https://doi.org/10.1007/s12207-021-09431-z

Motor Reaction Times as an Embedded Measure of Performance Validity: a Study with a Sample of Austrian Early Retirement Claimants

Manuel Czornik[1] · Doris Seidl[1] · Sophie Tavakoli[1] · Thomas Merten[2] · Johann Lehrner[1]

Received: 7 July 2021 / Accepted: 27 October 2021

Abstract
Among embedded measures of performance validity, reaction time parameters appear to be less common. However, their potential may be underestimated. In the German-speaking countries, reaction time is often examined using the *Alertness* subtest of the Test of Attention Performance (TAP). Several previous studies have examined its suitability for validity assessment. The current study was conceived to examine a variety of reaction time parameters of the TAP *Alertness* subtest with a sample of 266 Austrian civil forensic patients. Classification results from the Word Memory Test (WMT) were used as an external indicator to distinguish between valid and invalid symptom presentations. Results demonstrated that the WMT fail group performed worse in reaction time as well as its intraindividual variation across trials when compared to the WMT pass group. Receiver operating characteristic analyses revealed areas under the curve of .775–.804. Logistic regression models indicated the parameter *intraindividual variation of motor reaction time with warning sound* as being the best predictor for invalid test performance. Suggested cut scores yielded a sensitivity of .62 and a specificity of .90, or .45 and .95, respectively, when the accepted false-positive rate was set lower. The results encourage the use of the *Alertness* subtest as an embedded measure of performance validity.

Keywords Performance validity tests · Embedded measures of validity · Reaction time · Forensic assessment · Symptom validity tests · Alertness,

Introduction

Symptom validity tests (SVTs) and performance validity tests (PVTs) play an increasingly important role in the context of neuropsychological research as well as in clinical and forensic practice. They were developed to check the validity of patients' test profiles and the credibility of their symptom reports. Higher prevalence estimates of non-valid psychological test data can be expected regularly in forensic contexts (e.g., Larrabee, 2003; Mittenberg et al., 2002; and many subsequent studies), yet the exact base rate of invalid test results is unknown and estimates depend upon a large variety of factors (cf. Merten & Merckelbach, 2020). Recent studies indicate base rates of invalid test performance to be in the range of 25 to 46% among patients with external incentives (Martin & Schroeder, 2020; Schroeder et al., 2021). In a recent study, Mestdagh et al. (2019) found that a rate of 54.1% patients with claimed solvent-induced chronic encephalopathy failed on at least one of two PVTs.

In forensic contexts, neuropsychological evaluations are required to proactively assess the validity of test profiles and subjective symptom claims. Recent position papers summarize the current developments in conceptualization, empirical research, and professional practice (Sherman et al., 2020, Sweet et al., 2021).

Next to free-standing PVTs developed to make determinations about the validity of patients' neuropsychological test profiles, embedded measures of performance validity are playing an ever-increasing role, with several newly proposed measures every year (e.g., An et al., 2019; Eglit et al., 2020; for some recent research). These are scores or indices secondarily derived from neuropsychological tests given to patients in order to assess cognitive functions and neuropsychological symptoms. One of the advantages of embedded PVTs is their apparently higher resistance to coaching practices (e.g., Brennan et al., 2009; Suhr & Gunstad, 2007). Greve et al. (2009) pointed out that embedded PVTs contribute to the improvement of the overall sensitivity of all used procedures without requiring additional testing time. Furthermore, they provide direct information regarding the validity of a given performance during a specific test. Another advantage is that they can retrospectively determine the validity for contexts in which no specific validity assessment was carried out. However, their intrinsic link to

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

[1] Department of Neurology, Medical University of Vienna, Vienna, Austria

[2] Department of Neurology, Vivantes Klinikum Im Friedrichshain, Berlin, Germany

Published online: 06 November 2021

cognitive performance often results in a higher risk of false-positive classifications in populations with significant genuine cognitive impairment; Erdodi and Lichtenstein (2017) termed this the *Invalid before Impaired paradox*.

The investigation of speed tests or reaction time tasks to serve as embedded PVTs was based on the observation that in analogue studies, healthy participants who were instructed to simulate sequelae of mild brain damage chose to respond significantly more slowly and with greater intraindividual variability than genuinely brain damaged patients (e.g., Steck et al., 2000). These authors examined the *Alertness* subtest of the Test of Attention Performance (TAP; Zimmermann & Fimm, 1995), an easy-to-use computerized motor reaction time task upon visual stimulation. They found that experimental malingerers presented with higher reaction times, but coaching participants with a more credible deception strategy was successful in diminishing the reaction time gap between simulators and full-effort participants. The same result was found for intraindividual variability: it was highest in naïve simulators, lower in coached simulators, and lowest in full-effort controls. It was concluded that a motor reaction task like TAP *Alertness* may yield a measure that reliably identifies invalid test performance and is relatively robust against coaching.

Based on previous research on elevated standard deviation in other speed tasks (e.g., Willison & Tombaugh, 2006, using the Computerized Tests of Information Processing by Tombaugh & Rees, 2000; Erdodi et al., 2014, using the Conners' Continuous Performance Test by Conners & Staff, 2004), Bodenburg (2014) selectively examined TAP *Alertness* intraindividual variability in 91 patients after traumatic brain injury who were referred for personal injury forensic evaluations. The mean age of the patients was 45.2 years. In his known-group design study, patients were classified according to their performance on the Word Memory Test (WMT; Green, 2003). Yet, in contrast to the *lege artis* classification rule (see below), performance was classified as invalid if the mean of the WMT three primary validity indicators was 82.5% or below ($n=23$). Mean scores above this cut score were classified as valid performance ($n=68$). As a result of a multivariate analysis of covariance, only the standard deviation (intraindividual variability) in TAP *Alertness* predicted group membership, whereas a number of other performance test, questionnaire, and validity data did not. Thus, standard deviations were proposed to serve as an embedded performance validity measure.

Using the same test (TAP *Alertness*), Fiene et al. (2015) analyzed data from 17 young healthy adults, 19 patients with multiple sclerosis, and 17 patients with mild cognitive impairment. The group of healthy adults was given the test on three different occasions, with three differing sets of instructions: either to (a) react as fast as possible (i.e., employing the full test effort condition), (b) try and simulate mild attention deficits (naïve malingering condition), or (c) try and simulate deficits in a sophisticated way and avoid detection (coached malingering condition). Naïve simulators' scores were highest, in both reaction times and standard deviation of reaction times, thereby excessively outperforming both clinical patient groups. Yet, after coaching, their test performance approached that of the patient groups.

Another study investigating the TAP *Alertness* subtest (and another TAP subtest, *Go/No-go*) as an embedded PVT was conducted by Stevens et al. (2016). In contrast to the two studies described previously, the approach was based on a logistic regression analysis which yielded acceptable results.

Remarkably slower reaction times on simple and choice RT tasks were also demonstrated by experimental simulators as compared to patients with brain injury in two studies by Woods et al., (2015a, 2015b). The simulators were clearly slower than patients with mild traumatic brain injury ($n=24$) and even slightly slower than patients with severe traumatic brain injury ($n=4$).

In PVT research, a number of other studies have investigated both reaction time parameters (e.g., Stevens & Merten, 2010; Willison & Tombaugh, 2006) and response times in neuropsychological tests (e.g., Patrick et al., 2021), free-standing validity tests (e.g., Jansen et al., 2020; Kanser et al., 2019; Lupu et al., 2018), and computerized questionnaires (e.g., Cerny et al., 2021). Several studies investigated performance validity using continuous performance tests, such as the Test of Variables of Attention (T.O.V.A., Leark et al., 2008; e.g., recently, Harrison & Armstrong, 2020; Pollock et al., 2021). In contrast to the TAP *Alertness* subtest, this test is much more time-consuming and has a component of monotony. The common denominator of most studies is the assumption that reaction and response times are under lower conscious control and, therefore, may be harder to manipulate, and such manipulations might be more difficult to coach (Lupu et al., 2018; Vagnini et al., 2008). It also relates to studies in other fields of deception detection where manipulations in response behavior led to elevated reaction times (Lukács et al., 2021; Suchotzki et al., 2017).

The current study was conceived to examine a variety of reaction time parameters of the TAP *Alertness* subtest in an attempt to replicate previous studies, in particular the Bodenburg (2014) and Fiene et al. (2015) studies, using a sample of Austrian civil forensic patients. During the last decade, there has been very little research on embedded measures of performance validity in Europe. On the basis of the two studies described above, the TAP may offer a window of opportunity in this area since it is already in widespread use by neuropsychologists in the German-speaking countries to test attention. While Bodenburg (2014) presented no cut scores on the basis of his sample, the cut scores proposed by Fiene et al. (2015) may have been too high, as they were based on the authors' research methodology with a low number of participants in a repeated measurement design and with insufficiently described clinical groups (no control of referral context and possible secondary gain expectations). Like Bodenburg, we resorted to a known-group design and also used the WMT as the external criterion for determining group membership. This choice was supported by the fact that the WMT is one of the best validated PVTs available. Moreover, it is in widespread use throughout the world, including in several European countries (e.g., Dandachi-FitzGerald et al., 2013; LaDuke et al., 2018), and a large amount of empirical research has been carried out with it. As dependent variables, we included both reaction time (RT) and its intraindividual variation across trials (SD of RT) in our analyses. With a sufficiently high number of patients in both groups, we also tried to identify possible cut scores for the recommended use of TAP *Alertness* as an embedded measure of performance validity.

Method

Archival data stemming from independent neuropsychological evaluations of patients claiming early retirement due to significant cognitive impairment were analyzed. 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.

Participants

All patients were referred for comprehensive neuropsychological assessment between December 2012 and December 2019; the referral source was either a general court or a pension insurance agency. Patients were tested using a complete neuropsychological test battery. For the current analyses, we selected the TAP *Alertness* subtest as well as Beck Depression Inventory–II scores and three validity tests, as indicated below.

In the context of Austrian pension insurance referrals, diagnoses are not systematically communicated to the examiners and, consequently, they were missing in many cases of the current study. As a general rule for referrals, the patients either presented 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. The formulation of a final medical diagnosis and the integration of all data on a case, including the neuropsychological work-ups, were usually completed by a psychiatrist at a later time, but with no regular feedback to the neuropsychologist. The mix of referral or background primary diagnoses was estimated to be roughly equivalent to that reported from a larger sample of psychosomatic in-patients referred by the German Pension Fund (Merten et al., 2020), with a predominance of depressive disorder (about one third of cases), followed by adjustment disorders and ICD-10 neurasthenia (about one fourth of cases respectively), anxiety disorder (about one tenth of cases), somatoform disorders (about 5% of cases), and a minority of diverse diagnoses (including substance abuse). In both samples, patients with claims of "soft psychopathology" (Plomin, 1986) constituted the overwhelming majority of cases. (These are syndromes usually called neuroses in the pre-DSM-III era, in contrast to hard-core psychopathological problems, such as psychoses, delusions, confusional states, amnestic syndromes, dementia).

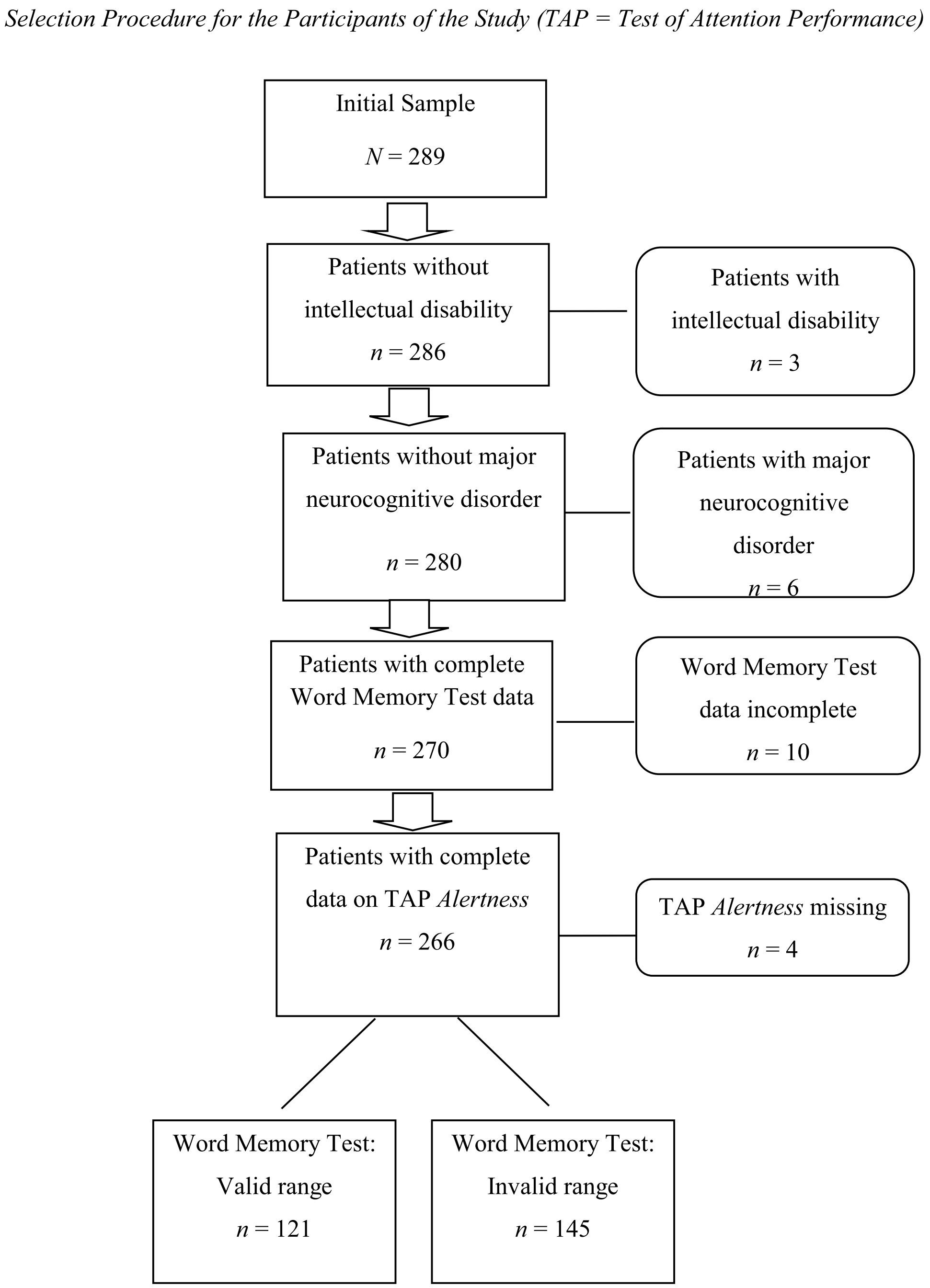
On this background, no detailed diagnostic information can be given for the sample composition. Only patients for which either the presence of primary intellectual disability ($n=3$) or a major neurocognitive disorder (such as dementia or amnestic syndrome; $n=6$) was communicated were excluded from the sample. This was done in order to eliminate, as much as possible, participants with a higher risk of scoring false positive on the WMT primary validity subtests. Also, patients with incomplete data on either WMT ($n=10$) or the TAP *Alertness* subtest ($n=4$) were excluded. Applying these criteria, of the 289 patients in the initial sample size, 266 patients were included in the final sample (Fig. 1).

Descriptive characteristics (demographic variables: age, sex, years of education) are presented in the upper part of Table 1. As outlined in the text below, the sample was divided on the basis of WMT performance into two groups with valid and invalid test profiles, respectively.

Instruments

Test Battery of Attention Performance (TAP; Zimmermann & Fimm, 1995), subtest Alertness. The TAP is a computerized test battery for the assessment of a large spectrum of components of attention that was originally developed to measure attentional deficits in patients with neurological deficits. It comprises 13 subtests focusing on attentional abilities such as alertness, divided attention, flexibility, and vigilance. In accordance with previous studies (Bodenburg, 2014; Fiene et al., 2015), we investigated the ability of the TAP *Alertness* subtest to be analyzed as an embedded validity measure (or embedded PVT). Alertness as a component of attention is defined as the state of wakefulness allowing the individual to respond quickly and adequately to concrete stimuli (Sturm & Willmes, 2001). It is a prerequisite for functional action and forms the basis of attentional performance. The TAP *Alertness* subtest, with a duration of approximately 10 min, measures patients' reaction times (RTs) in two different stimulus conditions over four blocks (two for each condition). The first condition is a simple RT task: a cross appears in the middle of the computer monitor at variable random time intervals. The participant is instructed to respond to the appearance of the cross as

Fig. 1 Selection procedure for the participants of the study (TAP = Test of Attention Performance)


Selection Procedure for the Participants of the Study (TAP = Test of Attention Performance)
Initial Sample
N = 289
Patients without intellectual disability
n = 286
Patients with intellectual disability
n = 3
Patients without major neurocognitive disorder
n = 280
Patients with major neurocognitive disorder
n = 6
Patients with complete Word Memory Test data
n = 270
Word Memory Test data incomplete
n = 10
Patients with complete data on TAP Alertness
n = 266
TAP Alertness missing
n = 4
Word Memory Test: Valid range
n = 121
Word Memory Test: Invalid range
n = 145

quickly as possible by pressing a key. This measures what is called the patient's "intrinsic alertness." The second test condition aims at what is called "phasic alertness." In this test condition, a computer-generated sound (beep) is first presented alerting the respondent (warning sound) to the target stimulus about to appear soon. Again, the participant is instructed to respond to the appearance of a cross by pressing a key as quickly as possible.

The four trials consist of 20 single measurements each. Trials 1 and 4 are conducted without a warning sound, whereas trials 2 and 3 are conducted with a warning sound. For each trial, the following results are computed: mean RT, median RT, standard deviation (SD) of RTs, number of correct responses (hits), omissions (misses), outliers (RT > [(mean RT + 2.35) x SD of RTs]), and anticipations (response given to the warning beep before the target cross was presented). For the phasic alertness reaction, the computer program automatically calculates a composite index (phasic alertness = (median RT without warning – median RT with warning)/median total RT). Normative data are available for the median and standard deviation of RTs and are separated for the two response conditions and for the composite index.

Table 1 Descriptive characteristics of the sample and results (raw scores) in relevant test parameters

	Total sample (*N* = 266)		Valid-range profiles: WMT negatives (passes) (*n* = 121)		Invalid-range profiles: WMT positives (fails) (*n* = 145)					
	M (*SD*)	Range	*M* (*SD*)	Range	*M* (*SD*)	Range	*t* (*df*)	*U*	*z*	Cohen's *d*
Age	47.8 (9.7)	18–64	46.5 (10.7)	18–64	48.9 (8.6)	20–63		9724.5	1.53	−0.25
Sex (% females)	133 (50%)		61 (50%)		72 (50%)					
Completed formal years of education	11.1 (3.6)	5–22	11.7 (3.9)	8–22	10.6 (3.2)	5–22		7353.5	− 2.27*	0.31
Beck Depression Inventory-II	28.2 (13.4)	0–58	23.8 (13.5)	0–58	32.0 (12.0)	0–58	− 5.13*** (235.6)			−0.65
Self-Report Symptom Inventory: Genuine Symptom Score	32.6 (10.7)	1–49	28.4 (11.4)	1–47	36.5 (8.2)	13–49		9851.5	5.83***	−0.83
Self-Report Symptom Inventory: Pseudosymptom Score	12.4 (10.8)	0–49	8.7 (8.3)	0–41	15.9 (11.6)	0–49		9595.5	5.34***	−0.70
Structured Inventory of Malingered Symptomatology: Total Score	22.0 (12.0)	3–64	16.9 (9.9)	3–57	26.7 (11.9)	5–64	− 6.68*** (227)			−0.89
Word Memory Test Immediate Recognition (%)	83.4 (17.9)	20.0–100	97.7 (3.5)	85.0–100	71.4 (16.2)	20.0–100		487.5	− 13.43***	2.16
Word Memory Test: Delayed Recognition (%)	82.0 (18.4)	22.5–100	97.5 (3.6)	87.5–100	69.0 (15.5)	22.5–100		241.5	− 13.81***	2.44
Word Memory Test: Consistency (%)	79.8 (17.1)	37.5–100	95.9 (4.9)	85.5–100	66.4 (11.0)	37.5–87.5		25.5	− 14.07***	3.36
Word Memory Test: Multiple Choice (%)	67.1 (24.3)	10.0–100	88.6 (12.1)	45.0–100	49.2 (15.9)	10.0–85.0		583.0	− 13.15***	2.76
Word Memory Test: Paired Associates (%)	60.26 (26.6)	0.0–100	82.5 (16.4)	35.0–100	41.3 (17.2)	0.0–85.0		854.0	− 12.61***	2.45
Word Memory Test: Free Recall (%)	35.4 (18.2)	0.0–95.0	47.1 (17.4)	12.5–95.0	24.7 (11.0)	0.0–57.5		2028.0	− 10.07***	1.57

[1] Years of education include years of formal schooling and college/university education, but not vocational training. No information was available on the latter

$^{*}p < .05$, $^{***}p < .001$

The test authors (Zimmermann & Fimm, 2002) stated that median reaction times of more than 400 ms (condition without warning sound) were indicative of a seriously reduced activation level. Such a reduction in basic RT was thought to affect the performance in other aspects of attention. One of the advantages of the TAP *Alertness* subtest is due to its minimal motor requirements and the very easy to understand test instruction; participants are not required to have prior experience with personal computers. Although there is a trend towards an increase in *Alertness* RTs usually beginning between 50 and 60 years of age, this slowing is far smaller if compared to more complex tasks. Further details on the TAP, its development, scoring, normative data, reliability and validity estimates, and research findings were felt to be beyond the scope of the current report; it can be retrieved in English language from Zimmermann and Fimm (2002).

Beck Depression Inventory–II (BDI-II; Beck et al., 1996). The German version (Hautzinger et al., 2006) is a self-administered 21-item questionnaire of depressive symptoms. On a four-point scale (0 to 3 points), patients are asked to describe the absence or presence of depressive symptoms within the past 2 weeks. The questionnaire is a paper-and-pencil test and takes approximately 5 to 10 min. The total score (varying from 0 to 63) is obtained by summing up the single scores. According to a German study by Köllner and Schauenburg (2012), the presence of clinically relevant depressive symptoms is indicated by total scores above 19.

Word Memory Test (WMT; Green, 2003; German version by Brockhaus & Merten, 2004). This is a computerized test for the simultaneous assessment of performance validity and memory functions. By means of a separate profile analysis, the WMT allows for a determination of whether or not the participant's test performance reliably reflects the true cognitive ability. The WMT comprises a series of subtests. For the purpose of the current study, we included only the three primary validity measures (Immediate Recognition, Delayed Recognition, and Consistency) since patients with major neurocognitive impairment for which profile analysis was developed were excluded from analysis. Invalid cognitive performance was diagnosed when a participant's score was 82.5% or below on at least one of the three WMT primary validity measures. In a subsequent step, the presence of a Genuine Memory Impaired Profile was checked. This incorporated both a profile analysis among WMT subtests (easy-hard difference) and information about the clinical presentation of patients and their participation in real-world activities.

Self-Report Symptom Inventory (SRSI; Merten et al., 2016). This is an SVT designed to detect possible symptom overreporting. The 107-item questionnaire consists of five scales of potentially genuine symptoms (e.g., cognitive, depressive, anxiety, and pain symptoms) and five scales of extreme, atypical, or bizarre pseudosymptoms (e.g., cognitive, motor, sensory, and pain pseudosymptoms) in a dichotomized response format (true or false). The professional test manual (Merten et al., 2019) recommends the use of either of two cut scores for the number of endorsed pseudosymptoms. At the standard cut score, overreporting was concluded in cases where ten or more pseudosymptoms were endorsed. At this point, the false-positive rate was less than 5%. Alternatively, a screening-level cut score (with a false-positive rate of less than 10%) was established at seven or more endorsed pseudosymptoms. The authors (Czornik et al., 2021) could recently demonstrate a satisfying resistance of SRSI pseudosymptom items against the presence of milder forms of cognitive impairment.

Structured Inventory of Malingered Symptomatology (SIMS; Smith & Burger, 1997). This SVT is a 75-item self-report questionnaire developed to diagnose malingered symptom presentations; however, in today's conceptual framework, it is better viewed as a measure of symptom overreporting or self-report symptom validity. The items mostly relate to extreme, atypical, or bizarre symptoms that apparently seem to fit into broad psychopathological domains. For every item, the examinee is asked to decide between true or false. The SIMS consists of five subscales of 15 items each (Neurological Impairment, Psychosis, Affective Disorders, Low Intelligence, and Amnestic Disorder). In the present study, the empirically established cut score for the German version (Cima et al., 2003) was used. With a total score of more than 16 endorsed pseudosymptoms, the test score is classified as positive and indicative of a noncredible symptom report. SIMS data were available for 239 of the 266 cases.

Results

On the basis of WMT classifications, the total sample was divided into a group of patients with credible test performance ($n=121$; WMT passes) and a group of patients with noncredible test performance ($n=145$; WMT fails). Hence, the WMT results yielded a base rate estimate of 54.5% for invalid test results (underachievement) in this special forensic population. Both subgroups were not different in terms of sex and age, but the group of valid WMT scorers presented with significantly more years of education.

About half of the participants (51.3%) with complete WMT data who failed on the primary validity measures of the test (Immediate Recognition, Delayed Recognition, and Consistency) presented with a difference between easy and hard subtests of at least 30. Yet, none of these patients presented clinically with major cognitive impairment. This led to the conclusion that none of them showed a Genuine Memory Impairment Profile.

Table 2 Chi-squared tests between performance and symptom validity test

		Word Memory Test						Structured Inventory of Malingered Symptomatology					
		Pass		Fail		Phi	χ^2	Pass		Fail		Phi	χ^2
		n	%	*n*	%			*n*	%	*n*	%		
Word Memory Test	Pass	–	–	–	–	–	–	58	53.2	51	46.8	.37	30.6***
	Fail							22	18.3	98	81.7		
Structured Inventory of Malingered Symptomatology	Pass	58	53.2	22	18.3	.37	30.6***	–	–	–	–	–	–
	Fail	51	46.8	98	81.7								
Self-Report Symptom Inventory: standard cut score	Pass	71	62.8	44	36.4	.27	16.4***	70	87.5	41	26.5	.57	72.3***
	Fail	42	37.2	77	63.7			10	12.5	114	73.5		
Self-Report Symptom Inventory: screening cut score	Pass	53	46.9	25	20.7	.28	18.1***	56	70.0	20	13.8	.57	72.8***
	Fail	60	53.1	96	79.3			24	30.0	125	86.2		

***$p < .001$

Table 1 also presents means, standard deviations, and ranges for a variety of test variables for the total sample as well as for the two subgroups. For all test variables, significant group differences were found. Patients who scored in the valid range of the WMT also endorsed fewer depressive symptoms (BDI-II) and fewer potentially genuine symptoms on the SRSI. At the same time, these patients also scored lower on the pseudosymptom report in two symptom validity tests (SIMS, SRSI).

Estimates of the rate of overreporting based on these SVTs amounted to 65.1% for the SIMS, 50.9% for the SRSI standard cut score, and 66.7% for the SRSI screening cut score. Chi-squared tests between the respective classification criteria of these tests and the WMT were all significant and are shown in Table 2.

In the motor reaction time task (TAP *Alertness*), patients with WMT results in the valid range scored significantly lower across all parameters, with the exception of the composite index *Phasic Alertness* (Table 3). Due to skewed distributions, all comparisons were done non-parametrically.

To examine the effectiveness of the TAP *Alertness* scores in predicting the diagnostic group and for the identification of the best fitting model, we ran a backward binary logistic regression analysis. Based on the *U* test analyses, we initially

Table 3 Results of a Variety of Test of Attention Performance (TAP), Subtest Alertness, Variables

	Total sample (*N* = 266)		Valid-range profiles: WMT negatives (passes) (*n* = 121)		Invalid-range profiles: WMT positives (fails) (*n* = 145)				
	M (*SD*)	Range	*M* (*SD*)	Range	*M* (*SD*)	Range	*U*	*z*	Cohen's *d*
Intraindividual variation (standard deviation) of motor reaction time: with warning (ms)	141.5 (126.1)	18–578	72.3 (59.6)	18–409	199.2 (137.6)	20–587	14,114	8.55***	−1.16
Intraindividual variation (standard deviation) of motor reaction time: without warning (ms)	161.2 (137.0)	15–651	87.7 (69.4)	15–438	222.5 (148.9)	18–651	13,969	8.32***	−1.13
Motor reaction time without warning: median scores (ms)	511.6 (324.8)	185–1584	341.3 (157.0)	185–1224	653.7 (359.0)	192–1584	13,798	8.04***	−1.10
Motor reaction time with warning: median scores (ms)	466.5 (306.6)	163–1793	316.4 (131.2)	163–884	591.8 (352.0)	181–1793	13,592	7.72***	−1.00
Composite index *Phasic Alertness*	0.08 (0.18)	−0.58–0.77	0.06 (0.14)	−0.56–0.47	0.10 (0.21)	−0.58–0.77	9848	1.85	−0.22

WMT Word Memory Test

***$p < .001$

added all TAP *Alertness* parameters that significantly ($p < 0.05$) predicted patients passing or failing the WMT from the strongest to the weakest in one block. As a result, all TAP *Alertness* parameters except *Phasic Alertness* were included in the initial model. In subsequent steps, the least significant parameter was removed at each step. This process continued until no non-significant variables remained and the accuracy of the final model, which only included significant predictors, was determined. With our data, the final model only contained the parameter *intraindividual variation (standard deviation) of motor reaction time: with warning sound* as the uniquely significant predictor to decide between WMT passing and failing (OR = 1.014, $p < 0.001$).

Subsequently, we included other potential predictor variables to determine if they could increase the predictive power of the model. Since completed school years and BDI-II results showed to be significantly different between the diagnostic groups ($p < 0.05$), a second and third models with the additional variables completed school years and BDI-II were computed. However, both parameters did not significantly improve the predictive value of the model, and the variable *intraindividual variation (standard deviation) of motor reaction time: with warning sound* remained as the only significant predictor.

In a subsequent step, we performed a receiver operating characteristic (ROC) analysis on the variable with the strongest predictive power of *intraindividual variation (standard deviation) of motor reaction time: with warning sound*. WMT classification results were used as the independent criterion for classification. The analysis revealed an area under the curve (AUC) of 0.804 (95% CI = 0.752–0.857, $p < 0.001$). While the Youden Index suggested a cut score of > 107.5 to yield the best overall classification results, specificity would be 0.85 at this point, and sensitivity 0.70. Yet, for PVTs, high specificity scores are essential. Table 4 comprises the more detailed results of the ROC analysis for the given test parameter. Table 5 contains estimates of positive predictive value (PPV) and negative predictive value (NPV) for a number of hypothetical base rate assumptions.

Table 4 Results of receiver operating characteristic (ROC) analysis for the Test of Attention Performance (TAP) subtest Alertness, intraindividual variation (standard deviation) of motor reaction time with warning sound (ms)

Possible cut scores (ms)	Sensitivity	Specificity
> 16	1.00	.000
…		
> 20	.993	.007
> 30	.959	.225
> 40	.910	.314
> 50	.855	.463
> 60	.807	.545
> 70	.772	.620
> 80	.745	.686
> 90	.724	.777
> 100	.703	.818
> 110	.683	.851
> 120	.648	.868
> 130	.628	.884
> 134	**.621**	**.900**
> 140	.621	.917
> 150	.586	.926
> 160	.559	.926
> 170	.510	.934
…		
> 193	**.448**	**.950**
…		
> 417	.090	1.00

The two cut scores that can be recommended for possible practical use were marked in bold

Discussion

The present study examined reaction times and their standard deviations of TAP *Alertness* parameters as possible indicators for invalid performance in neuropsychological assessments. The significance of symptom validation and the increasingly important role of embedded measures as opposed to freestanding PVTs are reflected by a multitude of recent published research. Although there are very few current German and European research activities in this area, the prominent use of the *Alertness* subtest in the German-speaking countries further denotes the significance of this study and the potential utility of the results for clinical and legal practice.

Based on patients' passing or failing of the WMT (Green, 2003), two groups were obtained which differed in all investigated PVTs and SVTs, BDI-II results, and completed formal years of education. Furthermore, the two diagnostic groups differed significantly in all TAP *Alertness* parameters besides the composite score of Phasic *Alertness*, independently of whether the reaction time was ascertained with or without an auditory warning signal. A backward binary logistic regression with all TAP *Alertness* parameters that differed significantly between the patients that passed or failed the WMT resulted in a single significant predictor, namely, *Alertness* with warning SD raw score. Adding the BDI-II and completed formal years of education, however, did not significantly improve the model's predictive power.

In the current study, we found the rate of positive validity test results to range between 51% (for the SRSI standard cut score) and 67% (for the SRSI screening cut score). WMT and SIMS results were within this range. Even when the legal setting and thus the readily identifiable secondary gains

Table 5 Positive and negative predictive values (PPV and NPV) for the Test of Attention Performance (TAP) subtest Alertness, intraindividual variation (standard deviation) of motor reaction time with warning sound (ms)

Hypothetical prevalence of invalid test performance	Cut score > 134 (specificity = .90)		Cut score > 193 (specificity = .95)	
	Positive predictive value	Negative predictive value	Positive predictive value	Negative predictive value
5%	0.25	0.98	0.32	0.97
10%	0.41	0.96	0.50	0.94
20%	0.61	0.90	0.69	0.87
30%	0.73	0.85	0.79	0.80
40%	0.81	0.78	0.86	0.72
50%	0.86	0.70	0.90	0.63
55%	0.88	0.66	0.92	0.58
60%	0.90	0.61	0.93	0.53

in the presented sample are considered, these results imply a high base rate of invalid symptom presentations. In the existing literature, a prevalence of probable malingering of 30 to 50% is suggested when similar tests and samples are used (Greiffenstein et al., 1994; Larrabee et al., 2008). However, some studies also found base rates far below or above this range. Against the background of the moderate concordance between the validity tests in the current study, our findings support the position that base rates vary in both dependence on the setting and possible secondary gains as well as on the applied validity test (also *cf.* Rogers et al., 1998).

A ROC analysis revealed a significant AUC of 0.804 for the parameter TAP *Alertness* with warning SD raw score. In this regard, analyses suggest a cut score of > 134.5 (PPV = 0.88; NPV = 0.66) or > 193.5 (PPV = 0.92; NPV = 0.58) rather than the score with the highest Youden Index due to the focus on a high specificity.

Fiene et al. (2015) previously identified less variable reaction times in the TAP *Alertness* subtest as a possible indicator for valid performance. The current study supports previous findings with a more representative sample due to a much greater sample size and further extends the knowledge to the identification of multiple embedded parameters. Our results are in line with the findings of Fiene et al. (2015) and indicate that not only the consistency but also the reaction times themselves may act as potential indicators, which are of important use for clinical practice as well as for legal issues, as the sensitivity to stimulation is no longer restricted to one single parameter. Rather, multiple parameters of the subtest may be used independently to judge the validity of performance profiles. This could also make it more difficult for coached patients to prepare for the assessment. The current study does not, however, make a statement as to how coaching would influence the results, but previous studies imply that it is not possible to achieve a standardized variability of the test values by deliberately simulating slower reaction times (Fiene et al., 2015; Steck et al., 2000).

Not only relating to reaction times as potential indicators for valid test performance, our results point in the same direction as the ones from Fiene et al. (2015) as well. The suggested cut scores are also in a remarkably similar range. In our sample, we suggest a cut score of > 134 or > 193 for the TAP *Alertness* with warning SD raw score corresponding to a specificity of 0.90 and 0.95, while Fiene et al. (2015) recommend a cut score of > 191 for the same task.

In the study by Bodenburg (2014), no cut scores were presented. Similar to our results, however, his study showed that the SDs of patients failing the WMT were remarkably larger than the ones of patients passing the WMT, independent of whether the RT was announced with or without a warning sound. Interestingly, in our study, the RT SDs of both tasks were around 20% larger for patients both passing and failing the WMT. This difference could be due to a greater heterogeneity in our sample of mental-health patients compared to the sample of traumatic brain injury patients in the Bodenburg (2014) study. The same applies to the Stevens et al. (2016) study in which no specific cut score for single TAP *Alertness* parameters were presented. Rather, a comprehensive model involving median RT as well as standard deviations of the TAP *Alertness* without warning sound was calculated. However, reaction times were much shorter and standard deviations were much smaller than in our study which, similar to the Bodenburg (2014) study, could be due to the inclusion of traumatic brain injury in the sample. It is indicated that depending on the type of the presented symptoms in clinical and forensic practice, the ideal cut score to distinguish between valid and invalid performance may vary, and that the applied cut scores for progressive cognitive impairments should be higher than the ones used in patients presenting with traumatic brain injury.

In contrast to other approaches to performance validity, intraindividual variability measures appear to play a minor role, yet their potential as embedded PVTs seems to be largely undetected. This will apply mostly to computerized neuropsychological tests for which RTs of single responses and intraindividual variability of RTs are either available or can easily be made available for analysis. In earlier stages of validity research, it was variability across different points of measurement that was judged to be a promising approach (e.g., Strauss et al., 2002).

To summarize, the current main results suggest further exploration of the *Alertness* subtest as a measure for performance validity testing. Consequently, in addition to neuropsychological research, regular use of TAP *Alertness* as an embedded PVT both in clinical work and in forensic referrals can be facilitated. Classification rates are moderately high, yet they can be judged as satisfying given that embedded PVTs usually perform below the potential of well-developed freestanding performance validity measures.

As shown in Table 2, the association between the WMT classification (passes vs. fails) and either of the two SVTs was only modest, whereas SIMS and SRSI results were much closer associated. Several previous studies with different samples demonstrated that the latter two self-report based validity tests apparently measure the same construct (e.g., Merten et al., 2016, 2020). In contrast, associations between performance-based validity measures and self-report scales are usually only moderate or, in some studies, even absent (e.g., Grossi et al., 2017; Ingram et al., 2020; Whiteside et al., 2020). Performance validity and symptom validity appear to constitute two related but distinct concepts (Ord et al., 2021; van Dyke et al., 2013).

Another result of the study deserves mentioning (Table 1). Participants with WMT results in the noncredible range scored much higher on the BDI-II ($M=32.0$; $SD=12.0$) than participants in the credible range ($M=23.8$; $SD=13.5$). This may signal a link of PVT failure to true mental problems, but it may also indicate a tendency of claimants who underperform on cognitive tests to exaggerate symptom report in mental-health scales (beyond SVTs). To further investigate this issue was outside the scope of the current study, but it can be identified as one of the core problems in current validity research. In two other recent studies, associations between extremely elevated BDI-II scores (above 40) and SVT failure was found (Kaminski et al., 2020; Merten et al., 2020), confirming a previous hypothesis by Groth-Marnat (1990). This author opined that Beck depression scores over 40 were significantly above the symptom claims even of severely depressed persons and suggestive of possible exaggeration.

A major limitation of the current study can be seen in the fact that only one PVT was used to define the fail and pass group although the WMT appeared to be one of the prime candidates for this. The study was performed with archival data, so the methodological scope was confined to available information. When the study protocol was introduced in 2012, the use of multiple PVTs was still in a relatively early phase; non-memory-based freestanding PVTs were not available in the German speaking countries. The Dot Counting Test (Rey, 1941) was discredited for performing far below modern PVT standards (Hartman, 2002), the b Test (Boone et al., 2002) was not available to German users, and the Groningen Effort Test (GET; Fuermaier et al., 2017) had not been published yet. Embedded PVTs were clearly under-investigated not only in the German-speaking countries but in the whole of Europe, and continue to be neglected.

The exclusion of patients with major neurocognitive impairment from the sample may counterbalance, to some degree, the risk of wrong classifications based on WMT results. Yet, the cut scores for the TAP *Alertness* subtest proposed in our paper will certainly benefit from being critically reviewed by other authors with different methodological approaches.

Funding Open access funding provided by the Medical University of Vienna.

Declarations

Conflict of Interest Thomas Merten is one of the authors of the Self-Report Symptom Inventory, which is commercially distributed by Hogrefe Publishers, Göttingen, Germany.

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*. The Psychological Corporation.

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., Lu, P., & Herzberg, D. (2002). *The b test*. 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

Cerny, B. M., Rhoads, T., Leib, S. I., Jennette, K. J., Basurto, K. S., Basurto, K. S., Durkin, N. M., Ovsiew, G. P., Resch, Z. J., & Soble, J. R. (2021). Mean response latency indices on the Victoria Symptom Validity Test do not contribute meaningful predictive value over accuracy scores for detecting invalid performance. *Applied Neuropsychology: Adult.* Advance online publication. https://doi.org/10.1080/23279095.2021.1872575

Cima, M., Hollnack, S., Kremer, K., Knauer, E., Schellbach-Matties, R., Klein, B., & Merckelbach, H. (2003). "Strukturierter Fragebogen Simulierter Symptome". Die deutsche Version des "Structured Inventory of Malingered Symptomatology: SIMS" [The German version of the Structured Inventory of Malingered Symptomatology]. *Der Nervenarzt*, *74*(11), 977–986. https://doi.org/10.1007/s00115-002-1438-5

Conners, C. K., & Staff, M. (2004). *Conners' Continuous Performance Test II (CPT II V. 5)*. Multi-Health Systems.

Czornik, M., Merten, T., & Lehrner, J. (2021). 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

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

Eglit, G. M. L., Jurick, S. M., Delis, D. C., Filoteo, J. V., Bondi, M. W., & Jak, A. J. (2020). Utility of the D-KEFS Color Word Interference Test as an embedded measure of performance validity. *The Clinical Neuropsychologist, 34*(2), 332–352. https://doi.org/10.1080/13854046.2019.1643923

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

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

Fuermaier, A. B. M., Tucha, L., Koerts, J., Aschenbrenner, S., & Tucha, O. (2017). *Vienna Test System (VTS): The Groningen Effort Test (GET).* Schuhfried.

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*(3), 218–224. https://doi.org/10.1037/1040-3590.6.3.218

Greve, K. W., Binder, L. M., & Bianchini, K. J. (2009). Rates of below-chance performance in forced-choice symptom validity tests. *The Clinical Neuropsychologist, 23*(3), 534–544. https://doi.org/10.1080/13854040802232690

Grossi, L. M., Green, D., Einzig, S., & Belfi, B. (2017). Evaluation of the Response Bias Scale and Improbable Failure Scale in assessing feigned cognitive impairment. *Psychological Assessment, 29*(5), 531–541. https://doi.org/10.1037/pas0000364

Groth-Marnat, G. (1990). *The handbook of psychological assessment* (2nd ed.). John Wiley & Sons.

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

Hartman, D. E. (2002). The unexamined lie is a lie worth fibbing. Neuropsychological malingering and the Word Memory Test. *Archives of Clinical Neuropsychology, 17*(7), 709–714. https://doi.org/10.1016/S0887-6177(01)00172-X

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.

Ingram, P. B., Golden, B. L., & Armistead-Jehle, P. J. (2020). Evaluating the Minnesota Multiphasic Personality Inventory–2–Restructured Form (MMPI-2-RF) over-reporting scales in a military neuropsychology clinic. *Journal of Clinical and Experimental Neuropsychology, 42*(3), 263–273. https://doi.org/10.1080/13803395.2019.1708271

Jansen, M. T., Tamm, S., Merten, T., Tamm, A., & Hofmann, M. J. (2020). Feigning memory impairment in a forced-choice task: Evidence from event-related potentials. *International Journal of Neurophysiology, 158*, 190–200. https://doi.org/10.1016/j.ijpsycho.2020.10.003

Kaminski, A., Merten, T., & Kobelt-Pönicke, A. (2020). Der Vergleich von drei Beschwerdenvalidierungstests in der stationären psychosomatischen Rehabilitation [Comparison of three symptom validity tests in a sample of psychosomatic inpatients]. *Zeitschrift Für Psychiatrie, Psychologie Und Psychotherapie, 68*(2), 96–105. https://doi.org/10.1024/1661-4747/a000408

Kanser, R. J., Rapport, L. J., Bashem, J. R., & Hanks, R. A. (2019). Detecting malingering in traumatic brain injury: Combining response time with performance validity test accuracy. *The Clinical Neuropsychologist, 33*(1), 90–107. https://doi.org/10.1080/13854046.2018.1440006

Köllner, V., & Schauenburg, H. (2012). *Psychotherapie im Dialog - Diagnostik und Evaluation [Psychotherapy in dialogue: Diagnosis and evaluation]*. Thieme.

LaDuke, C., Barr, W., Brodale, D. L., & Rabin, L. A. (2018). Toward generally accepted forensic assessment practices among clinical neuropsychologists: A survey of professional practice and common test use. *The Clinical Neuropsychologist, 32*(1), 145–164. https://doi.org/10.1080/13854046.2017.1346711

Larrabee, G. J. (2003). Detection of malingering using atypical performance patterns on standard neuropsychological tests. *The Clinical Neuropsychologist, 17*(3), 410–425. https://doi.org/10.1076/clin.17.3.410.18089

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. (2008). *T.O. V.A. professional manual. Test of Variables of Attention continuious performance test*. T.O.V.A. Company.

Lukács, G., Kawai, C., Ansorge, U., & Fekete, A. (2021). Detecting concealed language knowledge via response times. *Applied Linguistics* Review. https://doi.org/10.1515/applirev-2020-0130

Lupu, T., Elbaum, T., Wagner, M., & Braw, Y. (2018). Enhanced Detection of Feigned Cognitive Impairment Using per Item Response

Time Measurements in the Word Memory Test: Applied Neuropsychology: Adult, *25*(6), 532–542. https://doi.org/10.1080/23279095.2017.1341410

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/acaa017

Merten, T., Giger, P., Merckelbach, H., & Stevens, A. (2019). *Self-Report Symptom Inventory (SRSI) – deutsche Version [German version of the Self-Report Symptom Inventory]*. Hogrefe.

Merten, T., Kaminski, A., & Pfeiffer, W. (2020). Prevalence of overreporting 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

Merten, T., & Merckelbach, H. (2020). Factitious disorder and malingering. In J. R. Geddes, N. C. Andreasen, & G. M. Goodwin (Eds.), *The New Oxford Textbook of Psychiatry* (3rd ed., pp. 1342–1349). Oxford University Press.

Merten, T., Merckelbach, H., Giger, P., & Stevens, A. (2016). The Self-Report Symptom Inventory (SRSI): A new instrument for the assessment of symptom overreporting. *Psychological Injury and Law, 9*(2), 102–111. https://doi.org/10.1007/s12207-016-9257-3

Mestdagh, I., van Bergen, L., Kocken, C., Heyvaert, V., Cras, P., & van den Eede, F. (2019). Diagnosing solvent-induced chronic toxic encephalopathy: The effect of underperformance in neuropsychological testing. *International Journal of Psychiatry in Clinical Practice, 23*(3), 171–177. https://doi.org/10.1080/13651501.2019.1571210

Mittenberg, W., Patton, C., Canyock, E. M., & Condit, D. C. (2002). Base rates of malingering and symptom exeggeration. *Journal of Clinical and Experimental Neuropsychology, 24*(8), 1094–1102. https://doi.org/10.1076/jcen.24.8.1094.8379

Ord, A. S., Shura, R. D., Sansone, A. R., Martindale, S. L., Taber, K. H., & Rowland, J. A. (2021). Performance validity and symptom validity tests: Are they measuring different constructs? *Neuropsychology, 35*(3), 241–251. https://doi.org/10.1037/neu0000722

Patrick, S. D., Rapport, L. J., Kanser, R. J., Hanks, R. A., & Bashem, J. B. (2021). Performance validity assessment using response time on the Warrington Recognition Memory Test. *The Clinical Neuropsychologist, 35*(6), 1154–1173. https://doi.org/10.1080/13854046.2020.1716997

Pollock, B., Harrison, A. G., & Armstrong, I. T. (2021). What can we learn about performance validity from TOVA response profiles? *Journal of Clinical and Experimental Neuropsychology, 43*(4), 412–425. https://doi.org/10.1080/13803395.2021.1932762

Plomin, R. (1986). *Development, genetics, and psychology*. Lawrence Erlbaum.

Rey, A. (1941). L'examen psychologique dans les cas d'encéphalopathie traumatique [The psychological examination in cases of traumatic encephalopathy]. *Archives De Psychologie, 28*, 286–340.

Rogers, R., Salekin, R. T., Sewell, K. W., Goldstein, A. M., & Leonard, K. (1998). Acomparison of forensic and nonforensic malingerers: A prototypical analysis of explanatory models. *Law and Human Behavior, 22*(4), 353–367. https://doi.org/10.1023/A:1025714808591

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

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/acaa019

Smith, G. P., & Burger, G. K. (1997). Detection of Malingering: Validation of the Structured Inventory of Malingered Symptomatology (SIMS). *Journal of the American Academy Psychiatry and the Law, 25*(2), 183–189.

Steck, P., Reuter, B., Meir-Korrell, S., & Schçnle, P. (2000). Zur Simulierbarkeit von neuropsychologischen Defiziten bei Reaktions- 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., 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

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.

Strauss, E., Slick, D. J., Levy-Bencheton, J., Hunter, M., MacDonald, S. W. S., & Hultsch, D. F. (2002). Intraindividual variability as an indicator of malingering in head injury. *Archives of Clinical Neuropsychology, 17*(5), 423–444. https://doi.org/10.1016/S0887-6177(01)00126-3

Sturm, W., & Willmes, K. (2001). On the functional neuroanatomy of intrinsic and phasic alertness. *NeuroImage, 14*(1), 76–84. https://doi.org/10.1006/nimg.2001.083

Suchotzki, K., Verschuere, B., van Bockstaele, B., Ben-Shakhar, G., & Crombez, G. (2017). Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin, 143*(4), 428–453. https://doi.org/10.1037/bul0000087

Suhr, J. A., & Gunstad, J. (2007). Coaching and malingering: A review. Assessment of malingered neuropsychological deficits. 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., & Participants, C. (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)*. Carleton University.

Vagnini, V. L., Berry, D. T., Clark, J. A., & Jiang, Y. (2008). New measures to detect malingered neurocognitive deficit: Applying reaction time and event-related potentials. *Journal of Clinical and Experimental Neuropsychology, 30*(7), 766–776. https://doi.org/10.1080/13803390701754746

van Dyke, S. A., Millis, S. R., Axelrod, B. N., & Hanks, R. A. (2013). Assessing effort: Differentiating performance and symptom validity. *The Clinical Neuropsychologist, 27*(8), 1234–1246. https://doi.org/10.1080/13854046.2013.835447

Whiteside, D. M., Hunt, I., Choate, A., Caraher, C., & Basso, M. R. (2020). Stratified performance on the Test of Memory Malingering (TOMM) is associated with differential responding on the Personality Assessment Inventory (PAI). *Journal of Clinical and Experimental Neuropsychology, 42*(2), 131–141. https://doi.org/10.1080/13803395.2019.1695749

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: Instruction manual. 2010 edition. World Health Organization.

Zimmermann, P., & Fimm, B. (1995). *Testbatterie zur Aufmerksamkeitsprüfung (TAP)* [Test for Attention Performance]. Psytest.

Zimmermann, P., & Fimm, B. (2002). A test battery for attentional performance. In M. Leclercq & P. Zimmermann (Eds.), *Applied Neuropsychology of Attention: Theory, Diagnosis and Rehabilitation* (pp. 110–151). Taylor & Francis.

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