



Attention Performance as an Embedded Validity Indicator in the Cognitive Assessment of Early Retirement Claimants

Anselm B. M. Fuermaier¹ · Brechje Dandachi-Fitzgerald² · Johann Lehrner³

Received: 19 May 2022 / Accepted: 10 November 2022
© The Author(s) 2022

Abstract

The assessment of performance validity is essential in any neuropsychological evaluation. However, relatively few measures exist that are based on attention performance embedded within routine cognitive tasks. The present study explores the potential value of a computerized attention test, the Cognitrone, as an embedded validity indicator in the neuropsychological assessment of early retirement claimants. Two hundred and sixty-five early retirement claimants were assessed with the Word Memory Test (WMT) and the Cognitrone. WMT scores were used as the independent criterion to determine performance validity. Speed and accuracy measures of the Cognitrone were analyzed in receiver operating characteristics (ROC) to classify group membership. The Cognitrone was sensitive in revealing attention deficits in early retirement claimants. Further, 54% ($n = 143$) of the individuals showed noncredible cognitive performance, whereas 46% ($n = 122$) showed credible cognitive performance. Individuals failing the performance validity assessment showed slower ($AUC = 79.1\%$) and more inaccurate ($AUC = 79.5\%$) attention performance than those passing the performance validity assessment. A compound score integrating speed and accuracy revealed incremental value as indicated by $AUC = 87.9\%$. Various cut scores are suggested, resulting in equal rates of 80% sensitivity and specificity (cut score = 1.297) or 69% sensitivity with 90% specificity (cut score = 0.734). The present study supports the sensitivity of the Cognitrone for the assessment of attention deficits in early retirement claimants and its potential value as an embedded validity indicator. Further research on different samples and with multidimensional criteria for determining invalid performance are required before clinical application can be suggested.

Keywords Neuropsychology · Attention · Performance validity · Embedded validity testing

Introduction

The assessment of symptom and performance validity is considered an essential part of any clinical neuropsychological evaluation. Both, position papers and consensus reports, stated that all clinicians should proactively include validity measures in their assessments, regardless of the assessment context (Bush et al., 2005; Sherman et al., 2020; Sweet et al.,

2021). Noncredible cognitive performance, as indicated by cognitive underperformance, is not restricted to narrowly defined assessment contexts but occurs regularly in various settings of neuropsychological evaluation, including the psychiatric, neurological, or forensic context (Czornik et al., 2021a, b; Dandachi-FitzGerald et al., 2016; Merten et al., 2021; Stevens et al., 2018). Even though the exact base rate of noncredible performance is difficult to estimate as it depends on a great variety of factors, it can be assumed that base rates of invalid test data are particularly high in the forensic setting in individuals with external incentives (Martin & Schroeder, 2020; Schroeder et al., 2021).

Research over the last decades considered a large number of validity measures, so-called performance validity tests (PVTs), that can be classified into stand-alone (freestanding) validity tests and validity indicators embedded within routine measures of neuropsychological functions (for an overview, see Boone, 2021). While stand-alone PVTs are considered the gold standard for producing high diagnostic

✉ Johann Lehrner
johann.lehrner@meduniwien.ac.at

¹ Department of Clinical and Developmental Neuropsychology, Faculty of Behavioral and Social Sciences, University of Groningen, Groningen, The Netherlands

² Department of Clinical Psychological Science, University of Maastricht, Maastricht, The Netherlands

³ Department of Neurology, Medical University of Vienna, Vienna, Austria

accuracy in distinguishing credible and noncredible cognitive performance, embedded PVTs have certain advantages over stand-alone PVTs (An et al., 2019; Erdodi et al., 2014, 2018). Empirical evidence and current practice standards recommend the use of multiple PVTs to assess validity and stress the need to sample validity continuously throughout the assessment and across cognitive domains (e.g., Rhoads et al., 2021; Soble, 2021; Sweet et al., 2021). Relying solely on multiple stand-alone PVTs would require considerable additional test-taking time, which is problematic and may not be feasible in many situations. Further, a large number of stand-alone PVTs are memory-based, which may limit the possibility to cover various cognitive domains in validity assessment. These problems can be solved by combining stand-alone tests with embedded indicators of performance validity (e.g., Soble, 2021). Embedded indicators have the advantage that they do not require additional administration time and automatically sample various cognitive domains, since they are derived from routinely administered cognitive tests (e.g., tests for information processing speed, attention, executive functioning, or memory). Embedded PVTs have the additional advantage over stand-alone PVTs that they measure a clinically relevant construct, and if the performance measure is deemed valid, the routine test provides information on cognitive functioning. Moreover, embedded PVTs are more robust to coaching as embedded validity testing is inseparable from ability testing (i.e., preparation of individuals prior to the assessment, Brennan et al., 2009; Suhr & Gunstad, 2007). As the main disadvantage, embedded validity indicators are more sensitive to cognitive impairment compared to stand-alone PVTs, which requires more extensive validation research for specific conditions in order to avoid the confusion of invalid with impaired test performance (e.g., see invalid-before-impaired paradox, Erdodi & Lichtenstein, 2017). However, the disadvantage of a specific embedded validity test might be overcome in clinical practice by aggregating multiple indicators of performance validity under the condition that each indicator has sufficiently large specificity for the diagnostic condition evaluated (e.g., Erdodi & Lichtenstein, 2017).

Neuropsychological tests for attention and concentration constitute promising candidates for embedded validity testing as attention disorders occur frequently in patients with acquired brain injury and psychiatric populations (Fish, 2017; Fuermaier et al., 2019; Spikman et al., 2001; Sturm, 2009). The assessment of attention disorders has great value for neurorehabilitation and psychological treatment, as attention is an integral cognitive function for many aspects of daily living, such as academic performance, following and contributing to conversations, mobility and car driving, occupational functioning, or managing finances. Further, attention is also assumed to be the building block for various other aspects of cognition (Butzbach et al., 2019; Duncan,

2013; Mohamed et al., 2021; Spikman et al., 1996) and, thus, directly or indirectly, affects a broad range of cognitive operations of our daily lives. Therefore, the forensic and clinical neuropsychological practice commonly applies computerized neuropsychological tests in order to determine the presence of attention deficits. Widespread popularity and common availability in neuropsychological practice is recognized for continuous performance tests (CPT; Rosvold et al., 1956). Because CPTs were designed to assess one or several aspects of sustained attention, they, obviously, share several features with short-term tests for attention and concentration, such as stimulus configuration, outcome variables, and cognitive demands.

On these grounds, attention tests, including CPTs, became growingly popular in research for their potential value to serve as embedded PVTs (Erdodi et al., 2014; Fuermaier et al., 2018; Harrison & Armstrong, 2020; Leppma et al., 2018; Stevens & Merten, 2010; Tucha et al., 2015; Willison & Tombaugh, 2006; Woods et al. 2015a, b). Although attention tests vary widely in task characteristics, stimulus material and configuration, and administration test duration, the accuracy and speed of responses on computerized attention tests are the most promising for predictive accuracy for noncredible cognitive performance. Reaction times (RT) and the variability of reaction times (often expressed by the standard deviation, RTSD) using the *Alertness* subtest of the TAP battery of attention performance (Zimmermann & Fimm, 1995) were dedicated particular interest by several research groups (Bodenburg, 2014; Czornik et al., 2021a, b; Fiene et al. 2015; Steck et al., 2000; Stevens et al., 2016). By employing different research designs on various populations, it was concluded that both reaction times and the variability of reaction times on simple attention tasks are promising variables to serve as embedded validity indicators. Next to embedded validity indicators, there is also a growing body of research on the development and evaluation of stand-alone performance validity tests that are seemingly related to attention, such as the Groningen Effort Test (GET; Fuermaier et al., 2016, 2017), the Dot Counting Test (DCT; basic counting skills; Boone et al., 2002), or the Victoria Symptom Validity Test (VSVT; learning and recalling digits; Slick et al., 1996, 1997).

Against the background of previous research in this field, it can be hypothesized that also other routine measures of attention may constitute useful embedded validity indicators if they are (1) frequently administered in the neuropsychological assessment context, (2) include speeded response parameters, and (3) complement the speed of responses with the accuracy of task performance. The Cognitrone (Schuhfried, 2021) is a computerized test for attention and concentration that fulfills these requirements and may have the potential to serve as an embedded PVT in various neuropsychological assessment settings. The Cognitrone requires a continuous perception and processing of visual

information, while responses on the congruence between geometric target and reference figures must be given in a timely and correct manner. The present study explored its utility as an embedded PVT in an archival dataset of 281 early retirement claimants referred for a neuropsychological evaluation to private office in Vienna, Austria. Whereas the sample of this study has been described in earlier research (Czornik et al., 2021a, b), the present study seeks to explore the utility of the Cognitrone to serve as an embedded PVT in distinguishing credible from noncredible cognitive test performance. The design of this study follows the principles

of a known-groups comparison, with the Word Memory Test (WMT; Green, 2003) as the criterion to determine noncredible cognitive test performance. Based on earlier research on attention paradigms and RT tasks, we expect the Cognitrone as a speeded attention task to serve as a useful embedded PVT. Classification accuracy is expected to be in a similar range than that for TAP subtasks as shown by others in earlier and related research. Further, we expect incremental validity in detecting noncredible cognitive test performance if both accuracy and speed of responses (i.e., in terms of speeded response frequency) are considered.

Fig. 1 Flow chart of participant selection for the present study

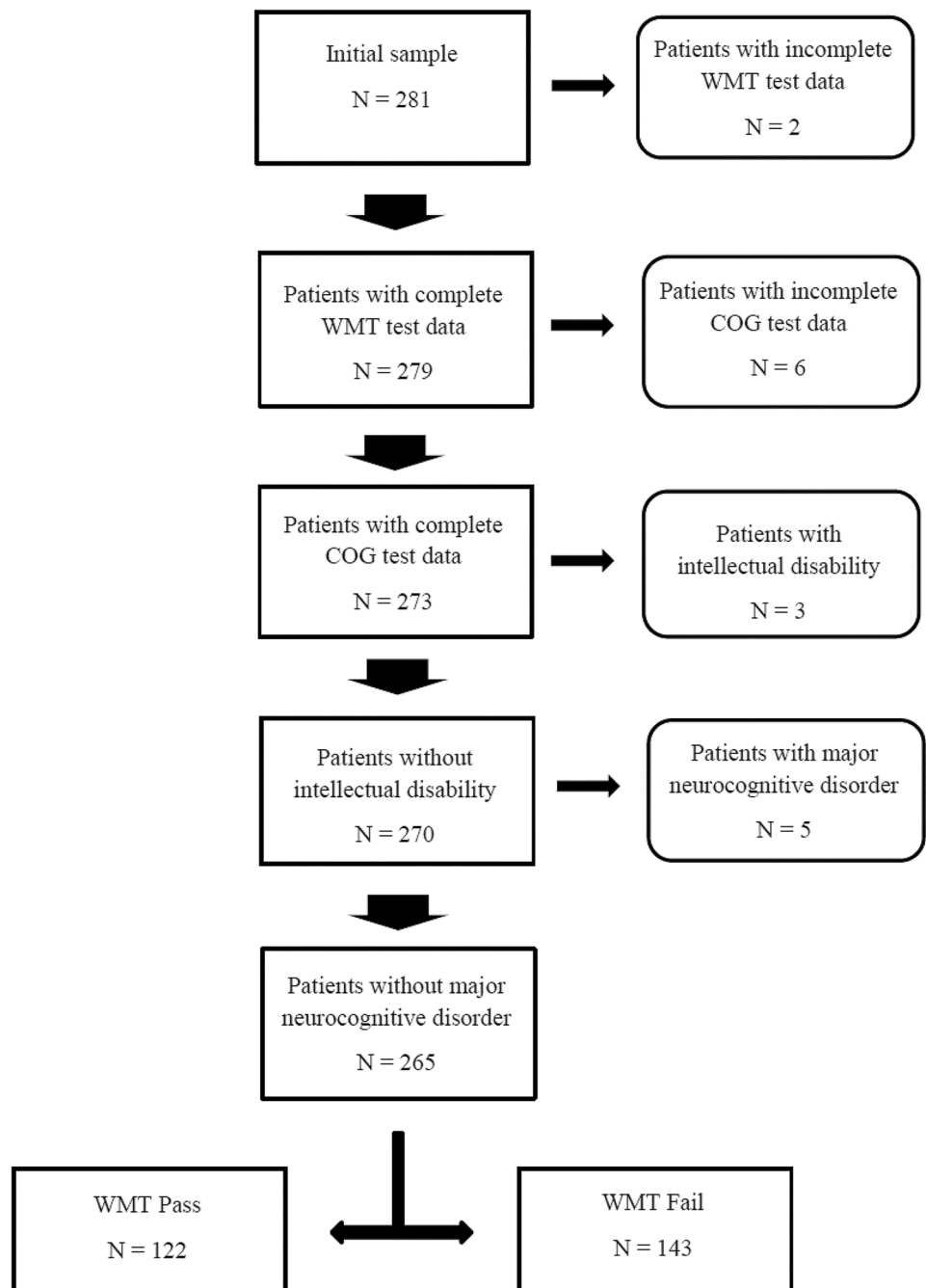


Table 1 Descriptive information and cognitive test performance of individuals passing and failing performance validity assessment (WMT)

	Total (<i>N</i> = 265)		WMT-pass (<i>n</i> = 122)		WMT-fail (<i>n</i> = 143)		<i>Z/X</i> ²	<i>p</i>	Cohen's <i>d</i>
	M±SD	PR ≤8	M±SD	PR ≤8	M±SD	PR ≤8			
Characteristics									
Age (in years)	47.8±9.7		46.6±10.8		48.8±8.6		1.312	0.190	0.23
Sex (f/m)	133/132		62/60		71/72		0.036	0.850	
Education (years)	11.1±3.7		11.7±4.1		10.6±3.2		2.291	0.022	0.30
Depression (BDI-II)	28.1±13.4		23.8±13.6		32.1±12.1		4.647	<.001	0.65
Word Memory Test									
WMT-IR	83.5±17.9		97.7±3.5		71.4±16.3		13.394	<.001	2.15
WMT-DR	82.0±18.5		97.5±3.6		68.7±15.4		13.931	<.001	2.49
WMT-CNS	79.9±17.1		95.8±4.9		66.4±11.1		14.054	<.001	3.34
WMT-MC	67.2±24.3		88.5±12.2		49.1±15.8		13.139	<.001	2.76
WMT-PA	60.3±26.6		82.3±16.5		41.1±17.2		12.617	<.001	2.44
WMT-FR	35.5±18.3		47.1±17.3		24.6±11.0		10.065	<.001	1.58
Cognitrone									
COG-Sum	266.6±124.5	49.8	331.6±111.6	26.2	211.1±107.1	69.9	8.155	<.001	1.10
COG-Incorrect	8.1±11.9	38.1	2.8±3.0	11.5	12.5±14.5	60.8	8.273	<.001	0.89

Education is indicated in years of school education including university/college

BDI-II Beck Depression Inventory–II, *WMT* Word Memory Test, *IR* immediate recall, *DR* delayed recall, *CNS* consistency, *MC* multiple choice, *PA* paired associates, *FR* free recall, *COG* Cognitrone, *Sum* Sum of Responses, *Incorrect* % incorrect responses, *PR* % of individuals having a percentile equal or smaller than 8 according to test norms as provided by the publisher

Methods

This retrospective study used an archival data sample stemming from the Department of Neurology of the Medical University of Vienna, Austria, collected in private office. The sample has been used in earlier research addressing related research questions and was described in detail by Czornik et al. (2021a, b). This sample contained data of independent neuropsychological evaluations of patients claiming early retirement due to significant cognitive impairment. Patients were referred by a general court or a pension insurance agency and were assessed between 2012 and 2019. The study protocol was approved by the Ethical Committee of the Medical University of Vienna and was conducted in accordance with the Declaration of Helsinki (registry number 2231/2020). All patients underwent a comprehensive neuropsychological evaluation, whereas the present study describes only materials and procedures for the current research question.

Participants and Procedure

The sample composition was presented in detail by Czornik et al. (2021a, b) and consisted mainly of patients with a confirmed ICD-10 (World Health Organization, 2011) chapter *F* (mental and behavioral disorders) diagnosis, a probable *F* diagnosis, or the claim of an *F*

diagnosis. However, the diagnostic status of individuals cannot be reported systematically for the present study, as diagnoses are not communicated to the examiners in the context of Austrian pension insurance referrals. The distribution of clinical disorders was estimated to be characterized by predominantly depressive disorders (about one-third), followed by adjustment disorders and ICD-10 neurasthenia (about one-fourth), anxiety disorder (about one-tenth), somatoform disorders (about 5% of cases), and a minority of diverse diagnoses (including substance abuse). Claims of softer psychopathology (Plomin, 1986) constituted the majority of cases in contrast to psychopathological syndromes of psychoses, delusions, confusional states, amnesic syndromes, or dementia.

Two hundred and eighty-one participants were initially considered for inclusion in the present study. Sixteen participants were excluded from data analysis because of incomplete WMT data (*n* = 2), incomplete COG data (*n* = 6), primary intellectual disability (*n* = 3), or a major neurocognitive disorder (such as dementia or amnesic syndrome; *n* = 5). Of the 265 participants entering data analysis, 143 failed the WMT (i.e., noncredible performance), and 122 passed the WMT (i.e., credible performance). Figure 1 depicts a flow chart of participant inclusion and exclusion. Descriptive information of the remaining sample is presented in Table 1, including age, gender, educational level, and depressive symptoms as assessed by the BDI-II.

Materials

BDI-II

Depressive symptoms were assessed with the Beck Depression Inventory–II (BDI-II; Beck et al., 1996; Hautzinger et al., 2006). The BDI-II contains 21 items, each rated on a 4-point scale (0–3), describing the experiences of depressive symptoms within the last 2 weeks. The BDI-II score is obtained by summing up individual item scores and, thus, ranges from 0 to 63. Any BDI-II score equal or higher than 19 indicates clinically relevant depressive symptoms in psychiatric patients (von Glischinski et al., 2019).

Word Memory Test

The Word Memory Test (WMT; Brockhaus & Merten, 2004; Green, 2003) is a computerized verbal memory test with built-in validity measures that intend to capture cognitive underperformance. The WMT involves memorizing a list of 20 semantically related simple word pairs, which is presented twice in the learning phase. Subsequently, three validity subtests and then three subtests for memory ability are administered. The present study considers the three validity subtests, i.e., immediate recognition, delayed recognition, and consistency, to classify the credibility of cognitive test performance. Per manual (Green, 2003), noncredible cognitive performance (WMT-fail) is concluded if the examinee scores equal to or below 82.5% on at least one of these three WMT subtests. Credible cognitive performance is concluded if the examinee scores above 82.5% on all three validity subtests (WMT-pass). The WMT was applied in numerous studies on diverse populations and was praised for its high sensitivity, with somewhat lower specificity rates (for an overview see Schroeder & Martin, 2021). A profile analysis (i.e., easy–hard difference) and the clinical presentation of individuals, as well as participation in real-world activities, were taken into consideration for the identification of major neurocognitive memory impairment, which was excluded from the main analysis.

Cognitrone (COG, Test Form S8)

The Cognitrone (abbreviation COG, test form S8, Schuhfried, 2021) is a computerized test for attention and concentration. The test requires the perception and processing of visual information and responding in a timely and correct manner. This test includes the defining characteristics of concentration tasks as it is composed of clearly visual stimuli, straightforward and easy-to-remember rules, and the differentiation between the speed and accuracy of responses as test outcome variables (Westhoff & Hagemeyer, 2005). A series of items (geometric figures) are presented in consecutive order in the COG test

administration. In each item, a target figure and four reference figures are presented on the screen, and the respondent is required to indicate as quickly as possible by pressing the correct of the two response buttons whether or not the target figure matches one of the reference figures. The next stimulus follows as soon as the answer is entered. Omitting items and going back to a preceding item is not possible. Working time on the COG is exactly seven minutes, in which as many items as possible should be evaluated. The COG's total administration time including the instruction phase is about 9 min. Test scores are the *Sum of Responses*, which is a quantitative indication of the number of all correct and incorrect responses given in the 7 min (speed indication) and the *Incorrect Responses* (% incorrect responses), which indicates the accuracy of the examinee's responses. The COG is normed based on a representative norm sample of 287 individuals aged 16 to 84 (Schuhfried, 2021).

Psychometric properties of the COG were examined by the test publisher and by independent follow-up research. The test yielded good internal consistency (0.87 and 0.98, for *Incorrect Responses* and *Sum of Responses*, respectively) and test–retest reliability (0.82; for a different test version, S11). Construct validity of the COG was shown by convergent and discriminant validity to near (e.g., choice reaction time tasks) and far-related (e.g., visual memory or visuospatial working memory) cognitive measures and by confirmatory factor analyses (Schuhfried, 2021). Further, the criterion validity of the COG was supported by studies implementing the COG in assessment batteries determining car-driving abilities in clinical and non-clinical samples (Ferreira, et al., 2013; Risser et al., 2008; Sommer et al., 2008, 2010).

Statistical Analysis

Characteristics of individuals and cognitive test performance were presented in descriptive statistics (means, standard deviations; frequencies for categorical data). For the Cognitrone test performance, individual test scores were additionally compared to test norms in order to identify those individuals scoring below average ($PR \leq 8$, Guilmette et al., 2020). Descriptive statistics were presented for the entire group ($N = 265$), as well as separately for those individuals passing ($n = 122$; credible performance) and failing ($n = 143$, noncredible performance) the WMT. Further, we considered a stricter criterion for passing the WMT by classifying test performance as credible only if all three WMT scores were above 90%. Group's statistics of this exploratory analysis are presented in the supplementary file. Next, all variables were statistically compared between both groups. Nonparametric group comparisons (i.e., Mann–Whitney U tests or chi-square tests for nominal data) were employed because of violations of assumptions for parametric testing (including normality and homogeneity of variances). The effect size

Cohen's d was calculated for each comparison to indicate the magnitude of the difference. Effect sizes were interpreted based on Cohen's classification as negligible ($d < 0.2$), small ($0.2 \leq d < 0.5$), medium ($0.5 \leq d < 0.8$), and large ($d \geq 0.8$; Cohen, 1988). The utility of the Cognitrone in classifying noncredible and credible test performance was explored in receiver operating characteristic (ROC) analyses, in which either of the COG variables was entered as test variable, and the WMT classification served as state variable. In case both models revealed significant classification accuracy, the incremental value when considering both COG variables was examined in binary logistic regression analysis with forward inclusion of predictor variables (COG variables) to distinguish credible from noncredible test performance (WMT classification). Because the groups showing credible and noncredible test performance also revealed small to moderate differences in years of education and depressive symptoms (see result section for a detailed presentation), additional hierarchical logistic regression models were carried out. In these models, either education or depressive symptoms were entered to the model as a second step after both COG variables have already been considered by the model in a first step. Bivariate correlation analyses further explored the effect of BDI-II scores and its association with attention performance and passing or failing the WMT.

In a subsequent step, in order to maximize the utility of both COG variables in distinguishing credible from noncredible test performance, we computed a *Cognitrone Compound Score* (CCS) by considering maximum value of each of the COG variables in classifying credible and noncredible performance. The CCS was derived by summing up the COG variables, each weighted with their unstandardized regression coefficients as obtained from the logistic regression model. The utility of the CCS in distinguishing credible from noncredible test performance was examined in another ROC analysis. Further, classification statistics sensitivity and specificity were presented for various cut scores of the CCS, together with their corresponding positive and negative predictive values (PPV, NPV) for a range of hypothetical base rate assumptions.

Results

Of the 265 individuals entering data analysis, 122 (46%) individuals passed, and 143 (54%) individuals failed the performance validity assessment based on the WMT. Descriptive

information and cognitive test performance of individuals passing and failing performance validity assessment are presented in Table 1. As expected, the group failing the WMT differed substantially from the group passing the WMT in all variables of this test, as indicated by Cohen's d effect sizes larger than two in five of the six WMT variables. Further, the group showing noncredible test performance indicated a slightly smaller number of years of education (small effect) and higher scores in depressive symptoms (medium effect) than the group showing credible test data; however, both groups did not differ significantly in age or gender. With regard to the Cognitrone, data analyses revealed in the group with noncredible test performance a significantly smaller number of *Sum of Responses* (indicating slower response speed) and a higher percentage of *Incorrect Responses* (indicating lower accuracy) compared to the group with credible test performance (large effect sizes). Group differences in attention performance are underlined by a substantial number of individuals failing the WMT who scored below average ($PR \leq 8$) in *Sum of Responses* (49.8%) and *Incorrect Responses* (38.1%) of the Cognitrone, compared to individuals passing the WMT of which only up to a quarter scored below average, i.e., 26.2% on *Sum of Responses* and 11.5% on *Incorrect Responses*. Additional exploratory analysis was performed by classifying test performance as credible (WMT-pass) only if all three WMT validity indices were above 90%. Descriptive statistics of this stricter WMT-pass criterion are presented in Table S1 of the supplementary file. One can see largely similar results than shown in the analysis with conventional WMT performance categorization, with the exception of a smaller proportion of individuals passing the WMT who scored in the below average range of the COG-Sum (20.0% vs. 26.2%) (Table 2).

Regarding the utility of the Cognitrone in distinguishing credible from noncredible test performance, ROC analyses indicated significant predictive power for either variable, i.e., *Sum of Responses*, $AUC = 0.791$, $SE = 0.028$, $p < 0.001$, 95% $CI = 0.736$ – 0.845 , and *Incorrect Responses*, $AUC = 0.795$, $SE = 0.027$, $p < 0.001$, 95% $CI = 0.741$ – 0.848 . Classification accuracies in terms of sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) are presented in Table 3. In order to capture possible incremental value when considering both variables in one model, we performed a logistic regression analysis with forward inclusion of predictor variables. This model considered first the speeded response frequency (i.e., *Sum of Responses*) and showed significant incremental predictive value when the

Table 2 Spearman correlation coefficients (r , p) between BDI-II and Cognitrone scores

	Total ($N = 265$)	WMT-pass ($n = 122$)	WMT-fail ($n = 143$)
COG-Sum	0.209, $< .001$	0.102, .269	0.074, .402
COG-Incorrect	0.296, $< .001$	0.132, .153	0.299, $< .001$

WMT Word Memory Test, COG Cognitrone, Sum Sum of Responses; Incorrect % incorrect responses

Table 3 Classification accuracy of *Sum of Responses*, *Incorrect Responses*, and the *Cognitrone Compound Score (CCS)* for various cut scores and hypothetical base rates of noncredible performance

Sensitivity		Specificity	BR 15%		20%		30%		40%		50%		60%	
			PPV	NPV	PPV	NPV	PPV	NPV	PPV	NPV	PPV	NPV	PPV	NPV
Sum of Responses ≤														
156	0.36	0.95	0.56	0.89	0.64	0.89	0.76	0.78	0.83	0.69	0.88	0.60	0.92	0.50
179	0.46	0.90	0.49	0.90	0.54	0.89	0.66	0.80	0.75	0.71	0.82	0.63	0.87	0.53
205	0.57	0.85	0.40	0.92	0.49	0.89	0.62	0.82	0.72	0.75	0.79	0.66	0.85	0.57
227	0.60	0.80	0.35	0.92	0.43	0.89	0.56	0.82	0.67	0.75	0.75	0.67	0.82	0.57
257	0.69	0.75	0.33	0.93	0.41	0.91	0.54	0.85	0.65	0.78	0.73	0.71	0.81	0.62
321	0.85	0.53	0.24	0.95	0.31	0.93	0.44	0.89	0.55	0.84	0.64	0.78	0.73	0.70
355	0.90	0.39	0.21	0.96	0.27	0.94	0.39	0.90	0.50	0.85	0.60	0.80	0.69	0.72
415	0.95	0.22	0.18	0.96	0.23	0.95	0.34	0.91	0.45	0.87	0.55	0.82	0.65	0.75
Incorrect Responses ≥														
7.335	0.54	0.95	0.66	0.92	0.73	0.89	0.82	0.83	0.88	0.76	0.92	0.67	0.94	0.58
5.500	0.61	0.90	0.52	0.93	0.60	0.90	0.72	0.84	0.80	0.78	0.86	0.70	0.90	0.61
4.745	0.63	0.85	0.43	0.93	0.51	0.90	0.64	0.84	0.74	0.78	0.81	0.70	0.86	0.61
4.390	0.65	0.80	0.37	0.93	0.45	0.90	0.58	0.84	0.68	0.77	0.77	0.70	0.83	0.61
3.735	0.69	0.75	0.33	0.93	0.41	0.91	0.54	0.85	0.65	0.78	0.73	0.71	0.81	0.62
1.700	0.85	0.41	0.21	0.94	0.27	0.92	0.38	0.87	0.49	0.80	0.59	0.73	0.68	0.65
1.375	0.90	0.32	0.19	0.95	0.25	0.93	0.36	0.88	0.47	0.83	0.57	0.76	0.67	0.68
1.005	0.95	0.24	0.18	0.97	0.24	0.95	0.35	0.92	0.46	0.88	0.56	0.83	0.65	0.76
CCS ≤														
0.494	0.64	0.95	0.69	0.94	0.76	0.91	0.85	0.86	0.90	0.80	0.93	0.73	0.95	0.64
0.734	0.69	0.90	0.55	0.94	0.63	0.92	0.75	0.87	0.82	0.81	0.87	0.74	0.91	0.66
1.165	0.77	0.85	0.48	0.95	0.56	0.94	0.69	0.90	0.77	0.85	0.84	0.79	0.89	0.71
1.297	0.80	0.80	0.41	0.96	0.50	0.94	0.63	0.90	0.73	0.86	0.80	0.80	0.86	0.73
1.435	0.82	0.75	0.37	0.96	0.45	0.94	0.58	0.91	0.69	0.86	0.77	0.81	0.83	0.74
1.639	0.86	0.70	0.34	0.97	0.42	0.95	0.55	0.92	0.66	0.88	0.74	0.83	0.81	0.77
1.948	0.90	0.62	0.30	0.97	0.37	0.96	0.50	0.94	0.61	0.90	0.70	0.86	0.78	0.81
2.546	0.95	0.44	0.23	0.98	0.30	0.97	0.42	0.95	0.53	0.93	0.63	0.90	0.72	0.85

In bold: cut score with specificity = 90%

CCS Cognitrone Compound Score, BR base rate, PPV positive predictive value, NPV negative predictive value

accuracy index (i.e., *Incorrect Responses*) of the Cognitrone was entered to the model, $X^2(2) = 135.057$, $p < 0.001$, with *Cox & Snell* R^2 increasing from 23.0 to 39.9% in the second step. Both variables contributed significantly to the model, with *Sum of Responses*, $B = 0.009$, $SE = 0.002$, $X^2(1) = 37.215$, $p < 0.001$, and *Incorrect Responses*, $B = -0.286$, $SE = 0.051$, $X^2(1) = 37.175$, $p < 0.001$. Correlation analyses between both attention scores further justified the relevance to include both in our prediction models. Whereas we observed a small effect on the total sample, $r = 0.275$, $p < 0.001$, only negligible effects were demonstrated for the WMT-pass group, $r = 0.092$, $p = 0.311$, and the WMT-fail group, $r = 0.166$, $p = 0.048$.

Further, hierarchical logistic regression models examined whether education (in years of schooling) or depressive symptoms (BDI-II) increased the predictive value of noncredible test performance. When entering both

Cognitrone variables first to the model, neither education, $X^2(1) = 2.278$, $p = 0.131$, nor depressive symptoms, $X^2(1) = 3.721$, $p = 0.054$, made a significant improvement to the model. Additional correlation analyses between depression and COG performance are presented in Table 2. Correlation statistics revealed negligible to small association between cognitive performance and depression, which turned significant in three of the six indices. A point-biserial correlation between BDI-II scores and passing or failing the WMT revealed a medium effect, $r = 0.307$, $p < 0.001$.

Next, in order to maximize the utility of the Cognitrone in distinguishing credible from noncredible test performance, we computed a *Cognitrone Compound Score (CCS)* by summing up the variable scores *Sum of Responses* and *Incorrect Responses* (% incorrect responses), each weighted with their unstandardized regression coefficients from the logistic regression model, i.e., $CCS = \text{Sum of Responses} \times 0.009$

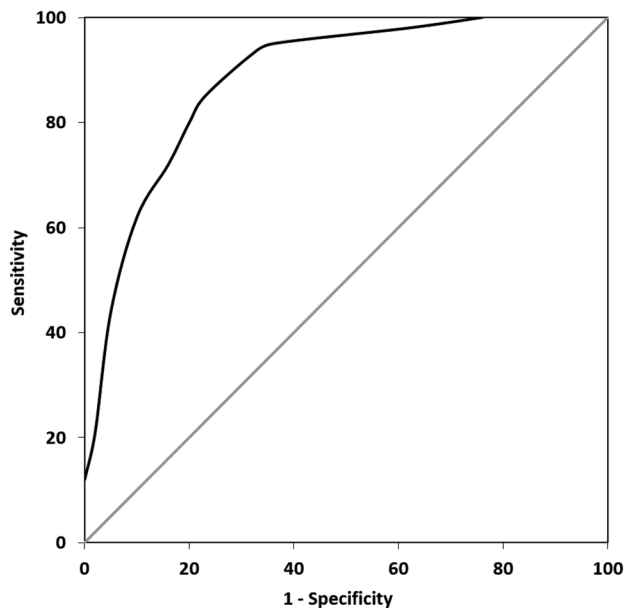


Fig. 2 Receiver operating characteristic (ROC) curve indicating diagnostic accuracy of the Cognitrone Compound Score (CCS) in distinguishing individuals with noncredible test performance ($n=143$) from individuals with credible test performance ($n=122$)

– *Incorrect Responses* $\times 0.286$. The classification accuracy of the CCS outperformed the models with single raw variables, $AUC=0.879$, $SE=0.021$, $p<0.001$, $95\% CI=0.839\text{--}0.920$. Figure 2 presents a graphical depiction of the ROC analysis and the corresponding AUC as derived from the CCS in distinguishing credible from noncredible test performance. Table 3 shows classification statistics sensitivity and specificity for various cut scores of the CCS. Table 3 additionally presents corresponding positive predictive values (PPV) and negative predictive values (NPV) for a range of hypothetical base rate assumptions of noncredible performance.

To illustrate the computation and interpretation of the CCS, a given individual with, for example, *Sum of Responses* = 357 and *Incorrect Responses* = 3.08 arrives at a $CCS = 357 \times 0.009 - 3.08 \times 0.286 = 2.33$, which falls above the suggested cut score of 0.734. Alternatively, another individual with *Sum of Responses* = 144 and *Incorrect Responses* = 4.17 yields a CCS of 0.103 (i.e., $144 \times 0.009 - 4.17 \times 0.286$), which falls below the suggested cut score of 0.734.

Discussion

The present study showed a 54% failure rate of performance validity testing based on the WMT as the criterion for noncredible cognitive performance. Czornik et al. (2021a, b) already reported this WMT failure rate on the same sample

of individuals, which falls within the broad range of estimated base rates of noncredible cognitive performance in forensic assessments reported in previous studies (Greiffenstein, et al., 1994; Larrabee et al., 2008; for a more comprehensive background, see Boone, 2021). Of note, it must be stressed that base rates are difficult to estimate as they depend on a great variety of factors, including the assessment setting, the presence of external gains, and the applied instruments for validity testing (Rogers et al., 1998).

Both the groups with credible and noncredible cognitive performance displayed higher rates of below average performance in attention assessment than it can be expected from a norm group, on the basis of a definition of “below average” with any score equal or lower the 8th percentile. Differences in below average rates emerged between groups, with only a slighter elevation of below average attention performance in the WMT-pass group (26% and 12% for speed and accuracy of responses, respectively; see also the supplementary file for a stricter criterion for passing the WMT) and much larger elevations in the WMT-fail group (50% and 38% for speed and accuracy, respectively). Below average speed and accuracy of attention performance in both groups demonstrates the sensitivity of this type of computerized attention assessment in the population of early retirement claimants.

The findings also give first indication that the attention assessment may be useful for embedded validity testing, because of the substantial higher rates of below average attention performance in the WMT-fail group compared to the WMT-pass group. Differences in attention performance between individuals showing credible and noncredible cognitive performance are also underlined by a comparison of raw scores indicating large effect sizes of Cohen’s $d=0.89$ (speed) and 1.10 (accuracy). Receiver operating characteristics (ROC) confirm the utility of the COG in differentiating between credible and noncredible cognitive performance, with similar values of AUCs for speed ($AUC=0.791$) and accuracy ($AUC=0.795$) of performance. From effect sizes of group differences, as well as classification accuracy as indicated by AUC, we learn that predictive power for each of the COG variables was similar to the strongest variable of the TAP *Alertness* task as reported by Czornik et al. (2021a, b), which further underscores the utility of attention tasks as embedded validity indicators. Both the TAP *Alertness* task and the COG provide a speeded measure (i.e., reaction time and speeded response frequency, respectively), whereas the TAP additionally indicates the variability of response speed and the COG reports the accuracy of performance. The accuracy of performance seems more important to indicate in the COG compared to the TAP *Alertness*, as the more complex task structure of the COG provokes a higher error rate than simple reaction time tasks such as the TAP *Alertness*. In the COG, individuals of the present study showed on average 267 responses of which 8% were incorrect, which equals to

about 21 incorrectly committed responses on average per person. In contrast, the entire TAP *Alertness* task with all its subtests contains 80 response-demanding cues, which are missed only in exceptions and incorrect responses occur rarely. This can be relevant as there is indication that processing speed is less suitable to serve as embedded validity indicator (i.e., low sensitivity) in populations with marked cognitive impairment (Ovsiew et al., 2020). In this case, a consideration of an accuracy measure (i.e., *COG-Incorrect*) as the primary embedded validity indicator may result in less false-positive classifications.

Considering both speed and accuracy of attention performance of the COG has proven to represent nonredundant information in the differentiation of credible from noncredible cognitive performance. Incremental value was observed when both variables are considered jointly, which increases explained variance from 23.0 to 39.9% in hierarchical regression models, and an AUC of 88% in the ROC analysis when considering the compound score (CCS). The classification accuracy of the CCS is superior to the classification accuracy of the variability of reaction times as reported by Czornik et al. (2021a, b) on the same sample and provides evidence for the advantage of jointly considering speed and accuracy of responses. A compound score integrating the speed and accuracy of attention performance may have another advantage that genuine cognitive dysfunction is more difficult to feign even for well-prepared individuals, as both variables and their trade-off need to be considered during task execution. Of note, it must be considered that the CCS is not readily available in the standard test output of the COG but must be computed manually. A compound score of attention variables for performance validity testing was also derived by Stevens et al. (2016) based on subtests of the TAP (*Alertness* and *Go/NoGo*) administered on individuals with and without brain lesion who were or were not involved in litigation. Based on factor analysis, the authors selected the variables representing the reaction time and variability of reaction time of the *Alertness* task while ignoring variables of the *Go/NoGo* task. The authors entered the selected variables into a logistic regression analysis to compute a compound score, which yielded good classification accuracy of noncredible performance in an independent validation sample, i.e., sensitivity = 79% and specificity = 87%. For the compound score of the COG in the present study (CCS), classification rates as presented in Table 3 suggest a cut score of 1.297 on the grounds that sensitivity and specificity are considered equally important (both reaching 80%). A cut score of 0.734 may be applied when specificity of at least 90% is desired (Boone, 2021), which results in an acceptable sensitivity rate of 69%.

Given the widespread availability and common use of CPTs in neuropsychological practice, their utility to serve as embedded PVTs has been studied in a considerable amount of research. Classification accuracy of the COG in the present study falls in a similar range and is thus broadly

consistent with the results of a variety of CPT research on different populations (i.e., traumatic brain injury (TBI) or adults with attention-deficit/hyperactivity disorder (ADHD)), research designs (i.e., known-groups comparisons or simulation designs), and applied variants of a CPT. For example, Harrison and Armstrong (2020) administered the Test of Variables of Attention (TOVA; Leark et al., 2007) in a known-groups comparison on adults at clinical evaluation of ADHD and reported most promising accuracy for the numbers of omissions (i.e., an accuracy measure) and the variability of reaction times, with effects falling in a same range than the ones of the present study. Similar conclusions were drawn from a simulation design in ADHD research (Fuermaier et al., 2018) when considering the speed and accuracy (omissions) of responses in the Perception and Attention Functions—Vigilance (WAFV; Schuhfried, 2013; Sturm, 2006). Variants of the Conners' CPT (2nd and 3rd edition, Conners, 2004, 2014) were applied in a number of studies on different populations, including individuals evaluated for ADHD (Scimeca et al., 2021; for a review see also Tucha et al., 2015) and patients with TBI (e.g., Erdodi et al., 2014). Using known-groups designs with independent criterion PVTs, effect sizes between groups and AUCs for the test variables omission errors, the reaction time, and the variability of reaction time roughly corresponded to the effects found in the present study. A slightly higher accuracy in detecting noncredible cognitive performance, especially on grounds of simple reaction time (i.e., 80% sensitivity and 87–93% specificity), was reported by Willison and Tombaugh (2006) in a simulation design on patients with TBI using the Computerized Test for Information Processing (CTIP, Tombaugh & Rees, 2000).

Finally, in the present study it was noted that people showing noncredible performance scored also higher on depressive symptoms than people with credible cognitive performance. Czornik et al. (2021a, b) explained this observation with the possibility that PVT failure may be linked to increased issues of psychopathology (i.e., depression) and/or that individuals failing PVT assessment may be prone to extreme symptom overreporting (see also Merten et al., 2020). Whereas the exact nature of this phenomenon is outside the scope of this paper, the effect of self-reported depressive symptoms on cognitive performance was only small in the present study, and BDI-II scores did not add significant predictive value to noncredible cognitive performance after COG performance has already been considered. A stronger and more prominent effect of depression may be observed on more heterogeneous samples with a larger spread in functioning. This can be speculated based on the observation of moderate depressive symptoms in the WMT-fail group, which may have restricted the range of scores and may have masked true existing relationships.

Limitations and Future Directions

As a major limitation of the present study, it must be noted that only one PVT (i.e., the WMT) was used as the criterion to determine noncredible cognitive performance. The use of a single criterion PVT has been shown to distort accuracy estimates of the predictor PVT, especially when high congruence between predictor and criterion is given (see domain-specificity hypothesis, Erdodi et al., 2018). In favor of the present study, the congruence between the cognitive domains of the predictor PVT (COG as an attention test) and criterion PVT (WMT as a verbal memory test) was low, and the WMT is presumably one of the best studied PVTs and de facto gold standard in clinic and research. Yet, the use of multiple PVTs (failure on ≥ 2 independent PVTs according to current standards, see Schroeder et al., 2019; Soble, 2021), or a multidimensional definition of noncredible symptom report or performance, became the standard in practice, and research would strengthen the rigor of this study (Rhoads et al., 2021; Sherman et al., 2020; Sweet et al., 2021).

Related to this first point of limitation, further validation of the COG as embedded validity indicator seems necessary, especially against embedded and stand-alone PVTs that are also seemingly related to attention and may thus be based on the same cognitive domain. The GET (attention-based PVT; Fuermaier et al., 2016, 2017), the DCT (basic counting skills, Boone et al., 2002), or the VSVT (learning and recalling digits; Slick et al., 1996) may be suitable stand-alone PVTs. Other well-established embedded validity measures that share cognitive characteristics with the COG would be the Reliable Digit Span (RDS; short-term and working memory; Greiffenstein et al., 1994) or variants of a CPT as discussed before.

Further, the key variable composed to serve as embedded validity indicator (CCS) was derived from a weighted sum of the speed (*COG-Speed*) and accuracy (*COG-Incorrect*) on the Cognitrone. Even though the variable weights were estimated from logistic regression on a sufficiently large sample, the reliability of the weights are still to be examined in replication research. In this context, the question also arises whether the computation of the compound score is context and population dependent, taking into consideration population-specific speed-accuracy tradeoffs to reach utmost classification accuracy. For this reason, the present study carried out ROC analysis also for the single variable scores representing speed and performance accuracy, which allows a determination of performance validity based on one performance type by ignoring the other. In addition to different weighting of speed and accuracy measures, also the cut scores determining noncredible performance may differ per assessment context and/or population. Depending on the severity of cognitive dysfunction, slower and more

inaccurate attention performance may be shown, which requires higher cut scores to maintain adequate specificity, though at the expense of sensitivity (see invalid-before-impaired paradox, Erdodi & Lichtenstein, 2017).

Finally, it could be considered as a limitation that the COG was designed with a fixed test duration of 7 min, during which participants respond to as many items as possible. This test design affects standardization of the COG as the number of evaluated items and the percentage of incorrect responses differed between participants.

Conclusion

This study underscores the sensitivity of computerized attention assessment, in particular the Cognitrone or tasks with similar characteristics, in the neuropsychological evaluation of early retirement claimants. Further, this study highlights the potential value of computerized attention tasks as embedded validity indicator. The broad application of attention tests in routine neuropsychological practice across assessment settings indicate favorable conditions for the implementation of embedded validity indicators. Whereas embedded validity indicators are generally more difficult to recognize because they are inseparable from routine function assessment, the computation of a compound score as suggested in this study may further increase its robustness to coaching. However, further research is needed on different samples, and using multidimensional classification criteria of invalid performance (Sherman et al., 2020) before clinical application can be recommended.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s12207-022-09468-8>.

Funding Open access funding provided by Medical University of Vienna.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- An, K. Y., Charles, J., Ali, S., Enache, A., Dhuga, J., & Erdodi, L. A. (2019). Reexamining 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>
- Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck Depression Inventory-Second Edition*. Psychological Corporation. https://doi.org/10.1007/978-1-4419-1005-9_441
- Bodenburg, S. (2014). Die Standardabweichung der Reaktionszeiten als ein Maß für negative Antwortverzerrungen [Standard deviation of simple reaction time as an indication of negative response bias]. *Zeitschrift Für Neuropsychologie*, 25(2), 89–97. <https://doi.org/10.1024/1016-264X/a000123>
- Boone, K. B. (2021). Assessment of feigned cognitive impairment: A neuropsychological perspective. Guilford Press. <https://doi.org/10.1080/23279095.2022.2064753>
- Boone, K. B., Lu, P., & Herzberg, D. (2002). *The Dot Counting Test*. Los Angeles, CA: Western Psychological Service.
- Brennan, A. M., Meyer, S., David, E., Pella, R., Hill, B. D., & Gouvier, W. D. (2009). The vulnerability to coaching across measures of effort. *The Clinical Neuropsychologist*, 23(2), 314–328. <https://doi.org/10.1080/13854040802054151>
- Brockhaus, R., & Merten, T. (2004). Neuropsychologische Diagnostik suboptimalen Leistungsverhaltens mit dem Word Memory Test [Neuropsychological assessment of suboptimal performance: The Word Memory Test]. *Der Nervenarzt*, 75(9), 882–887. <https://doi.org/10.1007/s00115-003-1592-4>
- Bush, S., Ruff, R., Troster, A., Barth, J., Koffler, S., Pliskin, N., Reynolds, C. R., & Silver, C. (2005). Symptom validity assessment: Practice issues and medical necessity - NAN policy & planning committee. *Archives of Clinical Neuropsychology*, 20(4), 419–426. <https://doi.org/10.1016/j.acn.2005.02.002>
- Butzbach, M., Fuermaier, A. B. M., Aschenbrenner, S., Weisbrod, S., Tucha, L., & Tucha, O. (2019). Basic processes as foundations of cognitive impairment in adult ADHD. *Journal of Neural Transmission*, 126(10), 1347–1362. <https://doi.org/10.1007/s00702-019-02049-1>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ: Lawrence Erlbaum Associates. <https://doi.org/10.4324/9780203771587>
- Conners, C. K. (2014). *Conners' Continuous Performance Test – 3rd Edition (CPT-3) manual*. Toronto, ON: Multi-Health Systems.
- Conners, K. C. (2004). *Conners' Continuous Performance Test (CPT II). Version 5 for Windows. Technical guide and software manual*. North Tonawanda, NY: Multi-Health Systems.
- Czornik, M., Merten, T., & Lehrner, J. (2021a). Symptom and performance validation in patients with subjective cognitive decline and mild cognitive impairment. *Applied: Neuropsychology Adult*, 28(3), 269–281. <https://doi.org/10.1080/23279095.2019.1628761>
- Czornik, M., Seidl, D., Tavakoli, S., Merten, T., & Lehrner, J. (2021b). Motor reaction times as an embedded measure of performance validity: A study with a sample of Austrian early retirement claimants. *Psychological Injury and Law*. <https://doi.org/10.1007/s12207-021-09431-z>
- Dandachi-FitzGerald, B., van Twillert, B., van de Sande, P., van Os, Y., & Ponds, R. W. H. M. (2016). Poor symptom and performance validity in regularly referred hospital outpatients: Link with standard clinical measures, and role of incentives. *Psychiatry Research*, 239, 47–53. <https://doi.org/10.1016/j.psychres.2016.02.061>
- Duncan, J. (2013). The structure of cognition: Attentional episodes in mind and brain. *Neuron*, 80(1), 35–50. <https://doi.org/10.1016/j.neuron.2013.09.015>
- 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., 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., Sagar, S., Seke, K., Zuccato, B. G., Schwartz, E. S., & Roth, R. M. (2018). The Stroop test as a measure of performance validity in adults clinically referred for neuropsychological assessment. *Psychological Assessment*, 30(6), 755–766. <https://doi.org/10.1037/pas0000525>
- Ferreira, I., Simões, M. R., & Marôco, J. (2013). Cognitive and psychomotor tests as predictors of on-road driving ability in older primary care patients. *Transportation Research Part F: Traffic Psychology and Behaviour*, 21, 146–158. <https://doi.org/10.1016/j.trf.2013.09.007>
- Fiene, M., Bittner, V., Fischer, J., Schwiecker, K., Heinze, H. J., & Zaehle, T. (2015). Untersuchung der Simulationssensibilität des Alertness-Tests der Testbatterie zur Aufmerksamkeitsprüfung (TAP) [Detecting simulation of attention deficits using the Alertness test of the Test of Attentional Performance (TAP)]. *Zeitschrift Für Neuropsychologie*, 26(2), 73–86. <https://doi.org/10.1024/1016-264X/a000125>
- Fish, J. (2017). Rehabilitation of attention disorders. In: B. A. Wilson, J. Winegardner, C. M. van Heugten, & T. Ownsworth (Eds.), *Neuropsychological rehabilitation: The international handbook, 1st ed.* (pp 172 – 186). Abingdon, Oxon; New York, NY: Routledge. <https://doi.org/10.4324/9781315629537>
- Fuermaier, A. B. M., Fricke, J. A., de Vries, S. M., Tucha, L., & Tucha, O. (2019). Neuropsychological assessment of adults with ADHD: A Delphi consensus study. *Applied Neuropsychology: Adult*, 26(4), 340–354. <https://doi.org/10.1080/23279095.2018.1429441>
- Fuermaier, A. B. M., Tucha, L., Koerts, J., Aschenbrenner, S., & Tucha, O. (2017). *Vienna Test System (VTS): The Groningen Effort Test (GET) for the assessment of noncredible cognitive performance in the assessment of adult ADHD*. Vienna, Austria: Schuhfried.
- Fuermaier, A. B. M., Tucha, O., Koerts, J., Grabski, M., Lange, K. W., Weisbrod, M., Aschenbrenner, S., & Tucha, L. (2016). The development of an embedded figures test for the detection of feigned attention deficit hyperactivity disorder in adulthood. *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0164297>
- Fuermaier, A. B. M., Tucha, O., Koerts, J., Send, T., Weisbrod, M., Aschenbrenner, S., & Tucha, L. (2018). Is motor activity during cognitive assessment an indicator for feigned ADHD in adults? *Journal of Clinical and Experimental Neuropsychology*, 40(10), 971–986. <https://doi.org/10.1080/13803395.2018.1457139>
- Green, P. (2003). *Green's Word Memory Test. User's manual*. Green's Publishing.
- Greiffenstein, M. F., Baker, W. J., & Gola, T. (1994). Validation of malingered amnesia measures with a large clinical sample. *Psychological Assessment*, 6, 218–224. <https://doi.org/10.1037/1040-3590.6.3.218>
- Guilmette, T. J., Sweet, J. J., Hebben, N., Koltai, D., Mahone, E. M., Spiegler, B. J., Stucky, K., Westerveld, M., & Participants, C. (2020). American Academy of Clinical Neuropsychology consensus conference statement on uniform labeling of performance test scores. *The Clinical Neuropsychologist*, 34(3), 437–453. <https://doi.org/10.1080/13854046.2020.1722244>
- Harrison, A. G., & Armstrong, I. T. (2020). Differences in performance on the test of variables of attention between credible vs. noncredible individuals being screened for attention deficit hyperactivity disorder. *Applied Neuropsychology: Child*, 9(4), 314–322. <https://doi.org/10.1080/21622965.2020.1750115>
- Hautzinger, M., Keller, F., Kühner, C., & Bürger, C. (2006). *Das Beck Depressions-Inventar II (BDI-II) [The Beck Depression Inventory II (BDI-II)]*. Harcourt Test Services.
- Larrabee, G. J., Millis, S. R., & Meyers, J. E. (2008). Sensitivity to brain dysfunction of the Halstead-Reitan vs an ability-focused

- neuropsychological battery. *The Clinical Neuropsychologist*, 22(5), 813–825. <https://doi.org/10.1080/13854040701625846>
- Leark, R. A., Greenberg, L. M., Kindschi, C. L., Dupuy, T. R., & Hughes, S. J. (2007). *T.O.V.A. professional manual. Test of variables of attention continuous performance test*. Los Alamitos, CA: The T.O.V.A Company.
- Leppma, M., Long, D., Smith, M., & Lassiter, C. (2018). Detecting symptom exaggeration in college students seeking ADHD treatment: Performance validity assessment using the NV-MSVT and IVA-Plus. *Applied Neuropsychology: Adult.*, 25(3), 210–218. <https://doi.org/10.1080/23279095.2016.1277723>
- Martin, P. K., & Schroeder, R. W. (2020). Base rates of invalid test performance across clinical non-forensic contexts and settings. *Archives of Clinical Neuropsychology*, 35(6), 717–725. <https://doi.org/10.1093/arclin/acia017>
- Merten, T., Dandachi-FitzGerald, B., Hall, V., Bodner, T., Giromini, L., Lehrner, J., Gonzalez-Ordi, H., Santamaria, P., Schmand, B., & Di Stefano, G. (2021). *Symptom and performance validity assessment in European countries: An update*. Advance online publication. <https://doi.org/10.1007/s12207-021-09436-8>
- Merten, T., Kaminski, A., & Pfeiffer, W. (2020). Prevalence of over-reporting on symptom validity tests in a large sample of psychosomatic rehabilitation inpatients. *The Clinical Neuropsychologist*, 34(5), 1004–1024. <https://doi.org/10.1080/13854046.2019.1694073>
- Mohamed, S. M. H., Butzbach, M., Fuermaier, A. B. M., Weisbrod, M., Aschenbrenner, S., Tucha, L., & Tucha, O. (2021). Basic and complex cognitive functions in adult ADHD. *PLoS ONE*, 16(9), e0256228. <https://doi.org/10.1371/journal.pone.0256228>
- Ovsiew, G. P., Resch, Z. J., Nayar, K., Williams, C. P., & Soble, J. R. (2020). Not so fast! Limitations of processing speed and working memory indices as embedded performance validity tests in a mixed neuropsychiatric sample. *Journal of Clinical and Experimental Neuropsychology*, 42(5), 473–484. <https://doi.org/10.1080/13803395.2020.1758635>
- Plomin, R. (1986). *Development, genetics, and psychology*. Lawrence Erlbaum.
- Rhoads, T., Neale, A. C., Resch, Z. J., Cohen, C. D., Keezer, R. D., Cerny, B. M., Jennette, K. J., Ovsiew, G. P., & Soble, J. R. (2021). Psychometric implications of failure on one performance validity test: A cross-validation study to inform criterion group definition. *Journal of Clinical and Experimental Neuropsychology*, 43(5), 437–448. <https://doi.org/10.1080/13803395.2021.1945540>
- Risser, R., Chaloupka, C., Grundler, W., Sommer, M., Häusler, J., & Kaufmann, C. (2008). Using non-linear methods to investigate the criterion validity of traffic-psychological test batteries. *Accident Analysis and Prevention*, 40(1), 149–157. <https://doi.org/10.1016/j.aap.2006.12.018>
- Rogers, R., Salekin, R. T., Sewell, K. W., Goldstein, A. M., & Leonard, K. (1998). A comparison of forensic and nonforensic malingers: A prototypical analysis of explanatory models. *Law and Human Behavior*, 22(4), 353–367. <https://doi.org/10.1023/A:1025714808591>
- Rosvold, H. E., Mirsky, A. F., Sarason, I., Bransome, E. D., Jr., & Beck, L. H. (1956). A continuous performance test of brain damage. *Journal of Consulting Psychology*, 20(5), 343. <https://doi.org/10.1037/h0043220>
- Schroeder, R. W., & Martin, P. K. (2021). Forced-choice performance validity tests. In K. Boone (Ed.), *Assessment of feigned cognitive impairment: A neuropsychological perspective* (pp. 34–63). Guilford Press.
- Schroeder, R. W., Clark, H. A., & Martin, P. K. (2021). Base rates of invalidity when patients undergoing routine clinical evaluations have social security disability as an external incentive. *The Clinical Neuropsychologist*. <https://doi.org/10.1080/13854046.2021.1895322>
- Schroeder, R. W., Martin, P. K., Heinrichs, 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>
- Schuhfried, G. (2013). *Vienna Test System (VTS) 8 (Version 8.2.00)*. [Computer software] Mödling, Austria: Schuhfried.
- Schuhfried, G. (2021). *Cognitrone (COG): Test Manual*. Schuhfried GmbH, Mödling.
- Scimeca, L. M., Holbrook, L., Rhoads, T., Cerny, B. M., Jennette, K. J., Resch, Z. J., Obolsky, M. A., Ovsiew, G. P., & Soble, J. R. (2021). Examining Conners' Continuous Performance Test-3 (CPT-3) embedded performance validity indicators in an adult clinical sample referred for ADHD evaluation. *Developmental Neuropsychology*. <https://doi.org/10.1080/87565641.2021.1951270>
- Sherman, E. M. S., Slick, D. J., & Iverson, G. L. (2020). Multidimensional malingering criteria for neuropsychological assessment: A 20-year update of the malingered neuropsychological dysfunction criteria. *Archives of Clinical Neuropsychology*, 35(6), 735–764. <https://doi.org/10.1093/arclin/acia019>
- Slick, D. J., Hopp, G., Strauss, E., & Spellacy, F. J. (1996). Victoria Symptom Validity Test: Efficiency for detecting feigned memory impairment and relationship to neuropsychological tests and MMPI-2 validity scales. *Journal of Clinical and Experimental Neuropsychology*, 18, 911–922. <https://doi.org/10.1080/01688639608408313>
- Slick, D. J., Hopp, G., Strauss, E., & Thompson, G. B. (1997). *Manual for the Victoria Symptom Validity Test*. Odessa, FL: Psychological Assessment Resources.
- Soble, J. R. (2021). Future directions in performance validity assessment to optimize detection of invalid neuropsychological test performance: Special issue introduction. *Psychological Injury and Law*, 14(4), 227–231. <https://doi.org/10.1007/s12207-021-09425-x>
- Sommer, M., Heidinger, C., Arendasy, M., Schauer, S., Schmitz-Gielsdorf, J., & Häusler, J. (2010). Cognitive and personality determinants of post-injury driving fitness. *Archives of Clinical Neuropsychology*, 25(2), 99–117. <https://doi.org/10.1093/arclin/acp109>
- Sommer, M., Herle, M., Häusler, J., Risser, R., Schützhofer, B., & Chaloupka, C. (2008). Cognitive and personality determinants of fitness to drive. *Transportation Research Part f: Traffic Psychology and Behaviour*, 11(5), 362–375. <https://doi.org/10.1016/j.trf.2008.03.001>
- Spikman, J. M., Kiers, H. A., Deelman, B. G., & van Zomeren, A. H. (2001). Construct validity of concepts of attention in healthy controls and patients with CHI. *Brain and Cognition*, 47(3), 446–460. <https://doi.org/10.1006/brcg.2001.1320>
- Spikman, J. M., van Zomeren, A. H., & Deelman, B. G. (1996). Deficits of attention after closed-head injury: Slowness only? *Journal of Clinical and Experimental Neuropsychology*, 18(5), 755–767. <https://doi.org/10.1080/01688639608408298>
- Steck, P., Reuter, B., Meir-Korrell, S., & Schenle, P. (2000). Zur Simulierbarkeit von neuropsychologischen Defiziten bei Reaktionen und bei Intelligenztests [About the simulation of neuropsychological deficits in reaction and intelligence tests]. *Zeitschrift Für Neuropsychologie*, 11(3), 128–140. <https://doi.org/10.1024/1016-264X.11.3.128>
- Stevens, A., & Merten, T. (2010). Psychomotor retardation: Authentic or malingered? A comparative study of subjects with and without traumatic brain injury and experimental simulators. *German Journal of Psychiatry*, 13(1), 1–8.
- 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 param-

- eter. *Psychiatry Research*, 245, 74–82. <https://doi.org/10.1016/j.psychres.2016.08.022>
- Stevens, A., Schmidt, D., & Hautzinger, M. (2018). Major depression – A study on the validity of clinicians' diagnoses in medicolegal assessment. *The Journal of Forensic Psychiatry & Psychology*, 29(5), 794–809. <https://doi.org/10.1080/14789949.2018.1477974>
- Sturm, W. (2006). Perception and Attention Functions – Vigilance (WAFV). Vienna, Austria: Schuhfried.
- Sturm, W. (2009). Aufmerksamkeitsstörungen. In W. Sturm, M. Herrmann & T. F. Münte (Hrsg.), *Lehrbuch der Klinischen Neuropsychologie* (S. 421 – 443). Würzburg: Spektrum.
- Suhr, J. A., & Gunstad, J. (2007). Coaching and malingering: A review. In G. J. Larrabee (Ed.), *Assessment of malingered neuropsychological deficits* (pp. 287–311). Oxford University Press.
- Sweet, J. J., Heilbronnner, R. L., Morgan, J. E., Larrabee, G. J., Rohling, M. L., Boone, K. B., Kirkwood, M. W., Schroeder, R. W., Suhr, J. A., & Conference Participants. (2021). American Academy of Clinical Neuropsychology (AACN) 2021 consensus statement on validity assessment: Update of the 2009 AACN consensus conference statement on neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist*, 35(6), 1053–1106. <https://doi.org/10.1080/13854046.2021.1896036>
- Tombaugh, T. N., & Rees, L. (2000). *Manual for the Computerized Tests of Information Processing (CTIP)*. Ottawa Ontario: Carleton University (unpublished test).
- Tucha, L., Fuermaier, A. B. M., Koerts, J., Groen, Y., & Thome, J. (2015). Detection of feigned attention deficit hyperactivity disorder. *Journal of Neural Transmission*, 122(1), 123–134. <https://doi.org/10.1007/s00702-014-1274-3>
- von Glischinski, M., von Brachel, R., & Hirschfeld, G. (2019). How depressed is “depressed”? A systematic review and diagnostic meta-analysis of optimal cut points for the Beck Depression Inventory revised (BDI-II). *Quality of Life Research*, 28, 1111–1118. <https://doi.org/10.1007/s11136-018-2050-x>
- Westhoff, K., & Hagemester, C. (2005). *Konzentrationsdiagnostik*. Pabst Science Publishers.
- Willison, J., & Tombaugh, T. N. (2006). Detecting simulation of attention deficits using reaction time tests. *Archives of Clinical Neuropsychology*, 21(1), 41–52. <https://doi.org/10.1016/j.acn.2005.07.005>
- Woods, D. L., Wyma, J. M., Yund, E. W., & Herron, T. J. (2015a). The effects of repeated testing, simulated malingering, and traumatic brain injury on high-precision measures of simple visual reaction time. *Frontiers in Human Neuroscience*, 9, 540. <https://doi.org/10.3389/fnhum.2015.00540>
- Woods, D. L., Wyma, J. M., Yund, E. W., & Herron, T. J. (2015b). The effects of repeated testing, simulated malingering, and traumatic brain injury on visual choice reaction time. *Frontiers in Human Neuroscience*, 9, 595. <https://doi.org/10.3389/fnhum.2015.00595>
- World Health Organization. (2011). *International statistical classification of diseases and related health problems – ICD-10, Tenth revision, Volume 2*. World Health Organization.
- Zimmermann, P., & Fimm, B. (1995). *Testbatterie zur Aufmerksamkeit-sprüfung (TAP)* [Test for Attention Performance]. Psytest.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.