




## Research Article

# Low rate of performance validity failures among individuals with bipolar disorder

Ariana Tart-Zelvin<sup>1</sup>, Bethany A. Navis<sup>1</sup>, Elena M. Lamping<sup>1</sup>, Scott A. Langenecker<sup>2</sup>, Kelly A. Ryan<sup>1</sup>, Melvin G. McInnis<sup>1</sup> and David F. Marshall<sup>1</sup> 

<sup>1</sup>Department of Psychiatry, University of Michigan/Michigan Medicine, Ann Arbor, MI, USA and <sup>2</sup>Department of Psychiatry, The University of Utah, Salt Lake City, UT, USA

### Abstract

**Objective:** Assessing performance validity is imperative in both clinical and research contexts as data interpretation presupposes adequate participation from examinees. Performance validity tests (PVTs) are utilized to identify instances in which results cannot be interpreted at face value. This study explored the hit rates for two frequently used PVTs in a research sample of individuals with and without histories of bipolar disorder (BD). **Method:** As part of an ongoing longitudinal study of individuals with BD, we examined the performance of 736 individuals with BD and 255 individuals with no history of mental health disorder on the Test of Memory Malingering (TOMM) and the California Verbal Learning Test forced choice trial (CVLT-FC) at three time points. **Results:** Undiagnosed individuals demonstrated 100% pass rate on PVTs and individuals with BD passed over 98% of the time. A mixed effects model adjusting for relevant demographic variables revealed no significant difference in TOMM scores between the groups,  $a = .07$ ,  $SE = .07$ ,  $p = .31$ . On the CVLT-FC, no clinically significant differences were observed ( $ps < .001$ ). **Conclusions:** Perfect PVT scores were obtained by the majority of individuals, with no differences in failure rates between groups. The tests have approximately >98% specificity in BD and 100% specificity among non-diagnosed individuals. Further, nearly 90% of individuals with BD obtained perfect scores on both measures, a trend observed at each time point.

**Keywords:** effort performance; performance validity; bipolar disorder; affective disorders; test of memory and malingering; neuropsychology  
(Received 27 January 2021; final revision 2 February 2022; accepted 9 February 2022; First Published online 11 April 2022)

### Introduction

Assessing protocol validity for performance-based measures is imperative in both clinical and research contexts as interpretation of data presupposes adequate participation on the part of examinees. Performance validity tests (PVT) are therefore recommended to identify instances where one cannot interpret obtained data at face value (Bush et al., 2014).

Among the most frequently used measures of performance validity are the Test of Memory Malingering (TOMM; Tombaugh, 1996) and the optional forced-choice recognition trial of the California Verbal Learning Test-second edition (CVLT-FC; Delis et al., 2000). These measures employ a forced-choice recognition memory paradigm and given the large number of stimuli, particularly in the case of the TOMM, give the impression that they are harder than they are (Tombaugh, 1996). As a result, clinicians become concerned that those who obtain low scores on the task are more likely to be performing poorly for reasons other than genuine cognitive impairment. Each measure has demonstrated strong psychometric properties across settings and study designs (Martin et al., 2020; Schwartz et al., 2016). As per practice guidelines to minimize false positive diagnoses (Boone, 2013), each test employs conservative cut scores that maximize

specificity and positive predictive power. As part of this study, additional cutoffs for the TOMM were explored to increase sensitivity.

Cognitive difficulties in learning and memory, verbal fluency, executive functioning, working memory, and attention, for example, have been associated with bipolar disorder (BD) (Chang et al., 2012; Grande et al., 2016). However, a debate exists as to whether such cognitive weaknesses represent manifestations of disease driven mechanisms, poor effort on cognitive testing due to variable mood states, or blatant malingering (Gierok et al., 2005; Lieberman et al., 2016; Moritz et al., 2017). While performance validity is commonly examined among specific populations, such as those with mild traumatic brain injury (TBI) (Carone & Bush, 2013), it is rarely a focus among individuals with affective disorders such as BD. Until there is adequate research establishing the rate with which affective disorders alone influence the rate of failures on PVTs, hypothesizing that poor performance on these measures is a function of these diagnoses is merely speculative and relies primarily on conjecture (Considine et al., 2011).

Previous researchers have examined performance validity among individuals with diagnoses of major depressive disorder (MDD) as well as those who endorse a high level of anxiety. In a small study of 20 individuals diagnosed with MDD and high levels of depression measured by the Beck Depression Inventory, only two individuals scored below the 45 cutoff on the TOMM (Yanez,

**Corresponding author:** David F. Marshall, email: [davimars@med.umich.edu](mailto:davimars@med.umich.edu)

**Cite this article:** Tart-Zelvin A., Navis B.A., Lamping E.M., Langenecker S.A., Ryan K.A., McInnis M.G., & Marshall D.F. (2023) Low rate of performance validity failures among individuals with bipolar disorder. *Journal of the International Neuropsychological Society*, 29: 298–305, <https://doi.org/10.1017/S1355617722000145>

Copyright © INS. Published by Cambridge University Press, 2022. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

Fremouw, Tennant, Strunk, & Coker, 2006). Their results suggest that the TOMM can be used with severely depressed individuals, although authors do suggest slight caution given the fact that performance was not above the cutoff for all participants. In a more recent study of individuals with MDD, all study participants including 45 patients with MDD and 32 healthy controls performed above the cutoff on trial 2 (100% in each group, Considine et al., 2011). Among older community-dwelling adults with both anxiety and depression symptoms, all participants within the study scored above the 45 cutoff regardless of mood symptoms (Ashendorf, Constantinou, & McCaffrey, 2004).

This study examines the rate of failures on two common PVTs, the TOMM and CVLT-FC, across multiple time points among a large sample of well-educated individuals with BD and individuals without psychiatric diagnosis involved in a longitudinal research study. Given the state and trait effects of mood states such as depression, including multiple time points allows for examination of potential patterns over time. We explore the rate with which people with BD fail these tests and, given that recognition memory presented in a forced-choice format is fairly simple, whether obtained scores even approach failing scores. Given existing literature, we expect that individuals with mood disorders will demonstrate similar performance to healthy controls on both PVTs. Lastly, we explore different cutoffs for the TOMM (trial 1 and 2) and CVLT-FC to increase sensitivity and to better inform clinical practice.

## Methods

### Participants

This study includes a subset ( $N = 991$ ) of participants enrolled in the Prechter Longitudinal Study of Bipolar Disorder at the University of Michigan (McInnis et al., 2018) between October of 2010 and December of 2018. The larger longitudinal study examines biological, phenotypic, and cognitive outcomes associated with BD (see McInnis et al., 2018 for a full description). In the current study, a subset of participants of the larger cohort included those who completed performance validity measures at a baseline assessment, 1 year following the baseline, and again 5 years following the baseline.

Demographic and diagnostic information for participants can be found in Table 1. The two groups were well matched on sex and ethnicity. The BD group was older ( $M = 39.13$ ,  $SD = 13.6$ , ages 18 to 84) than the healthy control group ( $M = 35.72$ ,  $SD = 15.5$ , ages 18 to 77),  $t(397.07) = -3.12$ ,  $p = .002$ . The healthy control group was more educated ( $M = 16.05$ ,  $SD = 2.0$ ) than the BD group ( $M = 15.29$ ,  $SD = 2.1$ ),  $t(462.01) = 5.158$ ,  $p < .001$ . The BD group had a significantly higher percentage of individuals who identified as White ( $M = 86.7\%$ ) compared to the healthy control group ( $M = 71.4\%$ ),  $t(301.76) = -5.023$ ,  $p < .001$ . The healthy control group had higher IQ scores on average ( $M = 112.95$ ,  $SD = 11.7$ ) when compared to the BD group ( $M = 108.77$ ,  $SD = 12.4$ ),  $t(989) = 4.713$ ,  $p < .001$ . Racial identification notwithstanding, these demographic differences were of relatively small effect size (age  $g = 0.24$ , education  $g = 0.37$ , race  $g = 0.41$ , and IQ  $g = 0.35$ ). Lastly, there was a significant difference in depression scores (Hamilton Depression Rating Scale [HDRS]),  $F(3) = 125.212$ ,  $p < .001$ , and mania scores (Young Mania Rating Scale [YMRS]),  $F(3) = 47.280$ ,  $p < .001$ , between the groups, with the BD group having higher scores compared to the healthy control group;

**Table 1.** Demographic and clinical characteristics of all participants

	BD (n = 736) M (SD)	Controls (n = 255) M (SD)	p-value
Diagnosis (n)			
BD I	508	-	
BD II	145	-	
BD NOS	58	-	
Schizoaffective BD	25		
Age	39.13 (13.6)	35.72 (15.5)	.002**
Education	15.29 (2.1)	16.05 (2.0)	<.001**
Gender (M/F)	255/481	95/160	.453
Race (% White)	86.7%	71.4%	<.001**
Ethnicity (% Non-Hispanic or Latino)	94.6%	92.5%	.999
HAMD-17			
Baseline	8.92 (6.1)	.994 (1.4)	<.001**
1 Year	8.35 (6.0)	1.16 (1.6)	<.001**
5 Year	8.25 (5.8)	1.14 (1.6)	<.001**
YMRS			
Baseline	3.32 (3.8)	.304 (.71)	<.001**
1 Year	3.17 (4.0)	.41 (1.0)	<.001**
5 Year	3.28 (4.2)	.39 (.94)	<.001**
WASI IQ	108.77 (12.4)	112.95 (11.7)	<.001**

Note. BD = Bipolar Disorder. HAMD = Hamilton Depression Rating Scale. YMRS = Young Mania Rating Scale. WASI = Wechsler Abbreviated Scale Intelligence. IQ = Intelligence quotient.

\*  $p < .05$ . \*\*  $p < .01$ .

however, participants were not assessed if acutely manic per study protocol (i.e., YMRS score above 30) given concerns with ability to provide informed consent and need for medical intervention.

### Study procedure

Recruitment for the study occurred through advertisements in community mental health centers, in an outpatient specialty psychiatric clinic, in an inpatient psychiatric clinic, in local newspapers, at community outreach events, and on the Internet. All participants provided informed consent, and each participant received incentive payment for participation in the longitudinal study. The Institutional Review Board of the University of Michigan Medical School approved the study (IRBMED: HUM00000606).

Once participants gave informed consent, all participants completed an in-person baseline assessment. Clinicians evaluated all participants during the baseline assessment to confirm Diagnostic and Statistical Manual of Mental Disorders, fourth edition (DSM-IV) diagnoses using the Diagnostic Interview for Genetic Studies (DIGS, Nurnberger et al., 1994). To establish a consensus diagnosis and BD subtype (bipolar disorder I (BD I), bipolar disorder II (BD II), bipolar disorder not otherwise specified (BD NOS), and Schizoaffective bipolar disorder (Schizoaffective BD)) for participants, three authors evaluated medical records and data from the DIGS. The study excluded participants with active substance dependence or neurological disease at the time of enrollment. Clinicians administered mood measures were given at each time point. The research-defined neuropsychological test battery included the TOMM and CVLT-FC as performance validity measures as well as a measure of estimated IQ. Choice of performance validity measures were determined by data available from a fixed research battery that was part of a larger study from which our sample was derived.

## Study measures

### Performance validity measures

#### Test of memory malingering

The TOMM (Tombaugh, 1996) is a 50-item recognition test designed to detect feigned, exaggerated, or fabricated memory impairment. Initial validation studies of the TOMM (see Rees et al., 1998; Tombaugh, 1996) indicated that this instrument is a useful test for detecting exaggerated or malingered memory impairment. Using a criterion cutoff score of <45 on trial 2 (i.e.,  $\geq 90\%$  correct responding), specificity values exceeded 91% for clinical cases and 100% for nonclinical cases (see Rees et al., 1998; Tombaugh, 1996). This test was initially validated with a clinical sample, including five major groups (no cognitive impairment, cognitive impairment, aphasia, TBI, and dementia), and while there was no specific psychiatric group those with depression were linked into the cognitive impairment group which also included Korsakoff's syndrome, Parkinson's disease, Post Traumatic Stress Disorder (PTSD), and Huntington's disease. The cognitive impairment group obtained a mean score of 43.9 (5.3) for trial 1 and a mean score of 48.6 (3.1) for trial 2, suggesting that performance on the TOMM is very resistant to different types of severe cognitive impairment (Tombaugh, 1996). While it is traditional to use the cutoff score of <45 on trial 2 as suggested in the manual, this study also examined alternative cutoffs for trial 2 based on existing literature (e.g., Erdodi et al., 2018; Martin et al., 2020) in addition to examining trial 1 data as part of secondary analyses.

#### California verbal learning test-second edition (CVLT-II), forced choice trial

The CVLT-II (Delis et al., 2000) is a measure of episodic memory and verbal learning. Participants are read a list of 16 words over five trials and are asked to recall the list after each repetition of the list. The measure also consists of a distractor list, short and long-delay free recall, short and long-delay category-cued recall trials, a delayed recognition trial, and lastly, a delayed forced choice trial. During the delayed forced-choice trial, the technician presents two words to the examinee and they must choose the one that was in the original list. Distracter words are either concrete nouns, as are the target words, or are abstract nouns. Traditionally, the forced choice trial is an embedded performance validity measure in which more than one error signals possible problematic engagement (Schwartz et al., 2016). As described by Delis et al. (2000), over 94% of the CVLT-II normative sample obtained perfect scores on the forced choice recognition test, suggesting that a low score on this trial may indicate insufficient effort and that possible indications for a low score may include serious psychiatric disturbance such as psychosis, severe pain, mental delirium, or conscious or unconscious exaggeration of memory deficits.

### Mood measures

#### Hamilton depression rating scale

The HAMD-17 (Hamilton, 1960) is a 17-item measure designed to assess current (i.e., in the past 1 week) symptoms of depression. Scores between 10 and 13 represent mild symptoms of depression, 14–17 represents mild to moderate symptoms, and scores above 17 represent moderate to severe symptoms.

#### Young mania rating scale

The YMRS (Young et al., 1978) contains 11 items and quantifies current (i.e., in the past 1 week) symptoms of mania. Additionally,

clinicians include information derived from clinical observation. Scores between 13 and 19 represent minimal symptoms of mania, 20–25 represent mild symptoms, 26–37 represent moderate symptoms, and 38–60 represent severe symptoms.

### Cognitive measures

Within the Prechter Longitudinal Study of Bipolar Disorder, participants received a research-defined neuropsychological evaluation that focused on areas known to be adversely impacted in BD, including memory, attention and executive functioning, psychomotor speed, and emotion processing (Langenecker et al., 2010). Standard data reduction techniques were utilized, given the large number of dependent variables, using conceptually, statistically, and theoretically derived variables (see Langenecker et al., 2010 for a full description). All test scores with negative scale properties (lower numbers reflect better performance) were inverted. Eight factor scores, based on raw scores, are utilized, including auditory memory (as measured by the CVLT-II (Delis et al., 2000)), visual memory (Rey–Osterrieth Complex Figure Test (Meyers and Meyers, 1995)), fine motor dexterity (Purdue Pegboard test (Lezak, 1995)), verbal fluency and processing speed (FAS/Phonemic fluency of the Controlled Oral Word Association Test (Benton and Hamsher, 1976), Category/Animal fluency (Benton and Hamsher, 1976), Digit Symbol-Coding from the Wechsler Adult Intelligence Scale-Third Edition (Wechsler, 1997), Stroop Color and Word Test (Golden, 1978), and Trail Making Test Form B from Trail Making Test-Parts A and B (Armitage, 1946)), conceptual reasoning and set-shifting (Wisconsin Card Sorting Test (Grant and Berg, 1948) and Parametric Go/No-go (Langenecker et al., 2005)), processing speed with interference resolution (Trail Making Test Forms A and B, Digit Symbol, Stroop Color Word Test, and Parametric Go/No-go), inhibitory control (Parametric Go/No-go), and emotion processing (Emotion Perception Test (Green and Allen, 1997) and Facial Emotion Perception Test (Rappaport et al., 2002)).

The scores utilized from the CVLT-II included total learning over five trials, short delay free recall, short delay cued recall, long delay free recall, long delay cued recall, and recognition hits. Scores from Rey–Osterrieth Complex Figure Test included immediate recall, delayed recall, and recognition. On the fine motor task, scores on the Purdue Pegboard test included performance with the dominant hand, nondominant hand, and bimanual (using both hands). On emotion processing measures, errors (inverted) were utilized from the Emotion Perception Test along with accuracy and response time (inverted) from the Facial Emotion Perception Test. Regarding the Stroop Color Word Test, performance from the Stroop Word Condition, the Stroop Color Condition, and the Interference Condition were used. Correct categories and perseverative errors (inverted) were used from the Wisconsin Card Sort test. Mean accuracy for target trials, mean target response times (inverted), and mean accuracy for inhibitory trials were used from the Parametric Go/No-go task.

#### Wechsler abbreviated scale intelligence

The WASI (Wechsler, 1999) is a standardized measure that provides a Full-Scale Intellectual Quotient (FSIQ) and contains four subtests: Vocabulary, Similarities, Matrix Reasoning, and Block Design. The average reliability coefficient for the FSIQ-4 is 0.98 and 0.96 for the FSIQ-2. Average test–retest reliabilities are 0.92 and 0.88, respectively.

### Method of analyses

The primary analyses included examining the frequencies with which participants, across diagnostic groups and time periods, obtained adequate scores on the TOMM and CVLT-FC. We calculated percentages for individuals that received specific score combinations (i.e., a 50 on the TOMM and a 16 on the CVLT-FC at baseline) on both the TOMM and CVLT-FC at each of the three time points.

This study used independent-samples *t*-tests to analyze demographic variables using version 24 of the Statistical Package for the Social Sciences (SPSS). This study also used independent-samples *t*-tests to compare mean scores on the TOMM trial 1 and trial 2 and CVLT-FC between the two primary groups (i.e., all individuals with BD and healthy controls) at each of the three time points as well as between each of the bipolar groups at each time point. While the primary focus of the analyses is on trial 2 of the TOMM, secondary analyses were also conducted on available TOMM trial 1 data as well. Furthermore, independent-samples *t*-tests were utilized among euthymic individuals (HDRS < 8 and YMRS < 8) who had passed symptom validity tests (TOMM trial 2 > 45 and CLVT-FC > 14) to compare neurocognitive test performance between the two groups (BD and controls). Additionally, we analyzed data using the lme4 package from the R statistical software program (Bates et al., 2015; R Core Team, 2018). The study conducted mixed effects models to analyze whether TOMM trial 2 scores differed between individuals with BD and healthy controls with data from three different time points. While the data are skewed, the model is quite robust and only requires that residuals are normally distributed (Schielzeth et al., 2020). The outcome variable was the TOMM score. The study used fixed effect of diagnostic group with time point, HAMD-17, YMRS, and WASI IQ as covariates and individual participants as a random effect.

## Results

### Mood measures

As expected, individuals with BD had significantly higher scores on both mood measures compared to the healthy controls at each of the three time points; all *ps* < .01. Table 1 includes these results.

### Performance validity results

Only five of 1316 total TOMM trial 2 scores fell below the clinical cut off of <45 on trial 2. These five scores belonged to four participants from the BD group (three with a diagnosis of BD I and one with a diagnosis of BD NOS). The performances of this small group of participants are further addressed in the discussion section. Regarding the CVLT-FC, 84 participants provided a total of 126 errors consisting of 88 concrete errors and 38 abstract errors. Table 5 contains the number of individuals who performed at various cutoffs on the CVLT-FC at three time points. Cutoff scores were determined using existing literature (e.g., Erdodi et al., 2018; Schwartz et al., 2016). One cutoff was based on a systematic review conducted by Schwartz and colleagues (2016) in which they applied a cutoff of 14 on the forced choice trial (sensitivity 50% and specificity 93%). Another, more stringent cutoff, was based on a study conducted with a group of mixed TBI individuals (75% mTBI) (Erdodi et al., 2018). In that study, they used a cutoff of 15 on the forced choice trial (sensitivity 56% and specificity 92%). Given the longitudinal nature of the study and attrition over time, not all participants completed the CVLT-FC over the three time points as can be seen in Table 5. Table 5 also contains the

**Table 2.** Comparison of the different bipolar disorder groups on the TOMM trial 2 and CVLT-FC at three time points: baseline testing, 1 year after baseline, and 5 years after baseline. TOMM trial 1 data was examined from the one time point it was collected

Comparison Groups and Time Point	<i>p</i> -value	Time Point	<i>p</i> -value
BD I versus Schizoaffective BD			
TOMM1	.542		
TOMM2 Baseline	.511	CVLT-FC Baseline	.813
TOMM2 1 Year	.779	CVLT-FC 1 Year	.62
TOMM2 5 Year	.417	CVLT-FC 5 Year	.51
BD I versus BD II			
TOMM1	.056		
TOMM2 Baseline	.05	CVLT-FC Baseline	.114
TOMM2 1 Year	.982	CVLT-FC 1 Year	.064
TOMM2 5 Year	.939	CVLT-FC 5 Year	.584
BD I versus BD NOS			
TOMM1	.663		
TOMM2 Baseline	.779	CVLT-FC Baseline	.392
TOMM2 1 Year	.788	CVLT-FC 1 Year	.698
TOMM2 5 Year	.446	CVLT-FC 5 Year	.813
Schizoaffective BD versus BD II			
TOMM1	.081		
TOMM2 Baseline	.494	CVLT-FC Baseline	.66
TOMM2 1 Year	.589	CVLT-FC 1 Year	.94
TOMM2 5 Year	.488	CVLT-FC 5 Year	.04*
Schizoaffective BD versus BD NOS			
TOMM1	.826		
TOMM2 Baseline	.417	CVLT-FC Baseline	.624
TOMM2 1 Year	.418	CVLT-FC 1 Year	.632
TOMM2 5 Year	.387	CVLT-FC 5 Year	.387
BD II versus BD NOS			
TOMM1	.089		
TOMM2 Baseline	.415	CVLT-FC Baseline	.929
TOMM2 1 Year	.557	CVLT-FC 1 Year	.435
TOMM2 5 Year	.456	CVLT-FC 5 Year	.946

Note. BD = Bipolar Disorder. TOMM2 = Test of Memory Malinger Trial 2. TOMM1 = Test of Memory Malinger Trial 1. CVLT-FC = California Verbal Learning Test forced choice trial. \* *p* < .05.

number of individuals who performed at various cutoffs on the TOMM (trial 1 and 2) at multiple time points.

This study compared the performance on the TOMM trials 1 and 2 and the CVLT-FC between individuals with BD I, BD II, BD NOS, and Schizoaffective BD (Table 2). As discussed in the methods section, the primary focus was on TOMM trial 2 scores based on the manual cutoff and the more commonly used CVLT-FC cutoff. However, TOMM trial 1 data were used when available as well as additional cutoffs for secondary analyses to explore alternative cutoffs that could increase the sensitivity of the measure. While the CVLT-FC and TOMM trial 2 scores were examined at three time points, TOMM trial 1 data was only available for one time point. There were no statistically or clinically significant differences between the groups on the TOMM trial 1 or 2 at any of the three time points. While the Schizoaffective BD group and the BD II group demonstrated a statistically significant difference in performance on the CVLT-FC at one time point (*p* = .044), this does not represent a clinically significant difference as the means differed by less than one point. Additionally, after correcting for multiple tests, this would no longer be considered statistically significant either. Table 3 summarizes the performance validity comparisons between the combined BD group and the healthy control group. The combined BD group of individuals did not demonstrate significant differences in their performance compared to that of healthy controls on the TOMM trial 2 across all three time points. While the healthy control group demonstrated statistically significant differences in their performance compared to the BD group on the CVLT-FC and TOMM trial 1 during

**Table 3.** Mean TOMM trial 1 and 2 and CVLT-FC scores for individuals with bipolar disorder and healthy controls

	BD	Controls	<i>p</i> -value
	( <i>n</i> = 736) M (SD)	( <i>n</i> = 255) M (SD)	
TOMM Trial 1			
Baseline (BP <i>n</i> = 463; HC <i>n</i> = 140)	47.79(3.0)	48.81(1.9)	<.001**
TOMM Trial 2			
Baseline (BP <i>n</i> = 381; HC <i>n</i> = 120)	49.84 (1.0)	49.94 (.27)	.077
1 Year (BP <i>n</i> = 346; HC <i>n</i> = 124)	49.85 (1.0)	49.91 (.38)	.514
5 Year (BP <i>n</i> = 264; HC <i>n</i> = 81)	49.89 (.52)	49.95 (.22)	.340
CVLT-FC			
Baseline (BP <i>n</i> = 736; HC <i>n</i> = 255)	15.91 (.38)	15.98 (.13)	<.001**
1 Year (BP <i>n</i> = 558; HC <i>n</i> = 194)	15.90 (.48)	15.99 (.07)	<.001**
5 Year (BP <i>n</i> = 263; HC <i>n</i> = 81)	15.94 (.26)	15.98 (.16)	.175

Note. BD = Bipolar Disorder. TOMM = Test of Memory Malingering. CVLT-FC = California Verbal Learning Test forced choice trial.

\*  $p < .05$ . \*\*  $p < .01$ .

baseline ( $p < .001$ ) and the 1-year follow-up ( $p < .001$ ), these differences do not represent clinically significant differences.

Results from the mixed effects model adjusting for mood and IQ revealed no significant difference in TOMM trial 2 scores between the healthy control group and the BD group,  $a = .07$ ,  $SE = .07$ ,  $p = .31$ . The groups demonstrated incredibly similar performance (BD  $M = 49.85$ ,  $SD = .926$  and HC  $M = 49.93$ ,  $SD = .307$ ). When this model was re-rerun without adjusting for mood using the HAMD-17 and YMRS, there were still no significant differences between the healthy control group and individuals with BD. Additionally, Table 4 outlines the percentage of participants who obtained specific score combinations on the TOMM trial 2 and CVLT-FC at each time point.

### Neurocognitive test results

After accounting for the impact of mood symptoms and performance validity scores, the euthymic BD group exhibited poorer performance than the control group on most cognitive factors, including fine motor control  $t(262) = -4.019$ ,  $p < .001$ , auditory memory  $t(265) = -2.768$ ,  $p = .006$ , verbal fluency and processing speed  $t(256) = -4.848$ ,  $p < .001$ , processing speed with interference resolution  $t(255) = -5.498$ ,  $p < .001$ , and inhibitory control  $t(259) = -5.317$ ,  $p < .001$ . Table 6 includes the means and standard deviations between the two groups for each of the eight factor scores.

### Discussion

Across all time points and among BD and controls, the rate of PVT failures was exceedingly low. None of the 255 undiagnosed individuals failed a single performance validity test at any of the assessment periods, whereas the 736 individuals with BD failed less than 2% of the time, and inconsistently over time. In fact, nearly 90% of individuals with BD obtained perfect scores on both performance validity measures at each time point.

Among those with imperfect scores, individuals typically only missed one item on either test. In fact, the mean scores on the CVLT-FC differed by less than one point between the BD group and the healthy control group. Such a negligible difference is not considered clinically meaningful. Such perfect and nearly perfect scores are far from the more stringent standards for failure of each test, indicating that each PVT is robust to possible momentary attentional lapses. Other research samples show a similar pattern of perfect, or nearly perfect scores (Rees et al., 1998; Tombaugh,

1997). Dementia notwithstanding, this study contributes to the literature indicating that having a clinical diagnosis is insufficient to account for failures on these PVT's (Etherton et al., 2005; Hunt et al., 2014; Kirk et al., 2011). The commonly employed cut scores for these measures are therefore sufficiently conservative that clinicians can interpret failing scores with high positive predictive power and specificity.

These results provide compelling evidence that the mere diagnosis of BD is insufficient to account for failing scores on measures of performance validity. The results are consistent with existing literature (Duncan, 2005; Gierok et al., 2005) and helped to extend this literature with the use of multiple PVTs, a much larger sample, and the inclusion of scores across multiple time points. Clinicians obtaining substandard scores for examinees, therefore, must account for low scores on these measures from factors other than a BD diagnosis. Further, based on the results of this study, authors encourage future discussion on potentially more useful cutoff scores among patients with BD. Specifically, an alternative cutoff for the TOMM trial 2 could be beneficial for clinical practice while maintaining high specificity and strong sensitivity. Others have suggested alternative cutoffs (e.g., Denning, 2012), and this study suggests that the cutoff for the TOMM trial 2 could be <46 which would increase sensitivity and represent a conservative change. A less conservative practice to consider would be to have clinicians utilize a cutoff of <49 which would provide higher sensitivity to poor effort while maintaining adequate specificity.

This study has several limitations that subsequent research can address. Given the longitudinal nature of the larger study the data were utilized from, attrition occurred over time as would be expected, and scores were not available for each participant at each of the three time points for the TOMM trials 1 and 2 and the CVLT-FC. Also, these findings pertain to well-educated young to early older adults who were mostly Caucasian and volunteered in a research context. As such, these results will need replication in clinical settings with older, less educated, more ethnically diverse individuals in order to provide more generalizability of this data. Each participant received incentive payment for their participation alone as opposed to performing at a particular level on cognitive testing. The fact that there was no incentive to perform well could have also affected participants' performance on cognitive testing. Additionally, individuals did not complete testing if, at the time of the neurocognitive evaluation, they were acutely manic. While Vrabie et al. (2015) concluded that assessing those in mania is challenging due to the associated distractibility and impulsivity, it is plausible that acute mania affects performance on PVTs. This is a subject for subsequent investigations to explore. There has also been an emerging literature regarding PVTs in research settings, and in particular PVT failures, though this literature has yielded mixed results with most studies having focused on a college population with higher failure rates attributed to using more liberal cut-offs (Roye et al., 2019). While we cannot assume that all of our participants were putting forth maximal effort, our study sample was older compared to those previously reported and we also utilized the recommended cutoffs with rates similar to those reported by other studies.

Another limitation of this study concerns the use of forced-choice recognition PVTs, which are easily passed by individuals with significant neuropathology outside dementia. Readers should not assume that failure of any PVT is unexplainable by BD, especially PVTs embedded within existing cognitive measures. Accordingly, these conclusions do not fully address effortful processing, investment in performing well, motivation, or level

**Table 4.** TOMM trial 2 and CVLT-FC performance for individuals with bipolar disorder and healthy controls. Percentage of individuals who obtained the score combinations below at baseline (0 Year), 1 year following baseline (1 Year), and at 5 years following baseline (5 Year)

TOMM, CVLT-FC	BD			Controls		
	0 Year (n = 381)	1 Year (n = 345)	5 Year (n = 263)	0 Year (n = 120)	1 Year (n = 122)	5 Year (n = 81)
50, 16	88.71%	88.99%	88.97%	92.50%	93.44%	92.59%
50, 15	3.94%	3.48%	3.80%	2.50%	-	2.47%
50, 14	0.52%	0.87%	0.76%	-	-	-
50, 13	-	0.29%	-	-	-	-
49, 16	3.15%	4.06%	4.18%	4.17%	4.92%	4.94%
49, 15	0.52%	0.29%	0.38%	-	-	-
49, 14	0.52%	-	-	-	-	-
49, 13	-	-	-	-	-	-
48, 16	1.05%	0.29%	0.76%	0.83%	0.82%	-
48, 15	0.26%	-	-	-	-	-
48, 14	-	-	-	-	-	-
48, 13	-	-	-	-	-	-
47, 16	0.52%	0.58%	0.76%	-	0.82%	-
47, 15	-	-	-	-	-	-
47, 14	-	-	-	-	-	-
47, 13	-	-	-	-	-	-
46, 16	0.26%	0.29%	-	-	-	-
46, 15	-	-	-	-	-	-
46, 14	-	-	-	-	-	-
46, 13	-	-	-	-	-	-
45, 16	-	-	-	-	-	-
45, 15	-	-	-	-	-	-
45, 14	-	-	-	-	-	-
45, 13	-	-	-	-	-	-
45-50, 10-12	-	0.29%	-	-	-	-
40-44, 13-16	0.26%	-	0.38%	-	-	-
30-39, 13-16	0.26%	0.29%	-	-	-	-
30-39, 10-12	-	0.29%	-	-	-	-

Note. BD = Bipolar Disorder. TOMM = Test of Memory Malinger. CVLT-FC = California Verbal Learning Test forced choice trial.

**Table 5.** This table explores the number of individuals who performed at various cutoffs on the TOMM (trial 1 and 2) and the CLVT-FC. CVLT-FC performance and TOMM trial 2 performance for individuals with bipolar disorder and healthy controls are included for three time points. Performance on TOMM trial 1 was included for one time point

	Pass ≥ 14		Pass ≥ 15	
	BD	HC	BD	HC
CVLT-FC				
Baseline (BP n = 733; HC n = 254)	729	254	724	254
1 Year (BP n = 555; HC n = 194)	549	194	544	194
5 Year (BP n = 264; HC n = 83)	264	83	263	82
TOMM				

	T1 BL		T2 BL		T2 1Yr		T2 5 Yr	
	BD (n = 463)	HC (n = 140)	BD (n = 379)	HC (n = 119)	BD (n = 345)	HC (n = 124)	BD (n = 262)	HC (n = 81)
>=38	456	140	379	119	344	124	262	81
>=39	454	140	379	119	344	124	262	81
>=40	451	139	379	119	343	124	262	81
>=41	447	139	378	119	343	124	262	81
>=42	442	138	378	119	343	124	262	81
>=43	433	137	378	119	343	124	262	81
>=44	430	135	378	119	343	124	262	81
>=45	411	133	378	119	343	124	261	81
>=46	391	131	378	119	343	124	261	81
>=47	361	126	377	119	342	124	261	81
>=48	319	120	375	119	340	123	259	81
>=49	258	105	370	118	339	122	257	81

Note. BD = Bipolar Disorder. TOMM = Test of Memory Malinger. T1 = TOMM Trial 1. T2 = TOMM Trial 2. BL = Baseline. HC = Healthy Control. CVLT-FC = California Verbal Learning Test forced choice trial. "Pass ≥ 14" represents the number of individuals who passed the first cutoff with a score equal to or greater than 14 on the CVLT-FC. "Pass ≥ 15" represents the number of individuals who passed a second cutoff with a score equal to or greater than 15 on the CVLT-FC.

**Table 6.** Mean neurocognitive factor scores for individuals with euthymic bipolar disorder and healthy controls

	BD (n = 144) M (SD)	Controls (n = 137) M (SD)	p-value
Fine Motor	-0.14 (0.92)	0.28 (0.82)	<.001**
Visual Memory	0.09 (0.92)	0.27 (0.96)	.119
Auditory Memory	-0.01 (0.88)	0.26 (0.76)	.006*
Emotion Processing	-0.04 (0.82)	0.14 (0.74)	.062
Executive Functioning			
Verbal Fluency with Processing Speed	-0.07 (0.63)	0.30 (0.64)	<.001**
Conceptual Reasoning/ Set-Shifting	-0.04 (0.65)	0.09 (0.51)	.064
Processing Speed with Interference	-0.07 (0.64)	0.34 (0.56)	<.001**
Inhibitory Control	-0.21 (0.63)	0.21 (0.67)	<.001**

Note. BD = Bipolar Disorder.

\*  $p < .05$ . \*\*  $p < .01$ .

of engagement. It should also be noted that while PVT sensitivity good, it is not 100%. It is possible that existing PVT's, both embedded and stand-alone, do not have the sensitivity required for all clinical circumstances. Further research should explore other PVTs and address these factors.

In conclusion, having a diagnosis of BD alone is not sufficient to explain failures on two widely used forced-choice recognition PVTs. These PVTs in particular require minimal time to administer and speak to the interpretability of a testing battery. Additionally, insufficient effort does not explain away or address reduced neurocognitive functioning associated with individuals with BD and this lends credibility to our groups previous extensive work and current findings showing poorer neurocognitive function in BD compared to controls on measures of memory, executive functioning, and motor abilities. Thus, disease driven mechanisms associated with cognitive weaknesses experienced by individuals with affective disorders such as BD should serve as a focus for research and clinical intervention.

**Acknowledgments.** With gratitude, we acknowledge the Prechter Bipolar Longitudinal Research participants for their contributions and the research staff for their dedication in the collection and stewardship of the data used in this publication. Additionally, we are incredibly grateful for the thoughtful contributions from Dr. Robert J. Spencer.

**Financial support.** This research was supported by the Heinz C Prechter Bipolar Program, the Richard Tam Foundation, and the Department of Psychiatry and Depression Center at the University of Michigan.

**Conflicts of interest.** Drs. Tart-Zelvin, Ryan, Langenecker, and Marshall report no competing interest related to the content of this study. Dr. McInnis has consulted for Janssen and Otsuka Pharmaceuticals in the past 3 years. The Heinz C. Prechter Bipolar Research Fund, the Richard Tam Foundation at the University of Michigan, the University of Michigan Depression Center, and the Department of Psychiatry of the University of Michigan supported this study. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not reflect the views of the funding agency.

## References

Armitage, S. G. (1946). An analysis of certain psychological tests used in the evaluation of brain injury. *Psych Mono*, 60, 1–48.

Ashendorf, L., Constantinou, M., & McCaffrey, R. J. (2004). The effect of depression and anxiety on the TOMM in community-dwelling older adults. *Archives of Clinical Neuropsychology*, 19, 125–130.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1–48. <https://doi.org/10.18637/jss.v067.i01>

Benton, A., & Hamsher, K. (1976). *Multilingual Aphasia Examination*. Iowa City, IA: The University of Iowa.

Boone, K. (2013). *Clinical practice of forensic neuropsychology: an evidence-based approach*. New York, NY: The Guilford Press.

Bush, S. S., Heilbronner, R. L., & Ruff, R. M. (2014). Psychological assessment of symptom and performance validity, response bias, and malingering: Official position of the association for scientific advancement in psychological injury and law. *Psychological Injury and Law*, 7, 197–205. <https://doi.org/10.1007/s12207-014-9198-7>

Carone, D. A., & Bush, S. S. (2013). *Mild traumatic brain injury: symptom validity assessment and malingering*. New York, NY: Springer Publishing Company.

Chang, Y.-H., Chen, S.-L., Lee, S.-Y., Hsu, Y.-W., Wu, J. Y., Chen, S.-H., Chu, C. H., Lee, I. H., Yeh, T. L., Tzeng, N. S., Huang, S. Y., Yang, Y. K., & Lu, R.-B. (2012). Neuropsychological functions in bipolar disorders I and II with and without comorbid alcohol dependence. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, 37, 211–216.

Considine, C. M., Weisenbach, S.L., Walker, S.J., McFadden, E.M., Franti, L.M., Bieliauskas, L.A., & Langenecker, S.A. (2011). Auditory memory decrements, without dissimulation, among patients with major depressive disorder. *Archives of Clinical Neuropsychology*, 26, 445–453. <https://doi.org/10.1093/arclin/acr041>

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (2000). *California Verbal Learning Test—Second Edition (CVLT-II)*. California Verbal Learning Test—Second Edition (CVLT-II). San Antonio, TX: The Psychological Corporation.

Denning, J. H. (2012). The efficiency and accuracy of the Test of Memory Malingering trial 1, errors on the first 10 items of the Test of Memory Malingering, and five embedded measures in predicting invalid test performance. *Archives of Clinical Neuropsychology*, 27, 417–432.

Duncan, A. (2005). The impact of cognitive and psychiatric impairment of psychotic disorders on the Test of Memory Malingering (TOMM). *Assessment*, 12, 123–129.

Erdodi, L. A., Abeare, C. A., Medoff, B., Seke, K. R., Sagar, S., & Kirsch, N. L. (2018). A single error is one too many: The forced choice recognition trial of the CVLT-II as a measure of performance validity in adults with TBI. *Archives of Clinical Neuropsychology*, 33, 845–860.

Etherton, J. L., Bianchini, K. J., Greve, K. W., & Ciota, M. A. (2005). Test of memory malingering performance is unaffected by laboratory-induced pain: implications for clinical use. *Archives of Clinical Neuropsychology*, 20, 375–384.

Gierok, S., Dickson, A. L., & Cole, J. A. (2005). Performance of forensic and non-forensic adult psychiatric inpatients on the test of memory malingering. *Archives of Clinical Neuropsychology*, 20, 755–760.

Golden, C. (1978). *Stroop Color and Word Test*. Chicago, IL: Stoelting.

Grande, I., Berk, M., Birmaher, B., & Vieta, E. (2016). Bipolar disorder. *The Lancet*, 387, 1561–1572. [https://doi.org/10.1016/S0140-6736\(15\)00241-X](https://doi.org/10.1016/S0140-6736(15)00241-X)

Grant, D. A., & Berg, E. A. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card sorting paradigm. *Journal of Experimental Psychology*, 38, 404–411.

Green, P. W., & Allen, L. M. (1997). *The Emotional Perception Test*. Durham, NC: CogniSyst Inc.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry*, 23, 56–62. <https://doi.org/10.1136/jnnp.23.1.56>

Hunt, S., Root, J. C., & Bascetta, B. L. (2014). Effort testing in schizophrenia and schizoaffective disorder: Validity indicator profile and test of memory malingering performance characteristics. *Archives of Clinical Neuropsychology*, 29, 164–172.

Kirk, J. W., Harris, B., Hutaff-Lee, C. F., Koelemay, S. W., Dinkins, J. P., & Kirkwood, M. W. (2011). Performance on the test of memory malingering (TOMM) among a large clinic-referred pediatric sample. *Child Neuropsychology*, 17, 242–254. <https://doi.org/10.1080/09297049.2010.533166>

Langenecker, S. A., Saunders, E. F., Kade, A. M., Ransom, M. T., & McInnis, M. G. (2010). Intermediate: cognitive phenotypes in bipolar disorder. *Journal of Affective Disorders*, 122, 285–293.

- Langenecker, S.A., Bieliauskas, L.A., Rapport, L.J., Zubieta, J.-K., Wilde, E.A., & Berent, S. (2005). Face emotion perception and executive functioning deficits in depression. *Journal of Clinical and Experimental Neuropsychology*, *27*, 320–333.
- Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed). Oxford University Press.
- Lieberman, L., Liu, H., Huggins, A. A., Katz, A. C., Zvolensky, M. J., & Shankman, S. A. (2016). Comparing the validity of informant and self-reports of personality using laboratory indices of emotional responding as criterion variables. *Psychophysiology*, *53*, 1386–1397.
- Martin, P. K., Schroeder, R. W., Olsen, D. H., Maloy, H., Boettcher, A., Ernst, N., & Okut, H. (2020). A systematic review and meta-analysis of the test of memory malingering in adults: Two decades of deception detection. *Clinical Neuropsychologist*, *34*, 88–119.
- McInnis, M. G., Assari, S., Kamali, M., Ryan, K., Langenecker, S. A., Saunders, E. F. H., Versha, K., Evans, S., O'Shea, K. S., Mower Provost, E., Marshall, D., Forger, D., Deldin, P., Zoellner, S., & Prechter Bipolar Clinical Research Collaborative. (2018). Cohort profile: The Heinz C. Prechter longitudinal study of bipolar disorder. *International Journal of Epidemiology*, *47*, 28–28n.
- Meyers, J., & Meyers, K. (1995). *Rey Complex Figure and Recognition Trial: Professional manual*. Odessa, FL: Psychological Assessment Resources.
- Moritz, S., Klein, J. P., Desler, T., Lill, H., Gallinat, J., & Schneider, B. C. (2017). Neurocognitive deficits in schizophrenia. Are we making mountains out of molehills? *Psychological Medicine*, *47*, 2602–2612. <https://doi.org/10.1017/S0033291717000939>
- Nurnberger, J. I., Blehar, M. C., Kaufmann, C. A., York-Cooler, C., Simpson, S. G., Harkavy-Friedman, J. P., Severe, J. B., Malaspina, D., & Reich, T. (1994). Diagnostic interview for genetic studies: Rationale, unique features, and training. *Archives of General Psychiatry*, *51*, 849–859. <https://doi.org/10.1001/archpsyc.1994.03950110009002>
- R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from Available online at <https://www.R-project.org/>
- Rapport, L. J., Friedman, S. L., Tzelepis, A., & Van Voorhis, A. (2002). Experienced emotion and affect recognition in adult attention-deficit hyperactivity disorder. *Neuropsychology*, *16*, 102–110.
- Rees, L. M., Tombaugh, T. N., Gansler, D. A., & Moczynski, N. P. (1998). Five validation experiments of the test of memory malingering (TOMM). *Psychological Assessment*, *10*(1), 10–20.
- Roye, S., Calamia, M., Bernstein, J. P. K., De Vito, A. N., & Hill, B. H. (2019). A multi-study examination of performance validity in undergraduate research participants. *The Clinical Neuropsychologist*, *33*, 1138–1155.
- Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allogue, H., Teplitsky, C., Reale, D., Dochtermann, N. A., Garamszegi, L. Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, *11*, 1141–1152.
- Schwartz, E. S., Erdodi, L., Rodriguez, N., Ghosh, J. J., Curtain, J. R., Flashman, L. A., & Roth, R. M. (2016). CVLT-II forced choice recognition trial as an embedded validity indicator: A systematic review of the evidence. *Journal of the International Neuropsychological Society*, *22*, 851–858.
- Tombaugh, T. (1996). *TOMM, Test of Memory Malingering*. North Tonawanda, NY: Multi-Health Systems Inc.
- Tombaugh, T. N. (1997). The test of memory malingering (TOMM): Normative data from cognitively intact and cognitively impaired individuals. *Psychological Assessment*, *9*, 260–268. <https://doi.org/10.1037/1040-3590.9.3.260>
- Vrabie, W., Marinescu, V., Talasman, A., Tautu, O., Drima, E., & Miclutia, I. (2015). Cognitive impairment in manic bipolar patients: Important, understated, significant aspects. *Annals of General Psychiatry*, *14*. <https://doi.org/10.1186/s12991-015-0080-0>
- Wechsler, D. (1997). *WAIS-III: Wechsler Adult Intelligence Scale-third edition Administration and Scoring Manual*. San Antonio, TX: Psychological Corporation.
- Wechsler, D. (1999). *Manual for the Wechsler Abbreviated Scale of Intelligence*. San Antonio, TX: Psychological Corporation.
- Yanez, Y. T., Fremouw, W., Tennant, J., Strunk, J., & Coker, K. (2006). Effects of severe depression on TOMM performance among disability-seeking outpatients. *Archives of Clinical Neuropsychology*, *21*, 161–165.
- Young, R. C., Biggs, J. T., Ziegler, V. E., & Meyer, D. A. (1978). A rating scale for mania: reliability, validity and sensitivity. *British Journal of Psychiatry*, *133*, 429–435.