy G, Civil Arslan F, Aydemir Ö, Cubukcuoglu Tas Z, Vahip S, Atalay A, Atasoy N, Ateşci F, Tümkaya S. Executive dysfunction and cognitive subgroups in a large sample of euthymic patients with bipolar disorder. Eur Neuropsychopharmacol 2016; 26(8):1338-47. doi: 10.1016/j.euroneuro.2016.04.002.

4. Jensen JH, Knorr U, Vinberg M, Kessing LV, Miskowiak KW. Discrete neurocognitive subgroups in fully or partially remitted bipolar disorder: Associations with functional abilities. J Affect Disord 2016; 205:378-386. doi: 10.1016/j.jad.2016.08.018.

5. Robinson LJ, Ferrier IN. Evolution of cognitive impairment in bipolar disorder: a systematic review of cross-sectional evidence. Bipolar Disord 2006; 8(2):103-16. doi: 10.1111/j.1399-5618.2006.00277.x.

6. Jones BDM, Fernandes BS, Husain MI, Ortiz A, Rajji TK, Blumberger DM, Butters MA, Gildengers AG, Shablinski T, Voineskos A, Mulsant BH. A cross-sectional study of cognitive performance in bipolar disorder across the lifespan: the cog-BD project. Psychol Med 2023; 53(13):6316-6324. doi: 10.1017/S0033291722003622.

7. Samamé C, Martino DJ, Strejilevich SA. Longitudinal course of cognitive deficits in bipolar disorder: a meta-analytic study. J Affect Disord 2014; 164:130-8. doi: 10.1016/j.jad.2014.04.028.

8. Bora E, Özerdem A. Meta-analysis of longitudinal studies of cognition in bipolar disorder: comparison with healthy controls and schizophrenia. Psychol Med 2017; 47(16):2753-2766. doi: 10.1017/S0033291717001490.

9. Torres IJ, Qian H, Basivireddy J, Chakrabarty T, Wong H, Lam RW, Yatham LN. Three-year longitudinal cognitive functioning in patients recently diagnosed with bipolar disorder. Acta Psychiatr Scand 2020; 141(2):98-109. doi: 10.1111/acps.13141.

10. Torres IJ, Boudreau VG, Yatham LN. Neuropsychological functioning in euthymic bipolar disorder: a meta-analysis. Acta Psychiatr Scand Suppl 2007; (434):17-26. doi: 10.1111/j.1600-0447.2007.01055.x.

11. Arts B, Jabben N, Krabbendam L, van Os J. Meta-analyses of cognitive functioning in euthymic bipolar patients and their first-degree relatives. Psychol Med 2008; 38(6):771-85. doi: 10.1017/S0033291707001675.

12. Bora E, Yucel M, Pantelis C. Cognitive endophenotypes of bipolar disorder: a meta-analysis of neuropsychological deficits in euthymic patients and their first-degree relatives. J Affect Disord 2009; 113(1-2):1-20. doi: 10.1016/j.jad.2008.06.009.

13. Robinson LJ, Thompson JM, Gallagher P, Goswami U, Young AH, Ferrier IN, Moore PB. A meta-analysis of cognitive deficits in euthymic patients with bipolar disorder. J Affect Disord 2006; 93(1-3):105-15. doi: 10.1016/j.jad.2006.02.016.

14. Martínez-Arán A, Vieta E, Colom F, Torrent C, Sánchez-Moreno J, Reinares M, Benabarre A, Goikolea JM, Brugué E, Daban C, Salamero M. Cognitive impairment in euthymic bipolar patients: implications for clinical and functional outcome. Bipolar Disord 2004; 6(3):224-32. doi: 10.1111/j.1399-5618.2004.00111.x.

15. Torres IJ, deFreitas CM, Yatham LN. Cognition and Functional Outcome in Bipolar Disorder In: Cognitive Dysfunction in Bipolar Disorder: A Guide for Clinicians. Edited by Goldberg JF, Burdick KE. Washington, DC, American Psychiatric Press, 2008, pp. 217–234.

16. Baune BT, Malhi GS. A review on the impact of cognitive dysfunction on social, occupational, and general functional outcomes in bipolar disorder. Bipolar Disord 2015; 17 Suppl 2:41-55. doi: 10.1111/bdi.12341.

17. Yatham LN, Torres IJ, Malhi GS, Frangou S, Glahn DC, Bearden CE, Burdick KE, Martínez-Arán A, Dittmann S, Goldberg JF, Ozerdem A, Aydemir O, Chengappa KN. The International Society for Bipolar Disorders-Battery for Assessment of Neurocognition (ISBD-BANC). Bipolar Disord 2010; 12(4):351-63. doi: 10.1111/j.1399-5618.2010.00830.x.

18. Van Rheenen TE, Rossell SL. An empirical evaluation of the MATRICS Consensus Cognitive Battery in bipolar disorder. Bipolar Disord 2014; 16(3):318-25. doi: 10.1111/bdi.12134.

19. Stroop JR. Studies of Interference in Serial Verbal Reactions. J Exp Psychol 1935; 18:643-662. https://doi.org/10.1037/h0054651

20. Golden CJ. Stroop Color and Word Test: A Manual for Clinical and Experimental Uses. Chicago, Stoelting Company, 1978.

21. Regard M. Cognitive Rigidity and Flexibility: A Neuropsychological Study. University of Victoria, Doctoral Dissertation.1981.

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

23. Emek Savaş DD, Yerlikaya D, G Yener G, Öktem Tanör Ö. Validity, reliability and normative data of the Stroop Test Çapa Version. Turk Psikiyatri Derg 2020; 31(1):9-21. doi: 10.5080/u23549.

24. Karakaş S, Erdoğan E, Sak L, Soysal AŞ, Ulusoy T, Yüceyurt Ulusoy İ, Alkan S. Stroop Testi TBAG Formu: Türk kültürüne standardizasyon çalışmaları, güvenirlik ve geçerlik. Klinik Psikiyatri Dergisi 1999; 2(2): 75-88.

25. Weintraub S. Neuropsychological Assessment of Mental State. Principles of Cognitive and Behavioral Neurology. Edited by Mesulam MM. New York, NY, Oxford University Press, 2000, pp. 130.

26. MacLeod CM. Half a century of research on the Stroop effect: an integrative review. Psychol Bull 1991; 109(2):163-203. doi: 10.1037/0033-2909.109.2.163.

27. Friedman NP, Miyake A. The relations among inhibition and interference control functions: a latent-variable analysis. J Exp Psychol Gen 2004; 133(1):101-135. doi: 10.1037/0096-3445.133.1.101.

28. MacLeod CM. The Stroop task: The ‘gold standard’ of attentional measures. J Exp Psychol Gen 1992; 121:12-14

29. Golden CJ, Freshwater SM. The Stroop Color and Word Test: A Manual for Clinical and Experimental Uses. Chicago, IL, Stoelting Company, 2002.

30. Bora E, Pantelis C. Domains of cognitive impairment in bipolar disorder: commentary on "The International Society for Bipolar Disorders-Battery for Assessment of Neurocognition (ISBD-BANC)". Bipolar Disord 2011; 13(2):217-8. doi: 10.1111/j.1399-5618.2011.00899.x.

31. Long DL, Prat CS. Working memory and stroop interference: an individual differences investigation. Mem Cognit 2002; 30(2):294-301. doi: 10.3758/bf03195290.

32. Kane MJ, Engle RW. Working-memory capacity and the control of attention: the contributions of goal neglect, response competition, and task set to Stroop interference. J Exp Psychol Gen 2003; 132(1):47-70. doi: 10.1037/0096-3445.132.1.47.

33. Periáñez JA, Lubrini G, García-Gutiérrez A, Ríos-Lago M. Construct validity of the Stroop Color-Word Test: Influence of speed of visual search, verbal fluency, working memory, cognitive flexibility, and conflict monitoring. Arch Clin Neuropsychol 2021; 36(1):99-111. doi: 10.1093/arclin/acaa034.

34. Van der Elst W, Van Boxtel MP, Van Breukelen GJ, Jolles J. The Stroop color-word test: influence of age, sex, and education; and normative data for a large sample across the adult age range. Assessment 2006; 13(1):62-79. doi: 10.1177/1073191105283427.

35. Moering RG, Schinka JA, Mortimer JA, Graves AB. Normative data for elderly African Americans for the Stroop Color and Word Test. Arch Clin Neuropsychol 2004; 19(1):61-71.

36. Sørensen L, Plessen KJ, Adolfsdottir S, Lundervold AJ. The specificity of the Stroop interference score of errors to ADHD in boys. Child Neuropsychol 2014; 20(6):677-91. doi: 10.1080/09297049.2013.855716.

37. Unal M, O'Mahony E, Dunne C, Meagher D, Adamis D. The clinical utility of three visual attention tests to distinguish adults with ADHD from normal controls. Riv Psichiatr 2019; 54(5):211-217. doi: 10.1708/3249.32185.

38. Homack S, Riccio CA. A meta-analysis of the sensitivity and specificity of the Stroop Color and Word Test with children. Arch Clin Neuropsychol 2004; 19(6):725-43. doi: 10.1016/j.acn.2003.09.003.

39. Lubrini G, Periañez JA, Rios-Lago M, Viejo-Sobera R, Ayesa-Arriola R, Sanchez-Cubillo I, Crespo-Facorro B, Álvarez-Linera J, Adrover-Roig D, Rodriguez-Sanchez JM. Clinical Spanish norms of the Stroop test for traumatic brain injury and schizophrenia. Span J Psychol 2014; 17:E96. doi: 10.1017/sjp.2014.90.

40. Hentschel U, Rubino IA, Bijleveld C. Differentiating clinical groups using the serial color-word test (S-CWT). Percept Mot Skills 2011; 112(2):629-38. doi: 10.2466/05.09.13.PMS.112.2.629-638.

41. Mackin RS, Ayalon L, Feliciano L, Areán PA. The sensitivity and specificity of cognitive screening instruments to detect cognitive impairment in older adults with severe psychiatric illness. J Geriatr Psychiatry Neurol 2010; 23(2):94-9. doi: 10.1177/0891988709358589.

42. Snyder HR, Miyake A, Hankin BL. Advancing understanding of executive function impairments and psychopathology: bridging the gap between clinical and cognitive approaches. Front Psychol 2015; 6:328. doi: 10.3389/fpsyg.2015.00328.

43. Zelazo PD. Executive function and psychopathology: A neurodevelopmental perspective. Annu Rev Clin Psychol 2020; 16:431-454. doi: 10.1146/annurev-clinpsy-072319-024242.

44. Levent N, Tümkaya S, Ateşçi F, Tüysüzoglu H, Varma G, Oguzhanoglu N. Bipolar bozukluk ve erişkin dikkat eksikliği hiperaktivite bozukluğunun nöropsikolojik açıdan karşılaştırılması. Turk Psikiyatri Derg 2014; 25(1):1-8.

45. Çökmüş FP, Aşçıbaşı K, Dikici DS, Çöldür EÖ, Avcı E,

Aydemir Ö. Bipolar bozuklukta mevsimsellik: Duygudurum belirtilerine, psikososyal işlevselliğe, nörokognisyon ve biyolojik ritim üzerine etkisi. Noro Psikiyatr Ars 2021; 58(1):41-47.

46. Duman T, Ateşçi F, Topak OZ, Şendur İ, Tümkaya S, Özdel O. Bipolar bozukluk hastaları ve birinci derece yakınlarında zihin kuramı ve yürütücü işlevler. J Clin Psy 2019; 22(4): 396-407 doi:10.5505/kpd.2019.78942

47. Aydemir Ö, Kaya E. Bipolar bozuklukta öznel bilişsel değerlendirme neyi ölçüyor? Nesnel bilişsel değerlendirme ile bağıntısı. Turk Psikiyatri Derg 2009; 20(4):332-338

48. Rossetti MG, Girelli F, Perlini C, Brambilla P, Bellani M. A critical overview of tools for assessing cognition in bipolar disorder. Epidemiol Psychiatr Sci 2022;31:e70. doi: 10.1017/S2045796022000555.

49. Hıdıroğlu C, Demirci Esen Ö, Tunca Z, Neslihan Gűrz Yalçìn S, Lombardo L, Glahn DC, Özerdem A. Can risk-taking be an endophenotype for bipolar disorder? A study on patients with bipolar disorder type I and their first-degree relatives. J Int Neuropsychol Soc 2013; 19(4):474-82. doi: 10.1017/S1355617713000015.

50. Ceylan D, Akdede BB, Bora E, Aktener AY, Hıdıroğlu Ongun C, Tunca Z, Alptekin K, Özerdem A. Neurocognitive functioning during symptomatic states and remission in bipolar disorder and schizophrenia: A comparative study. Psychiatry Res 2020; 292:113292. doi: 10.1016/j.psychres.2020.113292.

51. First MB. Structured Clinical Interview for DSM-IV Axis I Disorders: SCID-I: Clinician Version: Administration Booklet. Washington, DC, American Psychiatric Press, 1997.

52. Özkürkçügil A, Aydemir Ö, Yıldız M. DSM-IV Eksen I Bozuklukları İçin yapılandırılmış klinik görüşmenin Türkçe'ye uyarlanması ve güvenilirlik çalışması. İlaç ve Tedavi Dergisi 1999; 12: 233-236.

53. Hamilton M. A rating scale for depression. J Neurol Neorusurg Psychiatry 1960; 23: 56-62.

54. Akdemir A, Örsel S, Dağ I, Türkçapar HM, İşcan N, Özbay H. Hamilton Depresyon Derecelendirme Ölçeği (HDDÖ)'nin geçerliliği, güvenirliliği ve klinikte kullanımı. Psikiyatri Psikoloji Psikofarmakoloji Dergisi 1996; 4: 251-259.

55. Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for mania: reliability, validity and sensitivity. Br J Psychiatry 1978;133:429-35. doi: 10.1192/bjp.133.5.429.

56. Karadag F, Oral ET, Aran Yalçın F. Young Mani Derecelendirme Ölçeginin Türkiye'de geçerlik ve güvenirligi. Turk Psikiyatri Derg 2001; 13: 107- 114.

57. Rubinsztein JS, Michael A, Paykel ES, Sahakian BJ. Cognitive impairment in remission in bipolar affective disorder. Psychol Med 2000; 30(5):1025-36. doi: 10.1017/s0033291799002664.

58. Clark L, Iversen SD, Goodwin GM. Sustained attention deficit in bipolar disorder. Br J Psychiatry 2002; 180:313-9. doi: 10.1192/bjp.180.4.313.

59. Serafini G, Pardini M, Monacelli F, Orso B, Girtler N, Brugnolo A, Amore M, Nobili F, Team On Dementia Of The Irccs Ospedale Policlinico San Martino DM. Neuroprogression as an Illness Trajectory in Bipolar Disorder: A Selective Review of the Current Literature. Brain Sci 2021; 11(2):276. doi: 10.3390/brainsci11020276.

60. Bora E. A meta-analysis of data-driven cognitive subgroups in bipolar disorder. Eur Neuropsychopharmacol 2025; 90:48-57. doi: 10.1016/j.euroneuro.2024.10.008.

61. Lezak MD, Howieson DB, Bigler ED, Tranel D. Neuropsychological Assessment (5th ed.). Oxford University Press, 2012.

62. Fountoulakis KN, Vieta E, Bouras C, Notaridis G, Giannakopoulos P, Kaprinis G, Akiskal H. A systematic review of existing data on long-term lithium therapy: neuroprotective or neurotoxic? Int J Neuropsychopharmacol 2008; 11(2):269-87. doi: 10.1017/S1461145707007821.

63. Solé B, Jiménez E, Torrent C, Reinares M, Bonnin CDM, Torres I, Varo C, Grande I, Valls E, Salagre E, Sanchez-Moreno J, Martinez-Aran A, Carvalho AF, Vieta E. Cognitive impairment in bipolar disorder: Treatment and prevention strategies. Int J Neuropsychopharmacol 2017; 20(8):670-680. doi: 10.1093/ijnp/pyx032.