Archives of Assessment Psychology, Vol. 11, No. 1, (137-161)
Printed in U.S.A.

Old vs New: The Classic and D-KEFS Trails as Embedded Performance Validity Indicators and Measures of Psychomotor Speed/Executive Function

Laszlo Erdodi, Ph.D., Jessica Hurtubise, M.A., Maame Brantuo, M.A. Laura Cutler, B.A., Arianna Kennedy, B.S.W., and Rayna Hirst, Ph.D.

Abstract

This study sought to provide a direct comparison between old and new versions of the Trail Making Test (TMT). Eighty-five undergraduate student volunteers were administered the old and new TMT. A third of them were instructed to feign neuropsychiatric deficits. The classification accuracy of the TMTs was evaluated against experimentally induced and psychometrically defined invalid performance. Results showed that the old TMT demonstrated superior psychometric properties, both as a measure of cognitive ability and performance validity. In conclusion, newer and more sophisticated versions of a test are not necessarily better than older, established instruments. Replications in clinical samples are needed to verify the findings.

Introduction

Performance Validity: Definition, Importance, Brief History

Performance validity is the extent to which the examinee's score on a given test accurately reflects the underlying abilities the instrument was designed to measure. As such, it is a basic assumption underlying the validity of the clinical interpretations based on psychometric data (Lezak et al., 2012). Historically, valid performance was assumed by default, reinforced by the belief that non-credible responding would be readily apparent to the assessing clinician. As such, there was no perceived need for objective measures of performance validity. However, it has long been evident that clinical intuition is an inadequate detection method (Heaton et al., 1978). Although the proportion of practitioners who report relying on subjective judgment to determine the credibility of a given response set has been shrinking (Dandachi-FitzGerald et al., 2013), recent reports raise concerns about the veracity of self-reported clinical practices (MacAllister et el., 2019).

Although conversations about invalid performance are often uncomfortable, failing to address the issue can harm the examinee and a wide range of stakeholders (Chafetz et al., 2015; Kirkwood, 2015; Rogers & Bender, 2018; Lippa, 2018; Niesten, Merckelbach, Dandacni-FitzGerald, & Jelicic, 2020; Rickards et al., 2018), and ultimately, undermine the credibility of neuropsychological assessments as a whole. Professional organizations recognize the significant threat non-credible responding poses to the validity of diagnostic and treatment decisions. As consolidated in the National Academy of Neuropsychology consensus statement (NAN; Bush et al., 2005) and reinforced by the American Academy of Clinical Neuropsychology (AACN; Heilbronner et al., 2009), best-practice standards deem critical the inclusion of symptom (SVT) and performance validity testing (PVT) in neuropsychological assessments (Bush et al., 2014; Schutte et al., 2015). Whereas SVTs are employed to detect fabricated or exaggerated self-reported

symptoms, PVTs monitor the credibility of test-taking behavior on performance-based cognitive measures.

Free-standing vs Embedded Measures

By design, SVTs and PVTs are either free-standing like the Inventory of Problems (Viglione, et al., 2017) and the Word Choice Test (WCT; Pearson, 2009) or embedded like the Response Bias Scale of the MMPI-2 (Gervais et al., 2007) and the reliable digit span (Greiffenstein et al., 1994). Long considered gold standard instruments, free-standing PVTs tend to have more robust signal detection performance, as they were specifically designed and calibrated to separate credible from non-credible presentation (Alverson et al., 2019; Larrabee, 2008; Erdodi, Green, et al., 2019). However, their limitations are becoming increasingly apparent: the cost of test material and clinician time, vulnerability to coaching (Bauer & McCaffrey, 2006; Victor & Abeles, 2004), failure to provide information on cognitive ability (arguably, the ultimate goal of neuropsychological assessment) despite taxing the examinee's mental stamina (Mossman & Hart, 1996; Rickards et al., 2018), and the risky epistemological leap of rendering an entire neurocognitive profile invalid based on failures on a few isolated PVTs (Bigler, 2014).

The growing awareness of these issues has led to the recent proliferation of embedded validity indicators (EVIs). EVIs have the advantage of simultaneously measuring the target clinical construct and the credibility of test-taking behavior, providing a cost-effective and inconspicuous method for monitoring performance validity (Rai et al., 2019; Webber & Soble, 2018). In addition, EVIs are fused with standard neuropsychological tests. Therefore, they monitor performance validity from within measures of cognitive ability. Like their free-standing counterparts, EVIs use empirically derived cutoffs to distinguish credible from non-credible response sets.

By lowering the administrative cost of repeatedly sampling performance validity, EVIs provide an attractive alternative to resource-intensive free-standing PVTs while providing a multivariate approach to performance validity assessment. Boone (2009) advocated for the administration of multiple validity measures distributed throughout evaluation sessions, given the potential for fluctuating and domain-specific patterns of non-credible responding (Cottingham et al., 2014; Erdodi, Dunn et al., 2018; Erdodi, Tyson, Abeare et al., 2018; Schroeder et al., 2019). Numerous studies have shown that multivariate models provide robust probabilistic metrics of the veracity of the assessment results since they are more likely to detect transient episodes of invalid performance, ranging from unintentional drifts in the attention of compliant examinees to overt malingering (Boskovic et al., 2019; Dandachi-FitzGerald et al., 2020; Jones, 2016; Martin et al., 2015; Mossman & Hart, 1996; Tracy, 2014; Walczyk et al., 2018).

The Trail Making Test (TMT)

Originally developed by US Army psychologists, the TMT has two parts: A and B. TMT_A requires the examinee to draw lines to connect encircled numbers (1 through 25) scattered on an 8.5 x 11-inch page in increasing order, as quickly as possible. TMT_B increases the cognitive load by requiring the examinee to switch back and forth between numbers (increasing order) and letters (alphabetical order). The most commonly used current administration protocol was established by Reitan (1958) and requires the examiner to point out errors and direct the examinee to self-correct.

TMT_A is a measure of simple visual attention, scanning, and psychomotor speed. In contrast, TMT_B is a measure of complex/divided visual attention, and mental flexibility. The main outcome measure is time-to-completion (T2C) in seconds. Performance on TMT is influenced by age (Salthouse et al., 2000; Stuss et al., 1987) and level of education (Bornstein, 1985; Hester et al., 2005). The TMT is sensitive to neuropsychological deficits associated with a wide range of disorders (Craun et al., 2019; Greer et al., 2010) ranging from traumatic brain injury (Lange et al.,

2005) to psychiatric disorders (Gass & Daniel, 1990; Lamberty et al., 1994; Knowles et al., 2015; Neill & Rossell, 2013; Savla et al., 2011) to neurodegenerative conditions (Bialystok et al., 2014; Greenlief et al., 1985; Heidler-Gary et al., 2007).

The Delis-Kaplan Executive Systems (D-KEFS; Delis, Kaplan, & Kramer, 2001) adaptation ($TMT_{D\text{-}KEFS}$) is a modified and expanded version of the original test. The $TMT_{D\text{-}KEFS}$ contains five trials: (1) Visual Scanning, (2) Number Sequencing, (3) Letter Sequencing, (4) Number-Letter Switching, and (5) Motor Speed. Each trial is presented on a 17 x 11-inch paper. The four baseline trials (1, 2, 3, and 5) measure the requisite subskills (visual scanning, number sequencing, letter sequencing, and motor speed) needed to perform Trial 4 ($TMT_{D\text{-}KEFS\ 4}$), a task analogous to TMT_B.

The TMT as EVI

In response to concerns that TMT performance may be confounded by non-credible responding (Backhaus et al., 2004; Goebel, 1983; Lees-Haley & Fox, 1990; O'Bryant et al., 2003; Ruffolo et al., 2000), initial EVI development efforts proposed raw score cutoffs (Busse & Whiteside, 2012; Iverson et al., 2002; Powell et al., 2011; Shura et al., 2016; Whiteside et al., 2018). Across studies, the proposed cutoffs ranged considerably from ≥48" to ≥63" for TMT_A, and from ≥125" to ≥200" for TMT_B.

Some researchers experimented with derivative raw-score-based TMT variables as EVIs. Busse & Whiteside (2012) found that the ≥190" cutoff for the sum of TMT_A and TMT_B approached .90 specificity. Subsequently, Shura et al., (2016) reported that a much more liberal cutoff (≥137") achieved the same level of specificity. Alternatively, Iverson et al. (2002) proposed that a $TMT_{B/A}$ raw score ratio of <1.50 separated credible from non-credible responding. An independent cross-validation by Merten et al., (2007) found this cutoff to be highly specific but insensitive to invalid performance.

Ashendorf et al., (2017) were the first to introduce TMT validity cutoffs based on demographically adjusted scores: the TMT_A T ≤37 and TMT_B T ≤35 reached the benchmark .90 specificity at .50-.53 sensitivity. These cutoffs failed to achieve the same level of specificity in a subsequent replication by Abeare et al. (2019). In their study, the first cutoff that cleared .90 specificity against all criterion measures was T ≤33 for both TMT_A and TMT_B, with .33-.68 sensitivity. More recently, Erdodi & Lichtenstein (2020) found TMT_A T ≤35 (.32-.43 sensitivity at .91-.95 specificity) and TMT_B T ≤37 (.32-.47 sensitivity at .86-.95 specificity) to be optimal cutoffs.

To our knowledge, there has been only one attempt at introducing validity cutoffs for the $TMT_{D\text{-}KEFS}$. Erdodi, Hurtubise et al. (2018) found that age-corrected scaled score (ACSS) cutoffs of ≤5 on $TMT_{D\text{-}KEFS\ 1\text{-}3}$, ≤4 on $TMT_{D\text{-}KEFS\ 4}$, and ≤8 on $TMT_{D\text{-}KEFS\ 5}$ effectively separated valid and invalid response sets. On the $TMT_{D\text{-}KEFS\ 4/2}$ raw score ratio (the conceptual equivalent to the $TMT_{B/A}$), the original cutoff of <1.50 had high specificity (.95-.96) but negligible sensitivity (.00-.07).

The Present Study

The purpose of this research project was two-fold: to replicate the validity cutoffs introduced by Erdodi, Hurtubise et al. (2018) in a different, non-clinical sample and to provide a direct comparison between the classic TMT and the D-KEFS version – both as EVIs and measures of visuomotor speed/executive skills. We hypothesized that $TMT_{A\ \&\ B}$ would be more sensitive to non-credible responding than $TMT_{D\text{-}KEFS}$, and produce superior classification accuracy. We also predicted that $TMT_{A\ \&\ B}$ would be more sensitive to natural fluctuations in psychomotor speed and executive skills compared to the $TMT_{D\text{-}KEFS}$, given the difference in score metrics [$TMT_{A\ \&\ B}$

expressed as T-scores (0-100) vs $TMT_{D\text{-}KEFS}$ expressed as ACSS (1-19)]. Presenting data on both tests and validating them against a common criterion measure would allow assessors to make empirically informed decisions regarding the relative merits of the two TMT versions and their respective EVIs.

Method

Participants. The sample consisted of 85 undergraduate students who volunteered for academic research at a midsize Canadian university. Mean age was 21.7 years (*SD* = 5.6), while mean education was 14.5 years (*SD* = 1.3). The majority of the sample (86%) was female, reflecting the true gender imbalance among students majoring in psychology at the university. The data were collected for the first author's Master's Thesis project, and were used in a previous publication examining different topics (*reference blinded for review*).

Table 1

Recoded Components of the EI-7 and Corresponding Base Rates of Failure

	EI-7 Values			
	0	**1**	**2**	**3**
EI-7 Component	Pass	Fail	FAIL	**FAIL**
Animals	>33	21-33	15-20	≤14
Base Rate	*77.4*	*13.1*	*4.8*	*4.8*
BNT-15 T2C	<75	75-85	86-100	>100
Base Rate	*88.2*	*2.4*	*2.4*	*6.0*
$CD_{WAIS\text{-}III}$	>6	6	5	≤4
Base Rate	*80.7*	*4.8*	*8.4*	*6.0*
CIM	>9	9	8	≤7
Base Rate	*74.1*	*10.6*	*2.4*	*12.9*
FAS	>33	28-33	20-27	≤19
Base Rate	*72.6*	*19.1*	*3.6*	*4.8*
$GPB_{Dominant}$	>31	19-31	12-18	≤11
Base Rate	*69.4*	*20.0*	*5.9*	*4.7*
RDS	>7	7	6	≤5
Base Rate	*82.1*	*7.1*	*1.2*	*9.5*

Note. EI-7: Erdodi Index Seven; Animals: Category fluency (demographically adjusted T-score based on norms by Heaton et al., 2004; Hayward et al., 1987; Hurtubise et al., 2020; Sugarman & Axelrod, 2015); BNT-15: Boston Naming Test – Short Form (An et al., 2019); T2C: Time to completion (seconds); $CD_{WAIS\text{-}III}$: Coding subtest of the Wechsler Adult Intelligence Scale – Third Edition (age-corrected scaled score; Ashendorf et al., 2017; Erdodi, Abeare et al., 2017; Erdodi & Lichtenstein, 2017; Etherton et al., 2006; Inman & Berry, 2002; Kim et al., 2010; Trueblood, 1994); CIM: Raw score on the Complex Ideational Material subtest of the Boston Diagnostic Aphasia Battery (An et al., 2019; Erdodi, 2019; Erdodi et al., 2016; Erdodi & Roth, 2017); FAS: Letter fluency (demographically adjusted T-score based on norms by Heaton et al., 2004; Curtis et al., 2008; Hurtubise et al., 2020; Sugarman & Axelrod, 2014; Whiteside et al., 2015); $GPB_{Dominant}$: Grooved Pegboard Test dominant hand (demographically adjusted T-score based on norms by Heaton et al., 2004; Erdodi, Kirsch, et al., 2018; Erdodi, Seke et al., 2017); RDS: Reliable Digit Span (Greiffenstein et al., 1994; Pearson, 2009; Reese, et al., 2012; Schroeder et al., 2012; Webber & Soble, 2018).

Materials. All participants were administered a brief neuropsychological battery. Demographically adjusted T-scores for the $TMT_{A\&B}$ were calculated using norms by Heaton et al. (2004). ACSSs for the $TMT_{D\text{-}KEFS}$ were calculated using norms provided by the technical manual (Delis et al., 2001). The $TMT_{D\text{-}KEFS}$ was administered in the first block of tests, followed by the $TMT_{A\&B}$.

The main free-standing PVT was the Word Choice Test (WCT; Pearson, 2009). An additional validity composite [Erdodi Index Seven (EI-7)] was developed by aggregating embedded PVTs into a single-number summary of the credibility of a given cognitive profile (Erdodi, 2019). Each of the EI-7 components was recoded onto a 4-point ordinal scale: a score in the unequivocally *valid* range was coded as 0, whereas a score in the unequivocally *invalid* range was coded as 3. A score that only failed the most liberal cutoff available was coded as 1, whereas the next more conservative failure was coded as 2. If clear *a priori* cutoffs were not available, the EI levels 2 and 3 were calibrated to correspond to the 10th and 5th percentiles, respectively.

The value of the EI-7 is obtained by summing its recoded components. As such, it can range from 0 (all 7 PVTs passed) to 21 (all 7 PVTs failed at the most conservative cut-off). An EI-7 score ≤1 is considered an overall *Pass*, while scores ≥4 are considered an overall *Fail* (Erdodi, 2019). EI-7 scores 2 and 3 are considered *Borderline* and excluded from analyses requiring a dichotomous outcome, as they are significantly different from both *Pass* and *Fail* (Erdodi & Abeare, 2020; Erdodi, Seke et al., 2017; Erdodi, Tyson et al., 2018; Lichtenstein et al., 2019). The majority (61.2%) of the sample produced valid profiles, while 20% had psychometric evidence of non-credible performance.

Table 2

Distribution of the EI-7 Scores and their Clinical Classification

				Classification	
EI-7	*f*	%	*Cumulative %*	By Row	Overall
0	33	38.8	38.8	PASS	PASS
1	19	22.4	61.2	Pass	
2	10	11.8	72.9	Borderline	
3	6	7.1	80.0	Borderline	
4	1	1.2	81.2	Fail	FAIL
5	3	3.5	84.7	Fail	
6	0	0.0	84.7	FAIL	
7	1	1.2	85.9	FAIL	
8	1	1.2	87.1	**FAIL**	
9	3	3.5	90.6	**FAIL**	
≥10	9	10.6	100.0	**FAIL**	

Note. EI-7: Erdodi Index Seven; *f*: Frequency distribution. Capitalization and boldface indicate increased confidence in correctly classifying a given profile as *invalid*.

Procedure. Participants were recruited through the university's online system designed to match student volunteers with research studies. Students were randomly assigned to either the control condition (i.e., tests were administered under standard instructions) or the experimental

malingering (*exp*MAL) condition (i.e., participants were instructed to feign neuropsychological deficits without being detected) using a 2:1 ratio. Data from five participants in the *exp*MAL group were excluded from the study because during the manipulation check at the end of the study they indicated that they had failed to comply with the instructions. The vignette used to induce *exp*MAL has been published in previous papers by the same research group (*references blinded for review*).

Psychometric testing was administered by the first author or undergraduate research assistants trained in test administration and scoring. Participants were provided information on the condition to which they were assigned (i.e., control or *exp*MAL) in a sealed envelope, and were instructed not to disclose that information to the individual who was administering the tests. The study was approved by the University's Research Ethics Board. APA ethical guidelines governing research with human participants were observed throughout the study.

Data Analysis. Descriptive statistics [*M, SD*, base rate of failure (BR_{Fail})] were reported where relevant. The main inferential statistics were one-way ANOVAs, independent-sample *t*-tests (for continuous variables), and χ^2 (for categorical variables). Effect size estimates were provided in partial eta squared (η^2_p), Cohen's *d,* and Φ^2. The statistical significance of the difference between two *SD*s was determined using Levene's test of homogeneity of variance. Receiver operating characteristics [area under the curve (AUC), and the corresponding 95% CIs] were computed in SPSS 25.0. Classification accuracy [sensitivity, specificity, likelihood ratio (LR)] was computed using standard formulas.

Results

Cross-Validating the EI-7

Although the EI-7 has been previously validated against both *exp*MAL in cognitively intact participants (An et al., 2019; Rai et al., 2019) and psychometrically defined known-groups within clinical samples (Erdodi, 2019; Erdodi, Green et al., 2019; Rai & Erdodi, 2019), its classification accuracy was computed to demonstrate its predictive validity within the present sample. As shown in Table 3, the EI-7 produced very high AUCs (.88-.96) and good combinations of sensitivity (.71-.93) and specificity (.91-.96).

Table 3

Classification Accuracy of the EI-7 against Various Criterion Measures

EI-7	Criterion		
Statistics	*exp*MAL	WCT	Rey-15
AUC	.96	.93	.88
95% CI	.91-1.00	.87-1.00	.80-.96
Sensitivity	.93	.92	.71
Specificity	.93	.91	.96
χ^2	43.3	39.5	35.6
p	<.001	<.001	<.001
Φ^2	.64	.57	.52
+LR	12.5	10.3	17.1
−LR	0.08	0.08	0.30

Note. EI-7: Erdodi Index Seven (*Fail* defined as ≥4; Erdodi, 2019); *exp*MAL: Experimental malingering; WCT: Word Choice Test [Pearson, 2009; *Fail* defined

as ≤45 (Barhon et al., 2015; Bain & Soble, 2019; Davis, 2014; Erdodi et al., 2014; Erdodi & Lichtenstein, 2019; Zuccato et al., 2018) or time-to-completion ≥156" (Erdodi & Lichtenstein, 2019; Erdodi, Tyson et al., 2017)]; Rey-15: Rey Fifteen Item Test (Rey, 1941) combination score (free recall + recognition hits; *Fail* defined as ≤23; Boone et al., 2002; Poynter et al., 2019).

Between-Group Differences on the TMTs as a Function of Performance Validity

Participants in the control group performed within the Average range on both versions of the TMT, and significantly above the *exp*MAL group (Table 4). Effect sizes were larger for the TMT_A (d = 1.41) than the $TMT_{D\text{-}KEFS\ 2}$ (d = 0.88), but comparable for the TMT_B and $TMT_{D\text{-}KEFS\ 4}$ (d: 0.65-0.79). On the raw score ratio, a significant difference emerged on $TMT_{B/A}$ (d = 0.94), but not on $TMT_{D\text{-}KEFS\ 4/2}$ (p = .811).

Table 4

Between-Group Differences on the TMTs as a Function of Performance Validity

TMT Version & Score	Criterion							
	Control		*exp*MAL					
	n = 60		*n* = 25					
	M	*SD*	M	*SD*	*t*	*p*	*d*	σ_1 vs. σ_2
TMT_A T-Score	51.4	*14.7*	30.6	*14.9*	5.77	<.001	1.41	.899
TMT_B T-Score	49.0	*9.5*	39.9	*11.4*	3.63	.001	0.69	.341
$TMT_{B/A}$ Raw Score Ratio	2.57	*0.86*	1.88	*0.66*	3.50	.001	0.90	.151
$TMT_{D\text{-}KEFS\ 2}$ ACSS	9.6	*2.7*	6.4	*4.4*	4.09	<.001	0.88	.002
$TMT_{D\text{-}KEFS\ 4}$ ACSS	9.4	*2.7*	6.9	*3.8*	3.46	.001	0.79	.005
$TMT_{D\text{-}KEFS\ 4/2}$ Raw Score Ratio	2.41	*0.75*	2.36	*1.06*	0.24	.811	-	.045
	WCT							
	Pass		Fail					
	n = 65		*n* = 17					
TMT_A T-Score	50.5	*14.4*	24.1	*10.7*	6.85	<.001	2.08	.086
TMT_B T-Score	48.7	*9.7*	35.0	*8.9*	4.85	<.001	1.47	.550
$TMT_{B/A}$ Raw Score Ratio	2.52	*0.87*	1.70	*0.41*	3.61	<.001	1.21	.009
$TMT_{D\text{-}KEFS\ 2}$ ACSS	9.5	*2.7*	4.9	*4.1*	5.55	<.001	1.33	.007
$TMT_{D\text{-}KEFS\ 4}$ ACSS	9.2	*2.8*	6.5	*4.2*	3.20	.002	0.76	.003
$TMT_{D\text{-}KEFS\ 4/2}$ Raw Score Ratio	2.51	*0.88*	1.98	*0.62*	2.31	.023	0.70	.257
	EI-7							
	Pass		Fail					
	n = 52		*n* = 15					

TMT_A T-Score	52.6	*13.9*	26.0	*13.4*	6.58	<.001	1.95	.540
TMT_B T-Score	49.7	*9.0*	34.1	*8.6*	5.76	<.001	1.77	.645
$TMT_{B/A}$ Raw Score Ratio	2.59	*0.89*	1.87	*0.70*	2.91	.005	0.93	.097
$TMT_{D\text{-}KEFS\ 2}$ ACSS	10.2	*2.3*	4.9	*4.1*	6.74	<.001	1.59	<.001
$TMT_{D\text{-}KEFS\ 4}$ ACSS	9.9	*2.3*	6.1	*4.2*	4.67	<.001	1.12	<.001
$TMT_{D\text{-}KEFS\ 4/2}$ Raw Score Ratio	2.48	*0.77*	2.13	*0.82*	1.61	.113	-	.683

Note. TMT: Trail Making Test (Reitan, 1955; demographically adjusted T-scores based on norms by Heaton et al., 2004); D-KEFS: Delis-Kaplan Executive Function System (Delis et al., 2001); ACSS: Age-corrected scaled score; *exp*MAL: Experimental malingering; σ vs. σ: The *p*-value associated with Levene's test of homogeneity of variance; WCT: Word Choice Test [Pearson, 2009; *Fail* defined as ≤45 (Barhon et al., 2015; Bain & Soble, 2019; Davis, 2014; Erdodi et al., 2014; Erdodi & Lichtenstein, 2019; Zuccato, Tyson, & Erdodi, 2018) or time-to-completion ≥156" (Erdodi & Lichtenstein, 2019; Erdodi, Tyson et al., 2017)]; EI-7: Erdodi Index Seven (*Fail* defined as ≥4; Erdodi, 2019).

A similar pattern of performance was observed against the WCT as the criterion measure: an Average range mean score within the *Pass* group with significantly lower scores in the *Fail* group; effect sizes more pronounced on the contrasts involving the TMT_A and $TMT_{D\text{-}KEFS\ 2}$ (d = 2.08 vs. 1.33) than the TMT_B and $TMT_{D\text{-}KEFS\ 4}$ (d = 0.91 vs 0.76). An important deviation from the previous trend was that a significant difference emerged on the $TMT_{D\text{-}KEFS\ 4/2}$ raw score ratio, although the effect size was lower than that on $TMT_{B/A}$ (d = 0.70 vs 1.25).

Results based on the EI-7 as criterion mirrored the trend observed against *exp*MAL: larger effects on the TMT_A and $TMT_{D\text{-}KEFS\ 2}$ contrasts (d = 1.95 vs 1.59) compared to the TMT_B and $TMT_{D\text{-}KEFS\ 4}$ (d = 1.10 vs 1.12), significant difference on the $TMT_{B/A}$ (d = 1.00), but not on $TMT_{D\text{-}KEFS\ 4/2}$ raw score ratio (p = .113).

Classification Accuracy of TMTs against Criterion PVTs

All TMT scores except the $TMT_{D\text{-}KEFS\ 4/2}$ raw score ratio produced significant AUCs against the criterion measures (0.72 - 0.92). AUC values were significantly higher on the TMT_A compared to $TMT_{D\text{-}KEFS\ 2}$ against *exp*MAL (0.83 vs. 0.72) and WCT (0.93 vs. 0.82), narrowly missing the threshold against the EI-7 (0.91 vs. 0.83). Likewise, the $TMT_{B/A}$ consistently outperformed the $TMT_{D\text{-}KEFS\ 4/2}$ (0.78-0.83 vs. 0.61-0.70). AUC values were comparable on TMT_B and $TMT_{D\text{-}KEFS\ 4}$ (Table 5).

Table 5

Overall Classification Accuracy of the TMTs against Various Criterion Measures

TMT Version & Score	Criterion Measure					
	*exp*MAL		WCT		EI-7	
	AUC	95% CI	AUC	95% CI	AUC	95% CI
TMT_A T-Score	.83	.73-.93	.93	.87-.99	.91	.82-.99
TMT_B T-Score	.73	.60-.86	.84	.74-.95	.90	.81-.99
$TMT_{B/A}$ Raw Score Ratio	.77	.65-.89	.82	.71-.92	.78	.64-.92
$TMT_{D\text{-}KEFS\ 2}$ ACSS	.72	.57-.87	.82	.67-.97	.83	.69-.98

$TMT_{D\text{-}KEFS\ 4}$ ACSS	.72	.59-.85	.72	.54-.89	.78	.61-.95
$TMT_{D\text{-}KEFS\ 4/2}$ Raw Score Ratio	.61	.46-.76	.70	.55-.85	.63	.45-.80

Note. TMT: Trail Making Test (Reitan, 1955; demographically adjusted T-scores based on norms by Heaton et al., 2004); D-KEFS: Delis-Kaplan Executive Function System (Delis et al., 2001); ACSS: Age-corrected scaled score; *exp*MAL: Experimental malingering; WCT: Word Choice Test [Pearson, 2009; *Fail* defined as ≤45 (Barhon et al., 2015; Bain & Soble, 2019; Davis, 2014; Erdodi et al., 2014; Erdodi & Lichtenstein, 2019; Zuccato et al., 2018) or time-to-completion ≥156" (Erdodi & Lichtenstein, 2019; Erdodi, Tyson et al., 2017)]; EI-7: Erdodi Index Seven (*Fail* defined as ≥4; Erdodi, 2019).

A TMT_A cutoff of T ≤37 produced a good combination of sensitivity (.80) and specificity (.87) against the EI-7, but failed to achieve the minimum specificity threshold against *exp*MAL and the WCT. There was no observed value for T = 35. At T ≤33, specificity was approaching .90 (.86-.88), at .61-.75 sensitivity. Lowering the cutoff to ≤31 further consolidated specificity (.88-.90) without sacrificing any sensitivity. T ≤29 achieved uniformly high specificity (.92-.95) at a small cost to sensitivity (.57-.69). Making the cutoff even more conservative (T ≤27) reached the point of diminishing return: unchanged specificity but a decline in sensitivity (.52-.63).

On the TMT_B , a cutoff of T ≤37 fell similarly short of minimum specificity standards against *exp*MAL and the WCT. However, lowering the cutoff to T ≤35 produced good combinations of sensitivity (.42-.60) and specificity (.87-.94). Making the cutoff even more conservative resulted in a predictable trade-off of near-ceiling specificity (.95-1.00) and diminished sensitivity (.33-.53).

A $TMT_{D\text{-}KEFS\ 2}$ ACSS ≤6 cutoff met the minimum specificity threshold against all criterion PVTs, with sensitivity ranging between .52 and .65. Lowering the cutoff to ≤5 resulted in significant improvement in specificity (.92-.98), at a reasonable cost to sensitivity (.36-.53). Making the cutoff even more conservative (≤4) had no meaningful effect on classification accuracy.

A $TMT_{D\text{-}KEFS\ 4}$ ACSS ≤6 cutoff produced good combinations of sensitivity (.44-.53) and specificity (.87-.92). Lowering the cutoff to ≤5 significantly improved specificity (.96-.98) at the expense of sensitivity (.28-.41). The next level of cutoff (≤4) was the point of diminishing return: no change in specificity, further decline in sensitivity (.20-.29).

A $TMT_{B/A}$ raw score ratio cutoff of ≤1.70 achieved good combinations of sensitivity (.56-.65) and specificity (.85-.88). Lowering the cutoff to ≤1.60 disproportionately sacrificed sensitivity (.44-.53) for specificity (.86-.90). Making the cutoff even more conservative (≤1.50) resulted in a sensible trade-off between markedly improved specificity (.94-.97) and relatively well-preserved sensitivity (.31-.41).

Finally, a $TMT_{D\text{-}KEFS\ 4/2}$ raw score ratio cutoff of ≤1.70 was highly specific (.92-.94), but not very sensitive (.20-.29) to psychometrically defined invalid performance. Lowering the cutoff to ≤1.60 resulted in the predictable trade-off between marginally improved specificity (.93-.96) at a proportional loss in sensitivity (.16-.24). Making the cutoff even more conservative (≤1.50) extended this trend by further consolidating specificity (.95-.98) and eroding sensitivity (.12-.18). Table 6 provides a detailed summary of the analyses.

Table 6

Sensitivity and Specificity of the TMTs against Various Criterion Measures

			Criterion Measure					
			*exp*MAL		WCT		EI-7	
		BR_{Fail}	29.4		20.2		24.6	
TMT Version & Score	Cutoff		SENS	SPEC	SENS	SPEC	SENS	SPEC
TMT_A T-Score	≤37	31.7	.65	.81	.88	.82	.80	.87
	≤35	-	-	-	-	-	-	-
	≤33	29.3	.61	.88	.75	.86	.73	.88
	≤31	24.4	.61	.90	.75	.88	.73	.90
	≤29	23.2	.57	.95	.69	.92	.67	.94
	≤27	19.5	.52	.95	.63	.92	.60	.94
TMT_B T-Score	≤37	27.6	.46	.81	.60	.80	.67	.85
	≤35	21.1	.42	.89	.53	.87	.60	.94
	≤33	11.8	.33	.98	.40	.95	.53	1.00
	≤31	7.9	.21	.98	.33	.98	.33	1.00
	≤29	5.3	.17	1.00	.27	1.00	.27	1.00
$TMT_{B/A}$ Raw Score Ratio	≤1.70	24.4	.56	.88	.65	.85	.56	.88
	≤1.60	20.7	.48	.90	.53	.86	.44	.90
	≤1.50	12.2	.36	.97	.41	.94	.31	.96
$TMT_{D\text{-}KEFS\ 2}$ ACSS	≤6	24.7	.52	.87	.59	.84	.65	.90
	≤5	16.5	.36	.92	.53	.93	.53	.98
	≤4	15.3	.36	.93	.53	.94	.53	.98
$TMT_{D\text{-}KEFS\ 4}$ ACSS	≤6	21.2	.44	.88	.53	.87	.53	.92
	≤5	17.6	.28	.97	.35	.96	.41	.98
	≤4	10.6	.20	.97	.24	.96	.29	.98
$TMT_{D\text{-}KEFS\ 4/2}$ Raw Score Ratio	≤1.70	11.8	.20	.92	.29	.93	.29	.94
	≤1.60	9.4	.16	.93	.24	.94	.24	.96
	≤1.50	7.1	.12	.95	.18	.96	.18	.98

Note. TMT: Trail Making Test (Reitan, 1955; demographically adjusted T-scores based on norms by Heaton et al., 2004); D-KEFS: Delis-Kaplan Executive Function System (Delis et al., 2001); ACSS: Age-corrected scaled score; BR_{Fail}: Base rate of failure (% of the sample that failed a given cut-off); *exp*MAL: Experimental malingering; WCT: Word Choice Test [Pearson, 2009; *Fail* defined as ≤45 (Barhon et al., 2015; Bain & Soble, 2019; Davis, 2014; Erdodi et al., 2014; Erdodi & Lichtenstein, 2019; Zuccato et al., 2018) or time-to-completion ≥156" (Erdodi & Lichtenstein, 2019; Erdodi, Tyson et al., 2017)]; EI-7: Erdodi Index Seven (*Fail* defined as ≥4; Erdodi, 2019); SENS: Sensitivity; SPEC: Specificity.

Optimal Cutoffs as a Function of BR_{Fail} across Versions of the TMT

As expected, a larger proportion of participants in the *exp*MAL condition failed both TMT_A and $TMT_{D\text{-}KEFS\ 2}$. However, the effect size was noticeably larger for TMT_A [Φ^2 = .281 vs. .116). Also, BR_{Fail} within the *exp*MAL group was 1.69 times higher on the TMT_A than $TMT_{D\text{-}KEFS\ 2}$. Similarly, there was a stronger association between criterion grouping and failing the TMT_B (Φ^2 =

.205) than failing the $TMT_{D\text{-}KEFS\ 4}$ ($\Phi^2 = .133$). While BR_{Fail} within the *exp*MAL group was similar on both versions of the test (33.3% vs. 28.0%), participants in the control group were 1.74 times more likely to fail the $TMT_{D\text{-}KEFS\ 4}$ compared to the TMT_B (Table 7).

Table 7

One-Way ANOVAs Comparing TMTs across Three Levels of Performance on Two Measures of Psychomotor Speed and Two Measures of Executive Functioning (Controls Only)

		Digit-Symbol Coding ACSS									Sig.	
		≤8[A]		9-12[B]		≥13[C]					*post*	
Test	Trial	M	*SD*	M	*SD*	M	*SD*	*F*	*p*	η^2_p	*hoc*s	*d*
TMT	A	43.4	*16.9*	50.8	*13.6*	58.7	*12.0*	3.62	.033	.116	A-C	1.04
	B	42.8	*8.7*	48.8	*9.5*	54.6	*6.8*	4.88	.012	.163	A-C	1.51
D-KEFS	2	8.7	*2.9*	9.8	*2.2*	10.1	*2.7*	1.06	.353	.037	-	-
	4	8.2	*2.6*	9.2	*2.7*	11.2	*1.7*	4.60	.014	.143	A-C	1.37
											B-C	0.89
		GPB Dominant Hand T-score										
		≤39[A]		40-50[B]		≥51[C]						
TMT	A	49.9	*15.5*	48.7	*14.7*	57.4	*12.4*	1.89	.160	.063	B-C	.64
	B	48.7	*10.5*	47.5	*9.2*	51.4	*8.3*	0.69	.504	.027	-	-
D-KEFS	2	8.6	*2.9*	9.6	*2.6*	10.8	*2.2*	2.95	.061	.094	A-C	.85
	4	8.6	*2.9*	9.4	*2.8*	10.7	*1.8*	2.95	.060	.094	A-C	.87
		WCST-64 Categories Completed										
		≤2[A]		3-4[B]		5[C]						
TMT	A	51.0	*15.7*	47.9	*13.6*	58.3	*14.5*	2.97	.060	.096	B-C	0.74
	B	46.8	*5.8*	47.5	*10.1*	52.9	*9.3*	2.02	.143	.075	-	-
D-KEFS	2	8.9	*2.2*	9.1	*3.1*	10.8	*1.4*	2.82	.068	.090	B-C	0.71
	4	9.7	*1.8*	8.9	*3.1*	10.3	*2.1*	1.47	.238	.049	-	-
		Longest Digit Span Backward										
		3[A]		4[B]		≥5[C]						
TMT	A	52.7	*13.0*	48.8	*15.4*	54.6	*15.7*	0.81	.451	.039	-	-
	B	46.9	*9.6*	46.9	*9.5*	53.8	*8.2*	3.00	.059	.109	A-C	0.77
											B-C	0.78
D-KEFS	2	9.9	*1.9*	9.1	*2.7*	9.7	*3.4*	0.51	.601	.018	-	-
	4	9.4	*1.7*	9.3	*3.0*	9.7	*3.3*	0.13	.875	.005	-	-

Note. ACSS: Age-corrected scaled score; TMT: Trail Making Test demographically adjusted T-scores based on norms by Heaton et al., 2004; D-KEFS: Delis-Kaplan Executive Function System ACSS; Sig. *post hocs*: Uncorrected significant *post hoc* pairwise contrasts; GPB: Grooved Pegboard Test (demographically corrected T-scores demographically adjusted T-scores based on norms by Heaton et al., 2004); WCST-64: Wisconsin Card Sorting Test, 64-card version.

Sensitivity to Various Levels of Performance on Tests of Psychomotor Speed and Executive Function (Controls Only)

Scores on two measures of processing speed [Digit-Symbol Coding (DSC) and Grooved Pegboard (GPB)] and executive skills [categories completed on the 64-card version of the Wisconsin Card Sorting Test (WCST-64$_{CAT}$) and Longest Digit Span Backward (LDB)] were trichotomized to establish three levels of performance: low, average, and high (Table 8). The three subsamples were defined either by established classification thresholds (DSC and LDB), the distribution of scores within the present sample (WCST-64$_{CAT}$), or a combination of both (GPB). One-way ANOVAs were performed using the three levels of these tests as the independent variable, and the two versions of the TMT as independent variables.

DSC – ACSSs

A significant ANOVA emerged on TMT$_A$ with a large main effect (η^2_p = .116). The only significant *post hoc* contrast was between the ACSS ≤8 and ACSS ≥13 subsamples (d = 1.04, large effect). Even larger effects were observed on TMT$_B$ [η^2_p = .163 (large) and d = 1.51 (very large)]. The ANOVA on TMT$_{D\text{-}KEFS\ 2}$ failed to reach significance, as did all *post hoc* contrasts. However, a significant ANOVA emerged on TMT$_{D\text{-}KEFS\ 4}$ with a large main effect (η^2_p = .114). The ACSS ≥13 subsample outperformed both of the other two groups (d: 0.89-1.37, large effects).

Dominant Hand GPB – Demographically Adjusted T-scores

ANOVAs were non-significant for all four trials of the TMTs. However, the main effect was more pronounced on TMT$_A$ (η^2_p = .063, medium) than TMT$_B$ (η^2_p = .027, small). Similarly, a significant *post hoc* contrast emerged between the T = 40-50 vs. T ≥51 on TMT$_A$ (d = 0.64, medium effect). In contrast, the same main effect (η^2_p = .094, medium) emerged on both trials of the TMT$_{D\text{-}KEFS}$, approaching statistical significance (p: .060-.061). *Post hoc* contrasts between the T ≤39 and T ≥51 subsamples were also significant (d: 0.85-0.87, large effects).

Categories Completed on WCST-64 – Raw Score

The ANOVA emerged on the TMT$_A$ was approaching significance (p = .060), with a medium main effect (η^2_p = .096). The contrast between participants who completed ≤2 and those who completed 5 categories was also significant (d = 0.74, large effect). The ANOVA on the TMT$_B$ was not significant (p = .143), and neither were any of the pairwise *post hoc* contrasts. A similar pattern emerged on the TMT$_{D\text{-}KEFS}$: marginally significant ANOVA on Trails 2 (p = .068), non-significant ANOVA on Trails 4 (p = .238). Participants who completed 5 categories outperformed those who completed ≤2 categories (d = 0.71, large effect) on TMT$_{D\text{-}KEFS\ 4}$.

LDB – Raw Score

The ANOVA on TMT$_A$ failed to reach significance, as did all *post hoc* contrasts. However, the ANOVA on TMT$_B$ was approaching significance (p = .109). Participants with LDB ≥5 outperformed the other two groups (d: 0.77-0.78, large effects). In contrast, ANOVAs were not significant for both trials of the TMT$_{D\text{-}KEFS}$ (p: .601-.875), and neither were any of the *post hoc* contrasts.

Discussion

Overview

The present study compared the sensitivity of the TMT$_{A\ \&\ B}$ and TMT$_{D\text{-}KEFS\ 2\ \&\ 4}$ to invalid performance and natural fluctuations in psychomotor speed/executive function in undergraduate students, some of whom were instructed to feign neuropsychological deficits. We predicted that TMT$_{A\ \&\ B}$ would be superior to the TMT$_{D\text{-}KEFS}$ in both applications. However, given the divergent findings, the results provide mixed support for these hypotheses, as discussed below.

Sensitivity to Invalid Performance as Continuous Variables

TMT $_A$ was more sensitive to both *exp*MAL and psychometrically defined non-credible responding than TMT$_{D\text{-}KEFS\ 2}$ (*d*: 1.41-2.08 vs. 0.88-1.59). However, invalid performance had a comparable effect size on both TMT$_B$ and TMT$_{D\text{-}KEFS\ 4}$ (*d*: 0.65-1.10 vs. 0.76-1.12). The advantage of TMT$_{A\ \&\ B}$ over TMT$_{D\text{-}KEFS}$ re-emerged on the ratio score: B/A consistently produced a large effect (*d*: 0.94-1.25), whereas 4/2 failed to reach statistical significance against two of the three criterion measures (see Table 4 above).

Equally important, statistically significant heterogeneity of variance was consistently observed with TMT$_{D\text{-}KEFS}$, but not TMT$_{A\ \&\ B}$. Increased within-group variability has been identified as an emergent marker of invalid performance (Elbaum et al., 2019; Erdodi, 2019; Fuermaier et al., 2018; Lichtenstein et al., 2019; Stevens et al., 2016), possibly reflecting a divergence in malingering strategies (Cottingham et al., 2014; Erdodi et al., 2014). As the *SD* is the ultimate source of measurement error, the observed heteroscedasticity has potentially far-reaching implications to the instruments' clinical utility.

TMT$_{A\ \&\ B}$ vs TMT$_{D\text{-}KEFS}$ as EVIs

The TMT$_A$ produced significantly higher AUC values than *TMT$_{D\text{-}KEFS\text{-}2}$* against two of the three criterion PVTs. However, there was no difference in the signal detection performance of TMT$_B$ and *TMT$_{D\text{-}KEFS\text{-}4}$*. The B/A had a superior AUC to the 4/2 raw score ratio against all three criterion PVTs (see Table 5 above).

The TMT$_A$ cutoff of T ≤35 had comparable classification accuracy to TM*T$_{D\text{-}KEFS\text{-}2}$* ACSS ≤6 (.61-.75 sensitivity at .86-.88 specificity and .52-.65 sensitivity at .84-.90 specificity, respectively). The first cutoff to unequivocally achieve the .90 specificity benchmark was TMT$_A$ T ≤29 (.57-.67 sensitivity) and TM*T$_{D\text{-}KEFS\text{-}2}$* ACSS ≤5 (.36-.53 sensitivity). Similarly, the TMT$_B$ cutoff of T ≤35 had comparable classification accuracy to TM*T$_{D\text{-}KEFS\text{-}4}$* ACSS ≤6 (.42-.60 sensitivity at .87-.94 specificity and .44-.53 sensitivity at .87-.92 specificity, respectively). The first cutoff to unequivocally achieve the .90 specificity benchmark was TMT$_B$ T ≤33 (.33-.53 sensitivity) and TM*T$_{D\text{-}KEFS\text{-}2}$* ACSS ≤5 (.28-.41 sensitivity).

In terms of the ratio score, the superiority of B/A over 4/2 was apparent, likely driven by the higher (roughly twofold) BR$_{Fail}$. At the original cutoff (≤1.50), both versions had comparably high specificity (.94-.97 and .95-.98, respectively), but B/A achieved notably higher sensitivity than 4/2 (.31-.41 vs. .12-.18, respectively). At the liberal cutoff of ≤1.70, 4/2 maintained higher specificity (.92-.94) than B/A (.85-.88). However, B/A had much higher sensitivity (.56-.65 vs. .20-.29).

Classification Error Rate of the Two Versions of TMT as EVIs

Around 10% of participants in the control group failed the TMT$_A$ and TMT$_{D\text{-}KEFS\ 2}$ at optimal cutoffs (T ≤31 and ACSS ≤5). This is considered an acceptable false positive rate (Boone, 2013; Donders & Strong, 2011). However, the TMT$_A$ correctly identified a larger proportion of the *exp*MAL group (61% vs. 36%). In contrast, a comparably low BR$_{Fail}$ was observed on TMT$_B$ (33%) and TMT$_{D\text{-}KEFS\ 4}$ (28%), while both versions of the test had a low false positive rate (≤3%). Overall, results suggest that the TMT$_A$ makes for a better EVI than the TMT$_{D\text{-}KEFS\ 2}$; both the TMT$_B$ and TMT$_{D\text{-}KEFS\ 4}$ are relatively insensitive to experimentally induced non-credible performance but are highly specific to it.

Sensitivity to Fluctuations in Processing Speed and Executive Function

Consistent with our prediction, TMT$_A$ demonstrated a superior ability to differentiate variability in psychomotor speed performance on DSC relative to TMT$_{D\text{-}KEFS\ 2}$. The TMT$_A$ and TMT $_{D\text{-}KEFS\ 2}$ were comparably insensitive to fluctuations in GPB scores. As expected, TMT$_A$ was more sensitive to fluctuations in DSC scores than TMT$_B$. However, the opposite pattern was

observed in $TMT_{D\text{-}KEFS}$: Trial 4 was more sensitive to fluctuations in DSC scores than Trial 2. These results raise concerns about the construct validity of the $TMT_{D\text{-}KEFS}$.

Contrary to expectations, both TMT_{B} and $TMT_{D\text{-}KEFS\ 4}$ were insensitive to fluctuations in WCST-64_{CAT}, while TMT_{A} and $TMT_{D\text{-}KEFS\ 2}$ were approaching significance at comparable effect sizes. Negative findings extended to TMT_{A} and $TMT_{D\text{-}KEFS\ 2}$ on LDB. However, a large effect emerged on TMT_{B}, whereas the effect size was virtually zero on the $TMT_{D\text{-}KEFS\ 4}$. Once again, these findings suggest that $TMT_{A\ \&\ B}$ have stronger construct validity than $TMT_{D\text{-}KEFS\ 2\ \&\ 4}$.

Limitations

Results must be interpreted in the context of the study's inherent limitations. First, the *exp*MAL design has a number of well-known methodological flaws, the main one being that participants lack a proportional incentive and the experiential basis of real-world malingerers to produce credible impairment on testing (Erdal et al., 2004; Kanser et al., 2017; Tan et al., 2002). In addition, in clinical or forensic settings, there is a range of complicating factors such as psychogenic interference and somatoform disorders (Jurick et al., 2020; Merten & Merckelbach, 2013; Tarachow, 1947) that either hinder or facilitate the detection of non-credible responding. Undergraduate students cannot be reasonably expected to emulate such complex and ill-defined patterns of psychopathology. The variability in the wording of *exp*MAL instruction is another largely unknown source of error variance (Giromini et al., 2019).

Second, the study was based on cognitively healthy young adults. Although using normal controls to establish a proof of concept is a sensible strategy in the early stages of an investigation, it leaves unanswered questions about the generalizability of the findings. Namely, it is unclear whether the differences between the two versions of the TMT would extend to individuals with genuine neurocognitive deficits.

Third, even cognitively healthy young adults may possess mild forms of motor impairment (Kirby et al., 2008; Saban & Kirby, 2018) that may negatively impact TMT performance. Beyond looking for consistency between the GPB-D and $D\text{-}KEFS_{5}$ scores, it may be helpful to document idiosyncratic motoric behaviors such as peg drops, tremulousness, an inefficient pencil grasp, frequent pencil lifts, over and undershoots, slow oromotor speed, speech errors, and nystagmus. In the presence of apparent multi-domain weaknesses across measures, such an integrative approach that combines behavioral observations with psychometric data may assist in distinguishing genuine deficits from non-credible responding.

Finally, the sample was restricted to a single city. Given previous reports of geographic variations in cognitive functioning (Daugherty et al., 2017; Kura et al., 2013; Lichtenstein et al., 2019; McDaniel, 2006; Roth et al., 2015), the findings may not transfer to populations with different demographic characteristics.

Strengths

Nevertheless, to our knowledge, this is the first study to provide a systematic comparison between the classic and D-KEFS versions of the TMT, cautioning against uncritical adaptation of novel editions. It is also the first study to evaluate the classification accuracy of the $TMT_{D\text{-}KEFS}$ as a PVT using the *exp*MAL paradigm. The randomization scheme followed a single-blind design to minimize assessor bias and demand characteristics. Results also revealed that the TMT ratio continues to have potential as a derivative validity indicator, despite its poor track record in previous research (Abeare et al., 2019; Ashendorf et al., 2017; Iverson et al., 2002). Moreover, psychometrically defined criterion groups complemented the analyses to mitigate the methodological limitations inherent in expMAL paradigms. Nonetheless, the ultimate purpose of

including the *exp*MAL condition was to induce a certain amount of invalid performance and evaluate the TMTs' detection efficacy.

Conclusions

Results of the study reinforce the principle that new and more sophisticated editions of an old test are not necessarily psychometrically superior (Boone, 2013). In fact, there may be advantages of utilizing old, well-established instruments with a rich post-publication evidence base. Should the differential sensitivity of the classic TMT be upheld in future independent replications, clinical neuropsychologists will be provided with important actionable information about test selection and interpretation. Simultaneously, replication studies might help to explain some of the negative findings wherein the $TMT_{D\text{-}KEFS}$ were employed as an outcome measure. Regardless, the present findings indicate that greater consideration should be given to administering the $TMT_{A\ \&\ B}$ over the $TMT_{D\text{-}KEFS}$.

About the Authors

Laszlo Erdodi, PhD is with the Department of Psychology, University of Windsor, Windsor ON, Canada. Correspondence concerning this article should be addressed to Laszlo Erdodi at lerdodi(at)gmail.com.

Jessica Hurtubise, MA is with the Department of Psychology, University of Windsor, Windsor ON, Canada.

Maame Brantuo, MA is with the Department of Psychology, University of Windsor, Windsor ON, Canada.

Laura Cutler, BA is with the Department of Psychology, University of Windsor, Windsor ON, Canada.

Arianna Kennedy, BSW is with the Department of Psychology, University of Windsor, Windsor ON, Canada.

Rayna Hirst, PhD is with the Neuropsychology Program, University of Palo Alto, Palo Alto CA, USA

Author Note

This project received no financial support from outside funding agencies. Relevant ethical guidelines were followed throughout the project. All data collection, storage, and processing were done with the approval of relevant institutional authorities regulating research involving human participants, in compliance with the 1964 Helsinki Declaration and its subsequent amendments or comparable ethical standards.

References

Abeare, C., Sabelli, A., Taylor, B., Holcomb, M., Dumitrescu, C., Kirsch, N., & Erdodi, L. (2019). The importance of demographically adjusted cutoffs: Age and education bias in raw score cutoffs within the Trail Making Test. *Psychological Injury and Law, 12*(2), 170-182. https://doi.org/10.1007/s12207-019-09353-x

Alverson W. A., O'Rourke, J. J. F., & Soble, J. R. (2019). The Word Memory Test genuine memory impairment profile discriminates genuine memory impairment from invalid performance in a mixed clinical sample with cognitive impairment. *The Clinical Neuropsychologist, 21*(2), 209-231. https://doi.org/10.1080/13825580601025932

An, K. Y., Charles, J., Ali, S., Enache, A., Dhuga, J., & Erdodi, L. A. (2019). Re-examining performance validity cutoffs within the Complex Ideational Material and the Boston Naming Test-Short Form

using an experimental malingering paradigm. *Journal of Clinical and Experimental Neuropsychology, 41*(1), 15-25. https://doi.org/10.1080/13803395.2018.1483488

Ashendorf, L., Clark, E. L., & Sugarman, M. A. (2017). Performance validity and processing speed in a VA polytrauma sample. *The Clinical Neuropsychologist, 31*(5), 857-866. https://doi.org/10.1080/13854046.2017.1285961

Backhaus, S. L., Fichtenberg, N. L., & Hanks, R. A. (2004). Detection of sub-optimal performance using a floor effect strategy in patients with traumatic brain injury. *The Clinical Neuropsychologist, 18*(4)*, 591-603*. https://doi.org/10.1080/13854040490888558

Bain, K. M., & Soble, J. R. (2019). Validation of the Advanced Clinical Solutions Word Choice Test (WCT) in a mixed clinical sample: Establishing classification accuracy, sensitivity/specificity, and cutoff scores. *Assessment, 26*(7), 1320–1328. https://doi.org/10.1177/1073191117725172

Barhon, L. I., Batchelor, J., Meares, S., Chekaluk, E., & Shores, E. A. (2015). A comparison of the degree of effort involved in the TOMM and the ACS Word Choice Test using a dual-task paradigm. *Applied Neuropsychology: Adult, 22,* 114-123. https://doi.org/10.1080/23279095.2013.863775

Bauer, L., & McCaffrey, R. J. (2006). Coverage of the Test of Memory Malingering, Victoria Symptom Validity Test, and Word Memory Test on the internet: Is test security threatened. *Archives of Clinical Neuropsychology, 21*(1), 121-126. https://doi.org/10.1016/j.acn.2005.06.010

Bialystok, E., Craik, F., Binns, M. A., Ossher, L., & Freedman, L. (2014). Effects of bilingualism on the age of onset and progression of MCI and AD: Evidence from executive function test. *Neuropsychology, 28*(2), 290-304. https://doi.apa.org/doi/10.1037/neu0000023

Bigler, E. D. (2014). Effort, symptom validity testing, performance validity testing and traumatic brain injury. *Brain Injury, 28*(13-14), 1623-1638. https://doi.org/10.3109/02699052.2014.947627

Boone, K. B. (2009). The need for continuous and comprehensive sampling of effort/response bias during neuropsychological examination. *The Clinical Neuropsychologist, 23*(4), 729-741. https://doi.org/10.1080/13854040802427803

Boone, K. B. (2013). *Clinical practice of forensic neuropsychology*. The Guilford Press.

Boone, K. B., Salazar, X., Lu, P., Warner-Chacon, K., & Razani, J. (2002). The Rey 15-item recognition trial: A technique to enhance sensitivity of the Rey 15-item Memorization Test. *Journal of Clinical and Experimental Neuropsychology, 24*(5), 561-573. https://doi.org/10.1076/jcen.24.5.561.1004

Bornstein, R. A. (1985). Normative data on selected neuropsychological measures from a nonclinical sample. *Journal of Clinical Psychology, Issue/pages?*. https://doi.org/10.1002/1097-4679(198509)41:5%3C651::AID-JCLP2270410511%3E3.0.CO;2-C

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*(3), 197-205. https://psycnet.apa.org/doi/10.1007/s12207-014-9198-7

Bush, S. S., Ruff, R. M., Tröster, A. I., Barth, J. T., Koffler, S. P., Pliskin, N. H., Reynolds, C. R., & Silver, C. H. (2005). Symptom validity assessment: Practice issues and medical necessity (NAN Policy

and Planning Committees). *Archives of Clinical Neuropsychology, 20*(4), 419–426. https://doi.org/10.1016/j.acn.2005.02.002

Busse, M., & Whiteside, D. (2012). Detecting suboptimal cognitive effort: Classification accuracy of the Conners' Continuous Performance Test-II, Brief Test of Attention, and Trail Making Test. *The Clinical Neuropsychologist, 26*(4), 675-687. https://doi.org/10.1080/13854046.2012.679623

Chafetz, M. D., Williams, M. A., Ben-Porath, Y. S., Bianchini, K. J., Boone, K. B., Kirkwood, M. W., Larrabee, G. J., & Ord, J. S. (2015). Official position of the American Academy of Clinical Neuropsychology Social Security Administration policy on validity testing: Guidance and recommendations for change. *The Clinical Neuropsychologist, 29*(6), 723-740. https://doi.org/10.1080/13854046.2015.1099738

Cottingham, M. E., Victor, T. L., Boone, K. B., Ziegler, E. A., & Zeller, M. (2014). Apparent effect of type of compensation seeking (disability vs. litigation) on performance validity test scores may be due to other factors. *The Clinical Neuropsychologist, 28*(6), 1030-1047. https://doi.org/10.1080/13854046.2014.951397

Craun, E., Lachance, K., Williams, C., & Wong, M. M. (2019). Parent depressive symptoms and offspring executive functioning. *Journal of Clinical and Experimental Neuropsychology, 41*(2), 147-157. https://doi.org/10.1080/13803395.2018.1504893

Curtis, K. L., Thompson, L. K., Greve, K. W., & Bianchini, K. J. (2008). Verbal fluency indicators of malingering in traumatic brain injury: Classification accuracy in known groups. *The Clinical Neuropsychologist, 22*, 930-945. https://doi.org/10.1080/13854040701563591

Dandachi-Fitzgerald, B., Ponds, R. W. H. M., & Merten, T. (2013). Symptom validity and neuropsychological assessment: A survey of practices and beliefs of neuropsychologists in six European countries. *Archives of Clinical Neuropsychology, 28*(8), 771-783. https://doi.org/10.1093/arclin/act073

Daugherty, J. C., Puente, A. E., Fasfous, A. F., Hidalgo-Ruzzante, N., & Pérez-Garcia, M. (2017). Diagnostic mistakes of culturally diverse individuals when using North American neuropsychological tests. *Applied Neuropsychology: Adult, 24*(1), 16-22. https://doi.org/10.1080/23279095.2015.1036992

Davis, J. J. (2014). Further consideration of Advanced Clinical Solutions Word Choice: Comparison to the Recognition Memory Test – Words and classification accuracy on a clinical sample. *The Clinical Neuropsychologist, 28*(8), 1278-1294. https://doi.org/10.1080/13854046.2014.975844

Delis, D. C., Kaplan, E. F., & Kramer, J. H. (2001). *Delis-Kaplan Executive Function System* (D-KEFS): *Technical Manual.* Psychological Corporation.

Donders, J., & Strong, C. H. (2011). Embedded effort indicators on the California Verbal Learning Test – Second Edition (CVLT-II): An attempted cross-validation. *The Clinical Neuropsychologist, 25*, 173-184. https://doi.org/10.1080/13854046.2010.536781

Elbaum, T., Golan, L., Lupu, T., Wagner, M., & Braw, Y. (2019). Establishing supplementary response time validity indicators in the Word Memory Test (WMT) and directions for future research. *Applied Neuropsychology: Adult, 27*(5), 403-413. https://doi.org/10.1080/23279095.2018.1555161

Erdal, K. (2004). The effects of motivation, coaching, and knowledge of neuropsychology on the simulated malingering of head injury. *Archives of Clinical Neuropsychology, 19*(1), 73–88. https://doi.org/10.1016/S0887-6177(02)00214-7

Erdodi, L. A. (2019). Aggregating validity indicators: The salience of domain specificity and the indeterminate range in multivariate models of performance validity assessment. *Applied Neuropsychology: Adult, 26(*2), 155-172. https://doi.org/10.1080/23279095.2017.1384925

Erdodi, L. A., & Abeare, C. A. (2020). Stronger together: The Wechsler Adult Intelligence Scale – Fourth Edition as a multivariate performance validity test in patients with traumatic brain injury. *Archives of Clinical Neuropsychology, 35*(2), 188-204. https://doi.org/10.1093/arclin/acz032

Erdodi, L. A., Abeare, C. A., Lichtenstein, J. D., Tyson, B. T., Kucharski, B., Zuccato, B. G., & Roth, R. M. (2017). WAIS-IV processing speed scores as measures of non-credible responding – The third generation of embedded performance validity indicators. *Psychological Assessment, 29*(2), 148-157. https://doi.org/10.1037/pas0000319

Erdodi, L. A., Dunn, A. G., Seke, K. R., Charron, C., McDermott, A., Enache, A., Maytham, C., & Hurtubise, J. (2018). The Boston Naming Test as a measure of performance validity. *Psychological Injury and Law, 11*, 1-8. https://doi.org/10.1007/s12207-017-9309-3

Erdodi, L. A., Green, P., Sirianni, C., & Abeare, C. A. (2019). The myth of high false positive rates on the Word Memory Test in mild TBI. *Psychological Injury and Law, 12*(2), 155-169. https://doi.org/10.1007/s12207-019-09356-8

Erdodi, L. A., Hurtubise, J. L., Charron, C., Dunn, A., Enache, A., McDermott, A., & Hirst, R. (2018). The D-KEFS Trails as performance validity tests. *Psychological Assessment, 30*(8), 1081-1095. https://doi.apa.org/doi/10.1037/pas0000561

Erdodi, L. A., Kirsch, N. L., Sabelli, A. G., & Abeare, C. A. (2018). The Grooved Pegboard Test as a validity indicator – A study on psychogenic interference as a confound in performance validity research. *Psychological Injury and Law, 11*(4), 307-324. https://doi.org/10.1007/s12207-018-9337-7

Erdodi, L. A., & Lichtenstein, J. D. (2017). Invalid before impaired: An emerging paradox of embedded validity indicators. *The Clinical Neuropsychologist, 31*(6-7), 1029-1046. https://doi.org/10.1080/13854046.2017.1323119

Erdodi, L. A., & Lichtenstein, J. D. (2019). Information processing speed tests as PVTs. In K. B. Boone (Ed.), *Assessment of feigned cognitive impairment. A neuropsychological perspective* (pp. X-X). Guilford Publications.

Erdodi, L. A., & Roth, R. M. (2017). Low scores on BDAE Complex Ideational Material are associated with invalid performance in adults without aphasia. *Applied Neuropsychology: Adult, 24*(3), 264-274. https://doi.org/10.1080/23279095.2016.1154856

Erdodi, L. A., Roth, R. M., Kirsch, N. L., Lajiness-O'Neill, R., & Medoff, B. (2014). Aggregating validity indicators embedded in Conners' CPT-II outperforms individual cutoffs at separating valid from invalid performance in adults with traumatic brain injury. *Archives of Clinical Neuropsychology, 29*(5), 456-466. https://doi.org/10.1093/arclin/acu026

Erdodi, L. A., Seke, K. R., Shahein, A., Tyson, B. T., Sagar, S., & Roth, R. M. (2017). Low scores on the Grooved Pegboard Test are associated with invalid responding and psychiatric symptoms. *Psychology and Neuroscience, 10*(3), 325-344. https://psycnet.apa.org/doi/10.1037/pne0000103

Erdodi, L. A., Tyson, B. T., Abeare, C. A., Lichtenstein, J. D., Pelletier, C. L., Rai, J. K., & Roth, R. M. (2016). The BDAE Complex Ideational Material – A measure of receptive language or performance validity? *Psychological Injury and Law, 9*, 112-120. https://doi.org/10.1007/s12207-016-9254-6

Erdodi, L. A., Tyson, B. T., Abeare, C. A., Zuccato, B. G., Rai, J. K., Seke, K. R., Sagar, S., & Roth, R. M. (2018). Utility of critical items within the Recognition Memory Test and Word Choice Test. *Applied Neuropsychology: Adult, 25*(4), 327-339. https://doi.org/10.1080/23279095.2017.1298600

Erdodi, L. A., Tyson, B. T., Shahein, A., Lichtenstein, J. D., Abeare, C. A., Pelletier, C. L., Zuccato, B. G., Kucharski, B., & Roth, R. M. (2017). The power of timing: Adding a time-to-completion cutoff to the Word Choice Test and Recognition Memory Test improves classification accuracy. *Journal of Clinical and Experimental Neuropsychology*, *39*(4), 369-383. https://doi.org/10.1080/13803395.2016.1230181

Etherton, J. L., Bianchini, K. J., Heinly, M. T., & Greve, K. W. (2006). Pain, malingering, and performance on the WAIS-III Processing Speed Index. *Journal of Clinical and Experimental Neuropsychology, 28*(7)*,* 1218-1237. https://doi.org/10.1080/13803390500346595

Fuermaier, A. B. M., Tucha, O., Koerts, J., Send, T. S., Weisbrod, M., Aschenbrenner, S., & Tucha, L. (2018). Is motor activity during cognitive assessment an indicator for feigned attention-deficit/hyperactivity disorder (ADHD) in adults? *Journal of Clinical and Experimental Neuropsychology, 40*(10), 971-986. https://doi.org/10.1080/13803395.2018.1457139

Gass, C. S., & Daniel, S. K. (1990). Emotional impact on Trail Making Test performance. *Psychological Reports, 67*(2), 435-438. https://doi.org/10.2466%2Fpr0.1990.67.2.435

Gervais, R. O., Ben-Porath, Y. S., Wygant, D. B., & Green, P. (2007). Development and validation of a Response Bias Scale (RBS) for the MMPI-2. *Assessment, 14*(2), 196-208 https://doi.org/10.1177/1073191106295861

Giromini, L., Viglione, D. J., Pignolo, C., & Zennaro, A. (2019). An Inventory of Problems–29 sensitivity study investigating feigning of four different symptom presentations via malingering experimental paradigm. *Journal of Personality Assessment, 102*(4), 563-572. https://doi.org/10.1080/00223891.2019.1566914

Goebel, R. A. (1983). Detection of faking on the Halstead-Reitan Neuropsychological Test Battery. *Journal of Clinical Psychology, 39*(5), 731-742. https://doi.org/10.1002/1097-4679(198309)39:5%3C731::AID-JCLP2270390515%3E3.0.CO;2-T

Greenlief, C. L., Margolis, R. B., & Erker, G. J. (1985). Application of the Trail Making Test in differentiating neuropsychological impairment of elderly persons. *Perceptual and Motor Skills, 61*(3), 1283-1289. https://doi.org/10.2466%2Fpms.1985.61.3f.1283

Greer, S. E., Brewer, K. K., Cannici, J. P., & Pennett, D. L. (2010). Level of performance accuracy for core Halstead-Reitan measures by pooling normal controls from published studies: Comparison with existing norms in a clinical sample. *Perceptual and Motor Skills, 111*(4), 3-18. https://doi.org/10.2466%2F03.22.27.PMS.111.4.3-18

Greiffenstein, M. F., Baker, W. J., & Gola, T. (1994). Validation of malingered amnesia measures with a large clinical sample. *Psychological Assessment, 6*(3), 218-224. https://doi.apa.org/doi/10.1037/1040-3590.6.3.218

Hayward, L., Hall, W., Hunt, M., & Zubrick, S. R. (1987). Can localized brain impairment be simulated on neuropsychological test profiles? *Australian and New Zealand Journal of Psychiatry, 21*, 87-93. https://doi.org/10.3109%2F00048678709160904

Heaton, R. K., Miller, S. W., Taylor, M. J., & Grant, I. (2004). *Revised comprehensive norms for an expanded Halstead-Reitan battery: Demographically adjusted neuropsychological norms for African American and Caucasian adults*. Psychological Assessment Resources.

Heaton, R. K., Smith, H. H., Lehman, R. A. W., & Vogt, A. T. (1978). Prospects for faking believable deficits on neuropsychological testing. *Journal of Consulting and Clinical Psychology, 46*(5), 892-900. https://doi.org/10.1037/0022-006X.46.5.892

Heidler-Gary, J., Gottesman, R., Newhart, M., Chang, S., Ken, L., & Hillis, A. E. (2007). Utility of behavioral versus cognitive measures in differentiating between subtypes of frontotemporal lobar degeneration and Alzheimer's Disease. *Dementia and Geriatric Cognitive Disorders, 23*(3), 184-193. https://doi.org/10.1159/000098562

Heilbronner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., Millis, S. R., & Conference Participants (2009). American Academy of Clinical Neuropsychology consensus conference statement on the neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist, 23*, 1093-1129. https://doi.org/10.1080/13854040903155063

Hester, R. L., Kinsella, G. J., Ong, B., & McGregor, J. (2005). Demographic influences on baseline and derived scores from the Trail Making Test in healthy older Australian adults. *The Clinical Neuropsychologist, 19*, 45-54. https://doi.org/10.1080/13854040490524137

Hurtubise, J., Baher, T., Messa, I., Cutler, L., Shahein, A., Hastings, M., Carignan-Querqui, M., & Erdodi, L. (2020). Verbal fluency and digit span variables as performance validity indicators in experimentally induced malingering and real-world patients with TBI. *Applied Neuropsychology, Child*, does this have an issue number? 1-18. https://doi.org/10.1080/21622965.2020.1719409

Inman, T. H., & Berry, D. T. R. (2002). Cross-validation of indicators of malingering: A comparison of nine neuropsychological tests, four tests of malingering, and behavioral observations. *Archives of Clinical Neuropsychology, 17*, 1-23. https://doi.org/10.1016/S0887-6177(00)00073-1

Iverson, G. L., Lange, R. T., Green, P., & Franzen, M. D. (2002). Detecting exaggeration and malingering with the Trail Making Test. *The Clinical Neuropsychologist, 16*(3), 398-406. https://doi.org/10.1076/clin.16.3.398.13861

Jones, A. (2016). Cutoff scores for MMPI-2 and MMPI-2-RF Cognitive-Somatic validity scales for psychometrically defined malingering groups in a military sample. *Archives of Clinical Neuropsychology, 31*(7), 786-801. https://doi.org/10.1093/arclin/acw035

Jurick, S. M., Crocker, L. D., Merritt, V. C., Hoffman, S. N., Keller, A. V., Eglit, G. M. L., Thomas, K. R., Norman, S. B., Schiehser, D. M., Rodgers, C. S., Twamley, E. W., & Jak, A. J. (2020).

Psychological symptoms and rates of performance validity improve following trauma-focused treatment in veterans with PTSD and history of mild-to-moderate TBI. *Journal of International Neuropsychological Society, 26*(1), 108-118. doi: 10.1017/S1355617719000997.

Kanser, R. J., Rapport. L. J., Bashem, J. R., Billings, N. M., Hanks, R. A., Axelrod, B. N., & Miller, J. B. (2017). Strategies of successful and unsuccessful simulators coached to feign traumatic brain injury. *The Clinical Neuropsychologist, 31*(3), 644-653. https://doi.org/10.1080/13854046.2016.1278040

Kim, N., Boone, K. B., Victor, T., Lu, P., Keatinge, C., & Mitchell, C. (2010). Sensitivity and specificity of a Digit Symbol recognition trial in the identification of response bias. *Archives of Clinical Neuropsychology, 25*(5), 420-428. https://doi.org/10.1093/arclin/acq040

Kirby, A., Sugden, D., Beveridge, S., & Edwards, L. (2008). Developmental Co-ordination Disorder (DCD) in adolescents and adults in further and higher education. *Journal of Research in Special Educational Needs, 8*(3), 120-131. https://doi.org.10.1111/j.1471-3802.2008.00111.x

Kirkwood, M.W. (2015). *Validity testing in child and adolescent assessment: Evaluating exaggeration, feigning, and noncredible effort*. The Guilford Press. https://doi.org/10.1080/07317107.2017.1375723

Knowles, E. E. M., Mathias, S. R., McKay, D. R., Sprooten, E., Blangero, J., Almasy, L., & Glahn, D. C. (2015). Genome-wide analyses of working-memory ability: A review. *Current Behavioral Neuroscience Reports, 1*, 224-233. https://doi.org/10.1007/s40473-014-0028-8

Kura, K. (2013). Japanese north–south gradient in IQ predicts differences in stature, skin color, income, and homicide rate. *Intelligence, 41*(5), 512-516. https://doi.org/10.1016/j.intell.2013.07.001

Lamberty, G. J., Putnam, S. H., Chatel, D. M., Bieliauskas, L. A., & Adams, K. M. (1994). A preliminary report. *Cognitive and Behavioral Neurology, 7*(3), 230-234. https://psycnet.apa.org/record/1995-31624-001

Lange, R. T., Iverson, G. L., Zakrzewski, M. J., Ethel-King, P. E., & Franzen, M. D. (2005). Interpreting the Trail Making Test following traumatic brain injury: Comparison of traditional time scores and derived indices. *Journal of Clinical and Experimental Neuropsychology, 27*, 897-906. https://doi.org/10.1080/1380339049091290

Larrabee, G. J. (2008). Aggregation across multiple indicators improves the detection of malingering: Relationship to likelihood ratios. *The Clinical Neuropsychologist, 22*(4), 666-679. https://doi.org/10.1080/13854040701494987

Lees-Haley, P. R., & Fox, D. D. (1990). Neuropsychological false positives in litigation: Trail Making Test findings. *Perceptual and Motor Skills, 70*(3), 1379-1382. https://doi.org/10.2466/pms.1990.70.3c.1379

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment which edition?*. Oxford University Press.

Lippa, S. M. (2018). Performance validity testing in neuropsychology: A clinical guide, critical review, and update on a rapidly evolving literature. *The Clinical Neuropsychologist, 32*(3), 391-421. https://doi.org/10.1080/13854046.2017.1406146

Lichtenstein, J. D., Greenacre, M. K., Cutler, L., Abeare, K., Baker, S. D., Kent, K., J., Ali, S., & Erdodi, L. A. (2019). Geographic variation and instrumentation artifacts: In search of confounds in performance validity assessment in adults with mild TBI. *Psychological Injury and Law, 12*(2), 127-145. https://doi.org/10.1007/s12207-019-09354-w

MacAllister, W. S., Vasserman, M., & Armstrong, K. (2019). Are we documenting performance validity testing in pediatric neuropsychological assessment? A brief report. *Child Neuropsychology, 25*(8), 1035-1042. https://doi.org/10.1080/09297049.2019.1569606

Martin, P. K., Schroeder, R. W., & Odland, A. P. (2015). Neuropsychologists' validity testing beliefs and practices: A survey of North American professionals. *The Clinical Neuropsychologist, 29*(6), 741-746. https://doi.org/10.1080/13854046.2015.1087597

McDaniel, M. A. (2006). Estimating state IQ: Measurement challenges and preliminary correlates. *Intelligence, 34*(6), 607-619. https://doi.org/10.1016/j.intell.2006.08.007

Merten, T., Bossink, L., & Schmand, B. (2007). On the limits of effort testing: Symptom validity tests and severity of neurocognitive symptoms in nonlitigant patients. *Journal of Clinical and Experimental Neuropsychology, 29*(3), 308-318. https://doi.org/10.1080/13803390600693607

Merten, T., & Merckelbach, H. (2013). Symptom validity in somatoform and dissociative disorders: A critical review. *Psychological Injury and Law, 6*(2), 122-137. https://doi.org/10.1007/s12207-013-9155-x

Mossman, D., & Hart, K. J. (1996). Presenting evidence of malingering to courts: Insights from decision theory. *Behavioral Sciences & the Law, 14*(3), 271–291. https://doi.org/10.1002/(SICI)1099-0798(199622)14:3<271::AID-BSL240>3.0.CO;2-B

Neill, E., & Rossell, S. L. (2013). Executive functioning in schizophrenia: The result of impairments in lower order cognitive skills. *Schizophrenia Research, 150*(1), 76-80. https://doi.org/10.1016/j.schres.2013.07.034

Niesten, I. J. M., Merckelbach, H., Dandachi-FitzGerald, B., & Jelicic, M. (2020). The iatrogenic power of labeling medically unexplained symptoms: A critical review and meta-analysis of "Diagnosis Threat" in mild head injury. *Psychology of Consciousness: Theory, Research, and Practice*. https://doi.org/10.1037/cns0000224

O'Bryant, S. E., Hilsabeck, R. C., Fisher, J. D., & McCaffrey, R. J. (2003). Utility of the Trail Making Test in the assessment of malingering in a sample of mild traumatic brain injury litigants. *The Clinical Neuropsychologist, 17*(1), 69-74. https://doi.org/10.1076/clin.17.1.69.15624

Pearson (2009). *Advanced Clinical Solutions for the WAIS-IV and WMS-IV – Technical Manual*. Pearson.

Powell, M. R., Locke, D. E., Smigielski, J. S., & McCrea, M. (2011). Estimating the diagnostic value of the Trail Making Test for suboptimal effort in acquired brain injury rehabilitation patients. *The Clinical Neuropsychologist, 25*(1), 108-118. https://doi.org/10.1080/13854046.2010.532912

Poynter, K., Boone, K. B., Ermshar, A., Miora, D., Cottingham, M., Victor, T. L., Ziegler, E., Zeller, M. A., & Wright, M. (2019). Wait, there's a baby in this bath water! Update on quantitative and

qualitative cut-offs for Rey 15-Item Recall and Recognition. *Archives of Clinical Neuropsychology, 34*(8), 1367-1380. https://doi.org/10.1093/arclin/acy087

Rai, J., An, K. Y., Charles, J., Ali, S., & Erdodi, L. A. (2019). Introducing a forced choice recognition trial to the Rey Complex Figure Test. *Psychology and Neuroscience. 12*(4), 451-472. https://doi.org/10.1037/pne0000175

Rai, J., & Erdodi, L. (2019). The impact of criterion measures on the classification accuracy of TOMM-1. *Applied Neuropsychology: Adult*. Advance online publication. https://doi.org/10.1080/23279095.2019.1613994

Reese, C. S., Suhr, J. A., & Riddle, T. L. (2012). Exploration of malingering indices in the Wechsler Adult Intelligence Scale – Fourth Edition Digit Span subtest. *Archives of Clinical Neuropsychology, 27,* 176-181. https://doi.org/10.1093/arclin/acr117

Reitan, R. M. (1955). The relation of the Trail Making Test to organic brain damage. *Journal of Consulting Psychology, 19*, 393-394. https://doi.org/10.1037/h0044509

Reitan, R. M. (1958). The validity of the Trail Making Test as an indicator of organic brain damage. *Perceptual and Motor Skills, 8*, 271-276. https://doi.org/10.2466/PMS.8.7.271-276

Rey, A. (1941). L'examen psychologique dans les cas d'encephalopathie traumatique [Psychological examination in cases of traumatic encephalopathy]. *Archives de Psychologie, 28,* 286-340.

Rickards, T. A., Cranston, C. C., Touradji, P., & Bechtold, K. T. (2018). Embedded performance validity testing in neuropsychological assessment: Potential clinical tools. *Applied Neuropsychology: Adult, 25*(3), 219-230. https://doi.org/10.1080/23279095.2017.1278602

Rogers, R., & Bender, S. D. (2018). *Clinical assessment of malingering and deception* (4th ed.). The Guilford Press.

Roth, R. M., Erdodi, L. A., McCulloch, L. J., & Isquith, P. K. (2015). Much ado about norming the Behavior Rating Inventory of Executive Function. *Child Neuropsychology, 21*(2), 225-233. https://doi.org/10.1080/09297049.2014.897318

Ruffolo, L. F., Guilmette, T. J., & Willis, W. G. (2000). Comparison of time and error rates on the Trail Making Test among patients with head injuries, experimental malingerers, patients with suspect effort on testing, and normal controls. *The Clinical Neuropsychologist, 14*(2), 223-230. https://doi.org/10.1076/1385-4046(200005)14:2;1-Z;FT223

Saban, M. T., & Kirby, A. (2018). Adulthood in Developmental Coordination Disorder (DCD): A review of current literature based on ICF perspective. *Current Developmental Disorders Report. 5*, 9-17. https://doi.org/10.1007/s40474-018-0126-5

Salthouse, T. A., Toth, J., Daniels, K., Parks, C., Pak, R., Wolbrette, M., & Hocking, K. J. (2000). Effects of aging on efficiency of task switching in a variant of the Trail Making Test. *Neuropsychology, 14*(1), 102-111. https://doi.org/10.1037/0894-4105.14.1.102

Savla, G., Twamley, E. W., Thompson, W. K., Delis, D. C., Jeste, D. V., & Palmer, B. W. (2011). Evaluation of specific executive functioning skills and the processes underlying executive control

in schizophrenia. *Journal of the International Neuropsychological Society, 17*, 14-23. https://doi.org/10.1017/S1355617710001177

Schroeder, R. W., Martin, P. K., Heindrichs, R. J., & Baade, L. E. (2019). Research methods in performance validity testing studies: Criterion grouping approach impacts study outcomes. *The Clinical Neuropsychologist, 33*(3), 466-477. https://doi.org/10.1080/13854046.2018.1484517

Schroeder, R. W., Twumasi-Ankrah, P., Baade, L. E., & Marshall, P. S. (2012). Reliable Digit Span: A systematic review and cross-validation study. *Assessment, 19*(1), 21-30. https://doi.org/10.1177/1073191111428764

Schutte, C., Axelrod, B. N., & Montoya, E. (2015). Making sure neuropsychological data are meaningful: Use of performance validity testing in medicolegal and clinical contexts. *Psychological Injury and Law, 8*(2), 100-105. https:// https://doi.org/ 10.1007/s12207-015-9225-3

Shura, R. D., Miskey, H. M., Rowland, J. A., Yoash-Gatz, R. E., Denning, J. H. (2016). Embedded performance validity measures with postdeployment veterans: Cross-validation and efficiency with multiple measures. *Applied Neuropsychology: Adult, 23*, 94-104. https://doi.org/10.1080/23279095.2015.1014556

Stevens, A., Bahlo, S., Licha, C., Liske, B., & Vossler-Thies, E. (2016). Reaction time as an indicator of insufficient effort: Development and validation of an embedded performance validity parameter. *Psychiatry Research, 245*, 74-82. https://doi.org/10.1016/j.psychres.2016.08.022

Stuss, D. T., Stethem, L. L., & Poirier, C. A. (1987) Comparison of three tests of attention and rapid information processing across six age groups, *Clinical Neuropsychologist, 1*(2) 139-52, https://doi.org/10.1080/13854048708520046

Sugarman, M. A., & Axelrod, B. N. (2014). Utility of the Montreal Cognitive Assessment and Mini-Mental State Examination in predicting general intellectual abilities. *Cognitive & Behavioral Neurology, 27*(3), 148-154. https://doi.org/10.1097/WNN.0000000000000035

Sugarman, M. A., & Axelrod, B. N. (2015). Embedded measures of performance validity using verbal fluency tests in a clinical sample. *Applied Neuropsychology: Adult, 22*(2), 141-146. https://doi.org/10.1080/23279095.2013.873439

Tan, J. E., Slick, D. J., Strauss, E., & Hultsch, D. F. (2002). How'd they do it? Malingering strategies on symptom validity tests. *The Clinical Neuropsychologist, 16*(4), 495–5050. https://doi.org/10.1076/clin.16.4.495.13909

Tarachow, S. (1947). The syndrome of inhibition. *Psychiatric Quarterly, 21(2)*, 233-252. https://doi.org/10.1007/BF01641756

Tracy, D.K. (2014). Evaluating malingering in cognitive and memory examinations: A guide for clinicians. *Advances in Psychiatric Treatment, 20*(6), 405-412. https://doi.org/10.1192/apt.bp.114.012906

Trueblood, W. (1994). Qualitative and quantitative characteristics of malingered and other invalid WAIS-R and clinical memory data. *Journal of Clinical and Experimental Neuropsychology, 14*(4), 697-607. https://doi.org/10.1080/01688639408402671

Victor, T. L., & Abeles, N. (2004). Coaching clients to take psychological and neuropsychological tests: A clash of ethical obligations. *Professional Psychology: Research and Practice, 35*(4), 373-379. https://doi.org/10.1037/0735-7028.35.4.373

Viglione, D. J., Giromini, L., & Landis, P. (2016). The development of the Inventory of Problems–29: A brief self-administered measure for discriminating bona fide from feigned psychiatric and cognitive complaints. *Journal of Personality Assessment, 99*, 534–544. https://doi.org/10.1080/00223891.2016.1233882

Walczyk, J. J., Sewell, N., & DiBenedetto, M. B. (2018). A review of approaches to detecting malingering in forensic contexts and promising cognitive load-inducing lie detection techniques. *Front Psychiatry*. https://doi.org/10.3389/fpsyt.2018.00700

Webber, T. A., & Soble, J. R. (2018). Utility of various WAIS-IV Digit Span indices for identifying noncredible performance among performance validity among cognitively impaired and unimpaired examinees. *The Clinical Neuropsychologist, 32* (4), 657-670. https://doi.org/10.1080/13854046.2017.1415374

Whiteside D., M., Caraher, K., Hahn-Ketter, A., Gaasedelen, O., & Basso, M. R. (2018). Classification accuracy of individual and combined executive functioning embedded performance validity measures in mild traumatic brain injury. *Applied Neuropsychology: Adult, 26*(5), 472-481. https://doi.org/10.1080/23279095.2018.1443935

Whiteside, D. M., Kogan, J., Wardin, L., Philips, D., Franzwa, M. G., Rice, L., Basso, M., & Roper, B. (2015). Language-based embedded performance validity measures in traumatic brain injury. *Journal of Clinical and Experimental Neuropsychology, 37*(2), 220-227. https://doi.org/10.1080/13803395.2014.1002758

Zuccato, B. G., Tyson, B. T., & Erdodi, L. A. (2018). Early bird fails the PVT? The effects of timing artifacts on performance validity tests. *Psychological Assessment*, *30*(11), 1491-1498. https://doi.org/10.1037/pas0000596