# Performance validity in clinical neuropsychological assessment

# Performance Validity in Clinical Neuropsychological Assessment: Base Rates, Impact of Feedback, and Relevance to Outcomes

Jeroen Roor

**ISBN:** 978-94-6483-560-1
**Cover drawing by:** Patty Roor. The evolution of the concept of performance validity.
**Cover and inside layout by:** Bregje Jaspers | www.proefschriftOntwerp.nl
**Printed by:** Ridderprint

# Performance Validity in Clinical Neuropsychological Assessment: Base Rates, Impact of Feedback, and Relevance to Outcomes

DISSERTATION

to obtain the degree of Doctor at Maastricht University,
on the authority of the Rector Magnificus,
Prof.dr. Pamela Habibović
in accordance with the decision of the Board of Deans,
to be defended in public

on Friday January 19th 2024

at 13:00 hours

by

Jeroen Jan Roor

**Supervisors**

Prof. dr. R.W.H.M. Ponds

Prof. dr. B. Dandachi-FitzGerald — Maastricht University & Open University Heerlen

**Co-supervisor**

Dr. M.J.V. Peters

**Assessment Committee**

Prof. dr. H.L.G.J. Merckelbach (chair)

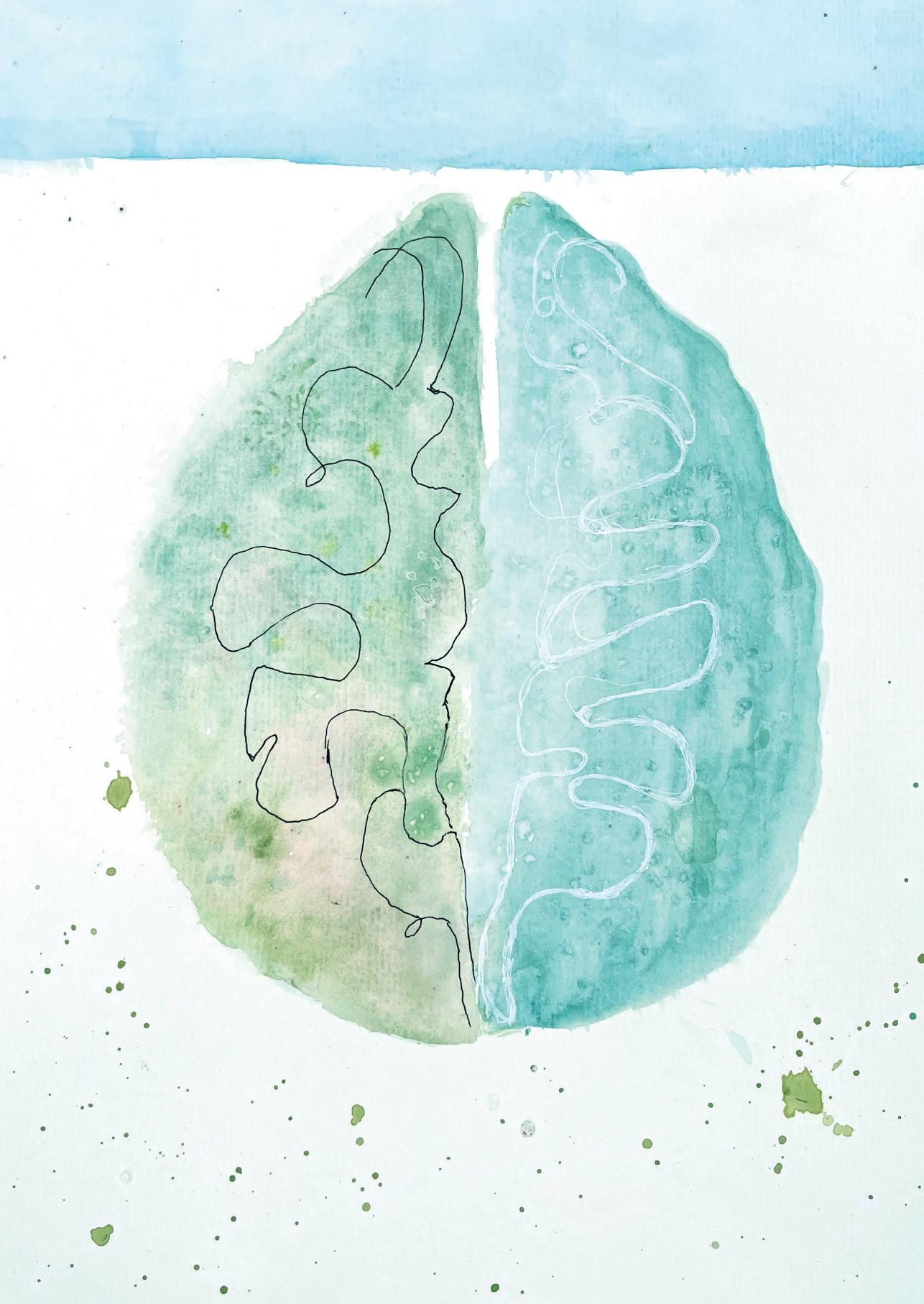
Dr. I. Bošković — Erasmus University Rotterdam

Dr. D.M.J.M. in de Braek

Prof. dr. M. Jelicic

Prof. dr. B. Schmand — University of Amsterdam

## CONTENTS



# CHAPTER 1

## General introduction and outline of the dissertation

## GENERAL INTRODUCTION

Psychological assessment is conducted to improve the understanding about the patient and his or her needs, to then use this information for diagnostic purposes and treatment recommendations. To be truly informative, results from the psychological assessment should provide accurate information about the symptoms a patient is experiencing and his/her true mental capabilities. When a patient is not answering accurately or honestly about experienced symptoms, or is performing below best of capabilities, the conditions necessary for obtaining accurate assessment results are not met.

Historically, clinicians believed that (nearly) all patients undergoing psychological assessment in routine clinical care would provide valid data. This belief stemmed from the assumption that patients are inherently motivated to provide accurate information regarding their symptoms and mental capacities. However, these assumptions are based on expectations that are unrealistically positive about patients' test-taking attitudes and general credibility of assessment outcomes. In the best-case scenario, these assumptions reflect understandable yet excessive optimism. However, in the worst-case scenario, they can be concerning as they have the potential to influence the perceived high value or validity that clinicians attribute to self-reported symptoms and cognitive task performance. This, in turn, may lead to the neglect of potential distortions in the presented clinical picture.

Recent survey studies confirmed that feigning symptoms (a) occurs regularly, and (b) that the associated motives are not restricted to external incentives (e.g., financial gain), but extent to those being more psychological in nature (e.g., to excuse a failure or seeking attention) (Dandachi-FitzGerald et al., 2020; Merten et al., 2023). These findings are well in line with current knowledge from the vast and growing body of research conducted within the field of symptom validity assessment, acknowledging that "most people engage in a variety of 'response styles' that reflect their personal goals in a particular setting" (p. 3, Rogers, 2008). Therefore, it can be argued that emphasizing or over-reporting symptoms, performing below best of capabilities on formal cognitive tests, omitting relevant details, or concealing anticipated disadvantageous behavior (e.g., substance abuse) at some time or the other are common in all people. It is not helpful to see this behavior as a moral failing. Clinicians may want to embrace the fact that non-credible reporting is ubiquitous and equip themselves with strategies that may help them in approaching patients displaying related behaviors (e.g., *see* Beach et al., 2017 for a detailed training program for psychiatric trainees). As described by Stone and Boone (2007), "recognition of feigning behaviors may prove to be the first therapeutic step to understanding the patient's actual needs" (p. 11).

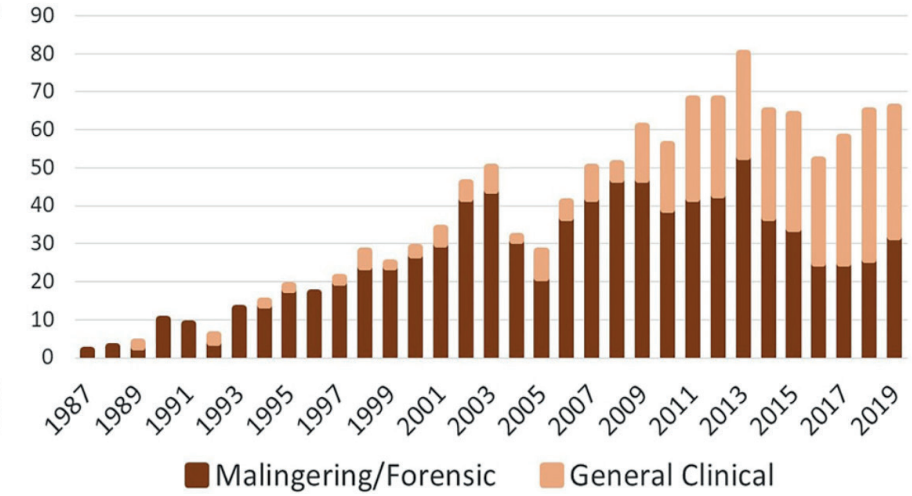
### Validity Assessment; from Forensics to Routine Clinical Care

Initially, research on assessing the credibility or reported symptoms and test performance mainly focused on the detection of malingering, typically in forensic contexts. The fourth edition of the Diagnostic and Statistical Manual of mental disorders (DSM-IV; American Psychiatric Association, 1994) defined malingering as the "intentional production of false or grossly exaggerated physical or psychological symptoms, motivated by external incentives such as avoiding military duty, avoiding work, obtaining financial compensation, evading criminal prosecution, or obtaining drugs" (p. 683). In the last two to three decades, there has been a steady and continuing growth in publications on validity assessment

(VA) in the context of routine clinical care, whereas literature focusing on forensic settings began to plateau (Figure 1).

**Figure 1**

*Number of Validity Assessment (VA) Publications in the Forensic Versus Clinical Context Identified in PsycINFO and MEDLINE Databases from 1987 through 2019*




*Note.* Reprinted from Suchy (2019), with permission.

These developments have recently led to the update of the highly influential diagnostic criteria for Malingered Neurocognitive Deficit (MND) that leaned heavily on the DSM-IV description of malingering (Slick et al., 1999). In the updated MND criteria (now termed Multidimensional Criteria for Neurocognitive, Somatic, and Psychiatric Malingering), it is stated that "external incentives from the MND model were too biased toward criminal and forensic settings" (Sherman et al., 2020, p. 737). Therefore, amongst others, these authors expanded the criteria for presence of external gains for malingering, by not only including high-stakes undesirable outcomes (e.g., avoiding criminal prosecution) but also adding lower-stakes undesirable outcomes, such as avoiding having to fulfill more basic duties and responsibilities such as avoiding work, school examinations, or home responsibilities. These latter "lower stakes" are likely far more prevalent in routine clinical care compared to the external gains criteria for malingering from the original MND model (Slick et al., 1999). This clearly illustrates the idea that the concept of malingering symptoms has evolved and is no longer reserved for forensic contexts alone, but may also apply to routine clinical care.

**Validity Tests**

In their seminal study, Heaton and colleagues (1978) send out test protocols containing the results on a self-report measure (Minnesota Multiphasic Personality Inventory; MMPI) and a battery of cognitive tests (Wechsler Adult Intelligence Scale; WAIS, and Halstead-Reitan Neuropsychological Test Battery for Adults) of 16 instructed malingerers and 16 bona fide head-injury patients to ten neuropsychologists. The healthy volunteers in the malingering group "were encouraged to fake the most severe [head-injury related] disabilities that they could, without making it obvious to the examiner that they were faking" (p. 894). The "success" of these neuropsychologists in identifying the test protocols as belonging to head-injury patients or instructed malingerers ranged from chance-level to about 20% better that chance. The study of Heaton and colleagues (1978) clearly demonstrated that clinicians would have concluded a substantial proportion of instructed malingerers as brain-injury patients and vice versa. This study underscores that, instead of exclusively relying on clinical judgment, validity assessment is essential in accurately evaluating cognitive abilities in neuropsychological assessments to prevent misdiagnosis.

Validity assessment encompasses the validity evaluation of self-reported symptoms as well as cognitive capabilities. Symptom validity tests (SVTs) are questionnaires designed to measure implausible symptom endorsement of items that are rare, atypical, or improbable. Performance validity tests (PVTs) are designed to measure implausible low performance on cognitive tests. Both SVTs and PVTs come in embedded and freestanding form. Embedded SVTs are 'built into' regular questionnaires for assessing all sorts psychological constructs (e.g., depression, personality traits), whereas freestanding SVTs are specifically designed to measure the accuracy of symptom reports. Embedded PVTs are 'built into' or derived from standard neuropsychological ability tests, and freestanding PVTs are specifically developed to address the validity status of cognitive performance. Freestanding PVTs are presented as regular tests of cognitive functioning, though they are - by definition - largely insensitive to cognitive disfunction (i.e., the floor effect strategy). In this dissertation, the focus is mainly on the identification of performance below best of capabilities (i.e., invalid performance) in clinical neuropsychological assessments using freestanding PVTs.

**Explanations for Performance Validity Test (PVT) Failure**

Previously, PVT failure was equated to malingering (Merten et al., 2022; Sherman et al., 2020). Nowadays, however, it is well established that PVT failure in itself does not provide information about the underlying motives of the examinee. In fact, PVT failure can be the result of a myriad of other factors conceptually different from malingering, such as careless responding, factitious disorder, or severe cognitive or psychiatric pathology (*see* Dandachi-FitzGerald et al., 2022 for proposed explanatory levels of validity test failure). It is important to recognize that in most cases, the cause(s) for PVT failure will remain unclear. In this instance, the clinician should stay abreast from seeking alternative reasons to try to explain-away this important finding. As clearly stated in the updated consensus statement on validity assessment, "empirical research currently does not support interpretation of multiple PVT failures as due to depression, anxiety, pain, fatigue, medication effects, or other putative conditions that might inaccurately be used to minimize or explain away PVT findings" (Sweet et al., 2021, p. 1069). Therefore, instead, the clinician should simply describe the performance below best of capabilities just as it is: invalid (Schroeder & Martin, 2022).

**Validity Testing in Clinical Practice**

The last two to three decades, it became increasingly clear that a substantial proportion of patients seen during routine clinical care failed SVTs and PVTs (e.g., Dandachi-FitzGerald et al., 2011). Moreover, in clinical assessments, the impact of PVT failure was found to obscure the expected brain-behavior relationship. For example, Rienstra and colleagues (2013) found an expected strong association between hippocampal volume using magnetic resonance imaging (MRI) and memory test performance in older patients seen for cognitive evaluation in a clinical context, but only *after* excluding those who did not perform to the best of their capabilities (i.e., failed a freestanding PVT). Research extending Heaton et al.'s (1987) study revealed that clinicians cannot accurately predict the added value of incorporating SVTs or PVTs into the test battery. In the Dandachi-FitzGerald et al. (2017) study, most clinicians indicated including validity tests in the battery only when they suspected distorted symptom reporting (e.g., based on known presence of external incentives or the clinical interview). However, their predictions made after reviewing the medical file and conducting a clinical interview but prior to testing were prone to errors. Therefore, in line with the mentioned conceptual developments regarding performance and symptom validity and related continued research, consensus statements on validity assessment by the American Academy of Clinical Neuropsychology and clinical recommendations on validity assessment by the British Psychological Society - both first published in 2009 - were updated (Sweet et al., 2021; Moore et al., 2021). One of the shared key recommendations from these guidelines, is to include PVTs in *every* assessment.

**Ongoing Questions Related to Performance Validity Assessment in Routine Clinical Care**

The shift from studying validity assessment in the forensic setting to routine clinical care, comes with various challenges. Contrary to the forensic setting, in routine clinical care, the clinician typically forms a doctor-patient relationship with the examinee. Due to the nature of this relationship, the clinician acts as an advocate for the patient and attempts to minimize harm, while maintaining confidentiality (American Psychological Association, 2017). Patients, on the other hand, do their best to provide accurate and complete information in this patient-doctor collaboration. However, when a patient shows signs of invalid symptom reporting or performance, the collaboration between the patient and the clinician is potentially compromised. It is important to keep in mind that, also in the presence of invalid responding, (neuro)psychological assessment in routine clinical care is primarily concerned with the question how to provide useful services. This is very different from the forensic setting, where a doctor-patient relationship is non-existent and the examiner reports back to a "third party" (American Psychological Association, 2017). Arguably, management of invalid performance in the clinical setting poses more challenges to the clinician as compared to the forensic context.

Survey results indicate that most clinicians acknowledge the mentioned consensus statements of professional organizations on validity assessment, and use PVTs during clinical neuropsychological assessment (Hirst et al., 2017). However, these consensus statements primarily focus on diagnostic features and research, providing little practical guidance. Therefore, despite clear improvements in the assessment of performance validity in routine clinical care, it is obvious that there still is a significant translational gap between the mentioned guidelines concerning the use of validity tests and their

application in clinical practice. Several lingering open ends have not yet been addressed properly, that may aid in assessing performance validity in routine clinical care.

First to mention is **the prevalence rate** of clinical patients who perform below the best of their capabilities. In the forensic setting, the base rate of PVT-failure is well explored and found to hover around 40% (Chafetz et al., 2007; Martin & Schroeder, 2020). However, empirical information about the base rate of PVT-failure in the clinical setting is lagging behind. In a frequently cited publication by Mittenberg and colleagues (2002), 131 practicing neuropsychologists were surveyed about the percentage of annual cases that involved probable symptom exaggeration or malingering. These authors found a *reported* base rate of 8% in general clinical cases, to about 30% in clinical cases that are potentially compensable (e.g., mild head injury or fibromyalgia). A more recent survey found a median estimated base rate of invalid performance and symptom reporting of 15% across non-forensic clinical evaluations (Martin & Schroeder, 2020). However, to make accurate decisions regarding performance validity, *empirical* base rate information about PVT failure is crucial in assisting the clinician to make accurate inferences about the validity of a patients' performance on cognitive tests. For example, clinically applied statistics such as positive predictive value (PPV; the likelihood that PVT failure reflects true invalid performance) and negative predictive value (NPV; the likelihood that passing a PVT reflects performance in line with true capabilities) can only be calculated using reliable empirical base rate information (Dandachi-FitzGerald & Martin, 2022). Therefore, it is essential that base rate information is available for each PVT in a specific clinical context and, ideally, for specific clinical patient groups (Schroeder et al., 2021). Besides, external gain incentives – a known driver of PVT-failure – are also present in patients undergoing routine clinical evaluations (Schroeder et al., 2022). Therefore, this factor should also be taken into account when establishing base rates.

Second, as mentioned as point 9.10 of the Ethical Principles and Code of Conduct of the American Psychological Association, "psychologists take reasonable steps to ensure that explanations of [assessment] results are given to the individual" (p. 14, American Psychological Association, 2017). In line with this guideline, most clinicians *reportedly* **provide feedback on invalid performance** with their patients -based upon surveys conducted in Europe (Dandachi-FitzGerald et al., 2013) and the United States (Martin et al. 2015). To date, however, the effects of such feedback interventions on invalid performance are largely unknown. The mentioned survey studies show large variability in the timing of the feedback (during and/or after completion of the neuropsychological assessment) and especially in the manner in which clinicians communicate indications of invalid performance. While most reported (a) to express that no firm diagnostic conclusions can be drawn, or (b) that test results are inconsistent with severity of injury, there was no consensus on these specific communication components. Obviously, the nature of the feedback intervention itself (i.e., that there are indications that the patient performed below best of capabilities) can be a reason for clinicians to struggle with providing feedback to patients about invalid performance.

Third, although the impact of performance invalidity on neuropsychological test performance is known to be as large or even greater than various medical and psychiatric conditions (e.g., Iverson, 2006), it's potential **relevance to treatment outcomes** remains largely unknown. This omission is striking, as inaccurate diagnostic conclusions (resulting from invalid performance) likely also result in

inappropriate recommendations for treatment. One can safely argue that treatments based upon invalid data are likely not targeting the expected (medical) condition/symptoms, and that these treatments therefore are less efficient and/or not in line with the specific needs of the patient. In a worst-case scenario, these treatments may even have iatrogenic effects.

**Dissertation Aims and Outline**

The main objective of the studies described in this dissertation was to gain more insight into the prevalence rate of invalid performance, the impact of feedback interventions upon indications of invalid performance, and the relevance of invalid performance to treatment outcome in the clinical setting. In Chapter 2, a detailed and illustrative case report serves as a starting point to demonstrate these aspects of performance validity assessment in routine clinical care, leading up to the research questions addressed in this dissertation. More specifically, these research questions are:

1. What is the prevalence of PVT failure in the clinical setting, and what are relevant mediating factors for calculating clinically applied statistics? (Chapter 3).
2. What are the effects of a feedback intervention upon psychometric evidence of invalid performance? (Chapters 4 and 5).
3. What is the impact of PVT failure on treatment outcome? (Chapter 6).

The main results, as well as methodological and clinical considerations, and directions for future research, are discussed in Chapter 7.

## REFERENCES


American Psychiatric Association (1994). Diagnostic and statistical manual of mental disorders (4th edition). American Psychiatric Press

American Psychological Association (2017). Ethical principles of psychologists and code of conduct. Retrieved May 19, 2023, from https://www.apa.org/ethics/code/

Beach, S. R., Taylor, J. B., & Kontos, N. (2017). Teaching psychiatric trainees to "think dirty": Uncovering hidden motivations and deception. *Psychosomatics*, *58*(5), 474–482. https://doi-org.mu.idm.oclc.org/10.1016/j.psym.2017.04.005

Chafetz, M. D., Abrahams, J. P., & Kohlmaier, J. (2007). Malingering on the Social Security disability consultative exam: A new rating scale. *Archives of Clinical Neuropsychology, 22*(1), 1–14, https://doi. 10.1080/13854046.2021.1895322

Dandachi-FitzGerald, B., Merckelbach, H., Bošković, I., & Jelicic, M. (2020). Do you know people who feign? Proxy respondents about feigned symptoms. *Psychological Injury and Law, 13*(3), 225–234. https://doi.org/10.1007/s12207-020-09387-6

Dandachi-FitzGerald, B., & Martin, P. K. (2022). Clinical judgement and clinically applied statistics: Description, benefits, and potential dangers when relying on either one individually in clinical practice. In R. W. Schroeder & P. K. Martin (Eds.), *Validity assessment in clinical neuropsychological practice; evaluating and managing noncredible performance* (pp. 107–125). The Guilford Press

Dandachi-FitzGerald, B., Merckelbach, H., & Merten, T. (2022). Cry for help as a root cause of poor symptom validity: A critical note. *Applied Neuropsychology. Adult*, 1–6. Advance online publication. https://doi.org/10.1080/23279095.2022.2040025

Dandachi-FitzGerald, B., Merckelbach, H., & Ponds, R. W. (2017). Neuropsychologists' ability to predict distorted symptom presentation. *Journal of Clinical and Experimental Neuropsychology*, *39*(3), 257–264. https://doi-org.mu.idm.oclc.org/10.1080/13803395.2016.1223278

Dandachi-FitzGerald, B., Ponds, R. W., Peters, M. J., & Merckelbach, H. (2011). Cognitive underperformance and symptom over-reporting in a mixed psychiatric sample. *The Clinical Neuropsychologist*, *25*(*5*), 812–828, https://doi.org/10.1080/13854046.2011.583280

Dandachi-FitzGerald, B., Ponds, R. W., & 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

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

Hirst, R. B., Han, C. S., Teague, A. M., Rosen, A. S., Gretler, J., & Quittner, Z. (2017). Adherence to validity testing recommendations in neuropsychological assessment: A survey of INS and NAN members. *Archives of Clinical Neuropsychology, 32*(4), 456–471, https://doi.org/10.1093/arclin/acx009

Iverson G. L. (2006). Ethical issues associated with the assessment of exaggeration, poor effort, and malingering. *Applied Neuropsychology*, *13*(2), 77–90. https://doi-org.mu.idm.oclc.org/10.1207/s15324826an1302_3

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.mu.idm.oclc.org/10.1093/arclin/acaa017

Martin, P. K., Schroeder, R. W., & Odland, A. P. (2015). Neuropsychologists' validity testing beliefs and practices: A survey of north American professionals. *The Clinical Neuropsychologist, 29*(6), 741-776, https://doiorg.mu.idm.oclc.org/10.1080/13854046.2015.1087597

Merten, T., Dandachi-FitzGerald, B., Hall, V., Bodner, T., Giromini, L., Lehrner, J., González-Ordi, H., Santamaría, P., Schmand, B., & Di Stefano, G. (2022). Symptom and performance validity assessment in European countries: an update. *Psychological injury and law*, *15*(2), 116–127. https://doi-org.mu.idm.oclc.org/10.1007/s12207-021-09436-8

Merten, T., Tucha, L., Giger, P., Niesten, I. J., Tucha, O., & Fuermaier, A. B. (2023). Laypeople's prevalence estimates of malingering: Survey data from the Netherlands. *Psychology & Neuroscience*. Advance online publication. https://dx.doi.org/10.1037/pne0000303

Mittenberg, W., Patton, C., Canyock, E. M., & Condit, D. C. (2002). Base rates of malingering and symptom exaggeration. *Journal of Clinical and Experimental Neuropsychology, 4*(8), 1094–1102.

Moore, P., Bunnage, M., Kemp, S., Dorris, L., & Baker, G. (2021). *Guidance on the assessment of performance validity in neuropsychological assessment*. The British Psychological Society.

Rienstra, A., Groot, P. F., Spaan, P. E., Majoie, C. B., Nederveen, A. J., Walstra, G. J., de Jonghe, J. F., van Gool, W. A., Olabarriaga, S. D., Korkhov, V. V., & Schmand, B. (2013). Symptom validity testing in memory clinics: Hippocampal-memory associations and relevance for diagnosing mild cognitive impairment. *Journal of Clinical and Experimental Neuropsychology*, *35*(1), 59–70. https://doi.org/10.1080/13803395.2012.751361

Rogers, R. (Ed.) (2008). *Clinical assessment of malingering and deception* (3rd edition). The Guilford Press.

Schroeder**,** R. W., Boone, K. B., & Larrabee, G. J. (2021). Design methods in neuropsychological performance validity, symptom validity, and malingering research. *In* K. B. Boone (Ed), *Assessment of feigned cognitive impairment: A neuropsychological perspective, 2nd ed*. (pp. 11-33). The Guilford Press,

Schroeder, R. W., Clark, H. A., & Martin, P. K. (2022). Base rates of invalidity when patients undergoing routine clinical evaluations have social security disability as an external incentive. *The Clinical Neuropsychologist*, *36*(7), 1902–1914, https://doi-org.mu.idm.oclc.org/10.1080/13854046.2021.1895322

Schroeder, R. W. & Martin, P. K. (2022). Explanations of performance validity test failure in clinical settings. *In* R. W. Schroeder, & P. K. Martin (Eds), *Validity assessment in clinical neuropsychological practice: Evaluating and managing noncredible performance* (pp. 11-30). The Guilford Press.

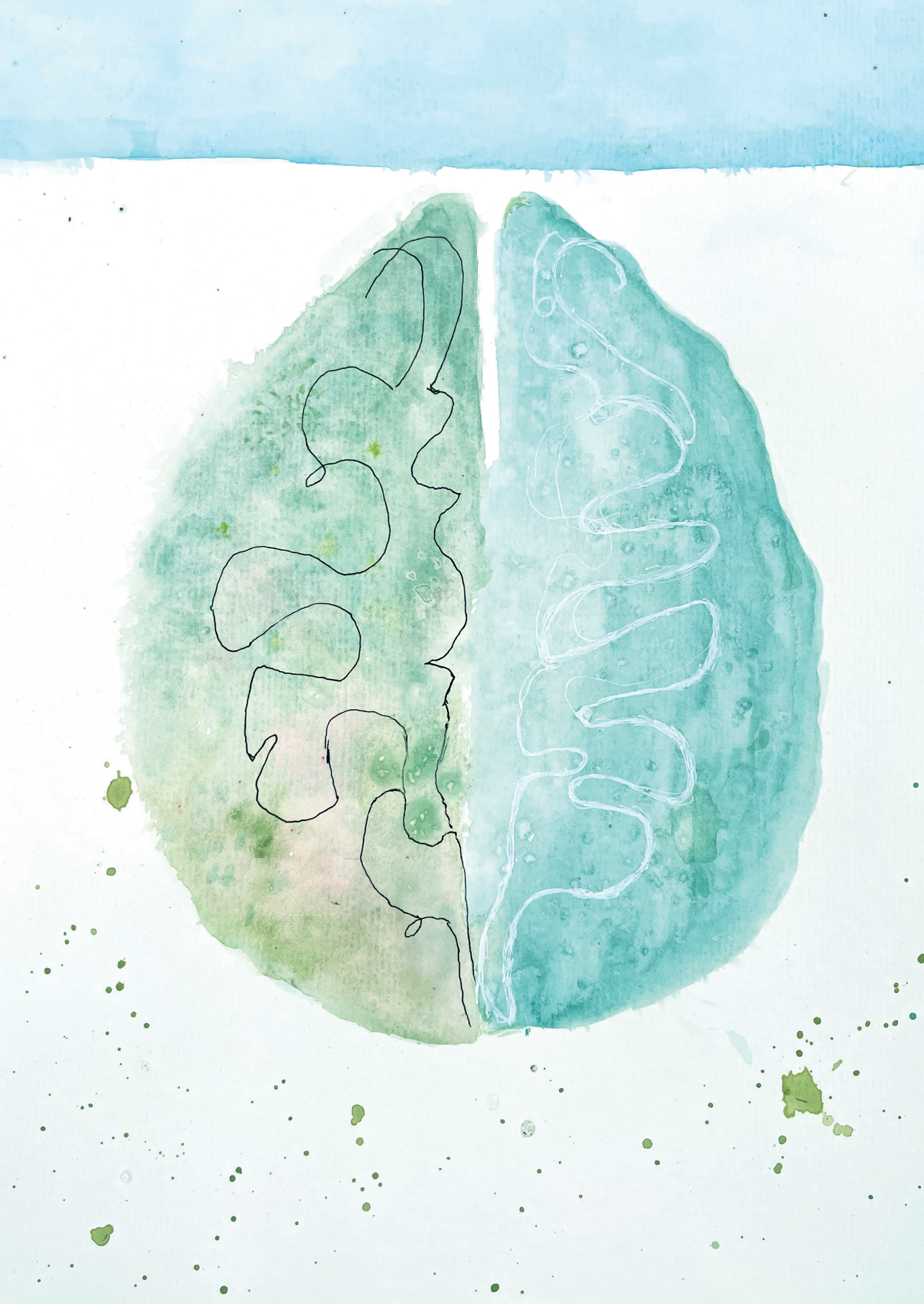
Sherman, E. M. S., Slick, D. J., & Iverson, G. L. (2020). Multidimensional malingering criteria for neuropsychological assessment: A 20-year update of the malingered neuropsychological dysfunction criteria. *Archives of Clinical Neuropsychology*, *35*(6), 735–764. https://doi.org/10.1093/arclin/acaa019

Slick, D. J., Sherman, E. M., & Iverson, G. L. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: proposed standards for clinical practice and research. *The Clinical Neuropsychologist*, *13*(4), 545–561. https://doi.org/10.1076/1385-4046(199911)13:04;1-Y;FT545

Stone, D. C. & Boone, K. B. (2007). Feigning of physical, psychiatric, and cognitive symptoms. Examples from history, the arts, and animal behavior. In Boone, K. B. (Ed.), *Assessment of feigned cognitive impairment. A neuropsychological perspective* (pp. 3-12). The Guilford Press.

Suchy, Y. (2019). Introduction to special issue: Current trends in empirical examinations of performance and symptom validity. *The Clinical Neuropsychologist, 33*(8), 1349-1353, https://10.1080/13854046.2019.1672334

Sweet, J. J., Heilbronner, 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



# CHAPTER 2

## A case of misdiagnosis of mild cognitive impairment: The utility of symptom validity testing in an outpatient memory clinic

Roor, J. J., Dandachi-FitzGerald, B., & Ponds, R. W. (2016). A case of misdiagnosis of mild cognitive impairment: The utility of symptom validity testing in an outpatient memory clinic. *Applied Neuropsychology: Adult, 23*(3), 172-178

## ABSTRACT

Noncredible symptom reports hinder the diagnostic process. This fact is especially the case for medical conditions that rely on subjective report of symptoms instead of objective measures. Mild cognitive impairment (MCI) primarily relies on subjective report, which makes it potentially susceptible to erroneous diagnosis. In this case report, we describe a 59-year-old female patient diagnosed with MCI 10 years previously. The patient was referred to the neurology department for reexamination by her general practitioner because of cognitive complaints and persistent fatigue. This case study used information from the medical file, a new magnetic resonance imaging brain scan, and neuropsychological assessment. Current neuropsychological assessment, including symptom validity tests, clearly indicated noncredible test performance, thereby invalidating the obtained neuropsychological test data. We conclude that a blind spot for noncredible symptom reports existed in the previous diagnostic assessments. This case highlights the usefulness of formal symptom validity testing in the diagnostic assessment of MCI.

## INTRODUCTION

Medical conditions based on subjective report of symptoms can be difficult to evaluate because the validity of reported symptoms can be difficult to discern. Mild cognitive impairment (MCI, Albert et al., 2011; Petersen et al., 1999) could be considered to be such a condition because the diagnostic criteria strongly emphasize subjective reports rather than objective measures (e.g., Han et al., 2011; Lenahan et al., 2012; Mitchell, 2008). MCI is considered to be a transitional stage between the cognitive changes of normal aging and dementia. The clinical definition of MCI includes: (a) cognitive concerns reflecting a change in cognition reported by the patient, informant, or clinician; (b) objective evidence of impairment in one or more cognitive domains; (c) preservation of independence in functional abilities; and (d) no dementia (Albert et al., 2011). For the second criterion, neuropsychological tests are used as an objective measure of cognitive functioning. However, a practical method for identifying such objective cognitive impairment is generally lacking. For example, the necessary neuropsychological tests have not yet been specified, no consensus has been reached on the appropriate cutoffs for neuropsychological tests (i.e., –1.0, –1.25, or –1.5 standard deviations below the mean), and potential confounding factors in neuropsychological assessment have not been identified. These facts make the diagnostic process for MCI susceptible to misclassification.

A significant proportion of patients diagnosed with MCI do not convert to dementia; some even revert to normal functioning (Visser et al., 2006; Vos et al., 2013). For example, Koepsell and Monsell (2012) found that 16% of patients with MCI reverted to normal cognitive functioning after 1 year. Although these patients satisfied the MCI criteria at baseline, their temporary cognitive decline was most likely not due to a brain disease (i.e., Alzheimer disease). The limited predictive validity of MCI is typically explained by differences in age and in the definition of MCI used (Visser & Verhey, 2008). However, the role of noncredible symptom reports in the incorrect diagnoses of MCI is underexamined.

Neuropsychological assessment plays a significant role in MCI diagnosis. In fact, neuropsychological assessment is the only "objective" measure of cognitive decline. It is crucial for clinicians to be confident that they have obtained a valid estimate of an individual's current functioning across a range of cognitive domains. However, noncredible performance undermines the validity of neuropsychological test results (Bush et al., 2005). If patients exert insufficient effort to perform to the best of their abilities, abnormal test results may not validly reflect cognitive impairment. Clinicians' ability to accurately judge a patient's motivation and effort during testing has been criticized (e.g., Faust, 1995), leading to the development of specialized symptom validity tests (SVTs; for an overview, *see* Boone, 2007). Guidelines state that adequate assessment of symptom validity is considered an essential part of every neuropsychological evaluation (Bush et al., 2005; Heilbronner et al., 2009). To the best of our knowledge, Rienstra et al. (2013) were the first to examine the symptom validity of neuropsychological test results in patients diagnosed with MCI. They found that an otherwise strong correlation between neuropsychological test results and hippocampal volume was nearly absent in patients who failed the SVTs. In other words, correcting for noncredible performance increases the accuracy with which cognitive impairments due to cerebral dysfunction are classified.

The present case concerns a patient diagnosed with MCI in an outpatient memory clinic. Ten years after the initial MCI diagnosis, repeated neuropsychological assessment that included SVTs clearly indicated noncredible test performance, which invalidated the data obtained from the standard cognitive tests (e.g., memory tests). This case demonstrates that MCI diagnosis is susceptible to false-positive errors when symptom credibility is not evaluated during the diagnostic assessment.

## METHOD

### Case

The patient was a 59-year-old woman who in 2012 was referred by her general practitioner to the neurology department of a general hospital to be treated for fatigue and cognitive complaints. The diagnostic evaluation consisted of a magnetic resonance imaging (MRI) brain scan, a neuropsychological assessment by the medical psychology department, and a study of her medical file, which contained two prior examinations performed at an outpatient memory clinic that included neuropsychological assessments (2001 and 2009). After the new neuropsychological evaluation was completed and feedback was provided, the patient signed an informed consent form to permit the anonymous publication of a case report including the information in the medical file, MRI brain scan, and neuropsychological test results.

### Background Information

The medical file showed a 30-year history of medically unexplained symptoms such as pain and fatigue. In the previous 30 years, the patient had undergone numerous medical examinations and treatments. In 2000, she was admitted to a psychiatric clinic. There, she was diagnosed with chronic fatigue syndrome (CFS) and fibromyalgia syndrome (FS). In 2001, the patient was referred to an outpatient memory clinic by her general practitioner because of memory complaints. A geriatrician, a neurologist, and a neuropsychologist arrived at a consensus diagnosis based on information from the medical chart, a computed tomography (CT) brain scan, and a neuropsychological examination (*see* Table 1). The CT brain scan showed no abnormalities. The neuropsychological assessment, however, showed notably low scores on tests measuring attention, executive functioning, and memory. Based on these test scores and the subjective cognitive complaints, the patient was diagnosed with MCI. In 2009, the patient contacted the memory clinic and requested a new examination. During the interview with the geriatrician, she stated that a new examination might result in an extension of her disability benefits, for which she qualified after being diagnosed with MCI in 2001. A second neuropsychological assessment was conducted. Again, the patient performed extremely poorly on the cognitive tests administered (*see* Table 1). The team reconfirmed the MCI diagnosis.

**Table 1**

*2001 and 2009 Neuropsychological Test Results*

| | 2001 | | 2009 | |
|---|---|---|---|---|
| **Tests** | **Scores** | **Percentiles** | **Scores** | **Percentiles** |
| CST-20 | 16/20 | | | |
| TMT-A | 49 s | 16 | 61 s | 7 |
| TMT-B | 216 s | 0 | Aborted | - |
| Stroop I | 66 s | 1 | 97 s | 0 |
| Stroop II | 91 s | 14 | 99 s | 0 |
| Stroop III | 214 s | 4 | 437 s | 0 |
| Digit Span forward | 5 | 30-40 | 4 | 20-30 |
| Digit span backward | 2 | 0 | 2 | 0 |
| VAT long | 33 | 34 | 32 | 25 |
| VLT | | | | |
| IR | 30 | 0 | 28 | 0 |
| DR | 6 | 10-20 | 3 | 0 |
| RC | 24/30 | | 21/30 | |
| RMT-F | 28/50 | 0 | 31 | 0 |
| RCFT | 34/36 | | 34/36 | |

*Note.* CST-20 = cognitive screening test-20; TMT-A and B = trail making test-a and b; VAT = visual association test long version; VLT = verbal learning test; IR = immediate recall; DR = delayed recall; RC = recognition; RMT-F = recognition memory test for faces; RCFT = rey complex figure test.

## Measures

Three well-validated SVTs were administered: the Amsterdam Short-Term Memory test (ASTM; Schmand & Lindeboom, 2005), the Word Memory Test (WMT; Green, 2003), and the Dutch version of the Structured Inventory of Malingered Symptomatology (SIMS; Merckelbach et al., 2001).

The ASTM is a forced-choice verbal recognition task that consists of 30 items. In each item, the participant is presented with five printed words from the same semantic category and is asked to read them aloud and remember them. Then, a simple calculation task that must be solved mentally is presented as a distractor. Finally, five words from the same semantic category as the five words shown earlier are presented. The participant must indicate which three words were also presented in the first series. The maximum score is 90 points (i.e., 30 items three words correct), and a total score of 85 was the original proposed cutoff. In the original validation studies, a cutoff score of 85 best distinguished between experimental simulators ($N = 57$) and the aggregated groups of patients with neurological disorders such as mild traumatic brain injury, multiple sclerosis, and severe epilepsy ($N = 57$) with a sensitivity of 84% and specificity of 90% (Schmand et al., 1999). A recent validation study by Rienstra and colleagues (2010) revealed that children older than 9 years of age all passed the ASTM.

The WMT is also a forced-choice recognition task. All five subtests of the WMT are based on a list of 20 pairs of semantically related words that are presented twice at the beginning of the task on a computer screen. In the Immediate Recall (IR) task, the participant is presented with 40 pairs of words and is asked to choose which word in each pair was shown earlier. This task is repeated 20 min later in the Delayed Recognition (DR) task. IR, DR, and the consistency between these two measures (CNS) are considered the three SVT measures of the WMT. The SVT measures are followed by a series of memory tests that gradually increase in difficulty and are sensitive to genuine verbal memory impairment. The original proposed cutoff was 82.5% for each of the three effort measures (i.e., IR, DR, and CNS). The WMT has undergone extensive validation in various populations (for an overview, see http:// www.wordmemorytest.com), which showed that the vast majority of patients with neurological conditions passed the SVT measures of the WMT (e.g., Carone et al., 2013).

The SIMS is a self-report questionnaire designed to screen for malingering psychiatric symptoms and/or cognitive impairment. The SIMS consists of 75 true/false items and contains five scales that assess commonly feigned conditions: amnesia, neurologic impairment, psychosis, affective disorder, and low intelligence. Using a cutoff of 16 points, Merckelbach and Smith (2003) found a sensitivity of 93% and a specificity of 98%.

In addition, the Self-Deceptive Enhancement subscale of the Balanced Inventory of Desirable Responding (SDE-BIDR; Paulhus, 2002) was administered to measure the tendency for self-deception. The SDE-BIDR is designed to measure exaggeration of one's positive attributes, which is related to poor introspective abilities. The scale consists of 20 statements. Using a 7-point scale, the participant indicates to what extent the items are true. The total score ranges from 20 to 140, with higher scores indicating greater self-deceptive tendencies. Merckelbach and colleagues (2011) found good internal consistency for the SDE-BIDR (Cronbach's alpha = .80).

The manuals of the aforementioned tests provide additional details regarding test procedures and materials. Standard neuropsychological tests (i.e., Montreal Cognitive Assessment, Trail-Making Test, Visual Association Test, clock drawing, bike drawing, and house drawing), psychological complaints (i.e., Symptom Checklist-90, Beck Depression Inventory), and personality questionnaires (i.e., Minnesota Multi-phasic Personality Inventory-2 [MMPI-2]) were also used in the evaluation.

## Interview

### *Symptoms*

The patient spontaneously reported a myriad of complaints, including fatigue, forgetfulness, and problems using her legs. Complaints of anxiety or depression were denied. She stated that she was sometimes unable to use her legs and had to use a wheelchair. When queried about the precise nature of her cognitive problems, she emotionally exclaimed, "Oh my god, this is what I mean. I cannot recall my own problems!" She reported that she cannot make simple calculations, such as adding 1 + 1.

### *Adaptive Functioning and Living Situation*

The patient was on welfare and received personal disability benefits. She was pursuing a career as a painter and a musical instrument maker and held frequent exhibitions. Her paintings are displayed on

her website, and she generates an income from selling these works. She also volunteers for an association for patients with CFS. The patient used a neck brace, wrist brace, orthopedic shoes, crutches, elbow crutches, and a wheelchair. At home, she had access to a triple chair and an adjustable chair and bed. For transportation, she had access to a scoot mobile, an electric bicycle, and a car. Her house was equipped with an adjustable kitchen and a staircase elevator. The patient received treatment at a day-care facility and home treatment from a physiotherapist and an occupational therapist. She participated in hydrotherapy as a member of a group for patients with rheumatic diseases. Home care services assisted her with housekeeping. The patient was able to manage her finances, her appointments, her travel, and her career as an artist.

***Observation***

During the interview, no cognitive problems were observed. The patient was able to thoroughly describe her medical history and both past and present details of her personal life. During testing, the patient stated that she could not continue with the assessment because of physical complaints, such as pain, dizziness, and fatigue. She needed encouragement to continue the tests. The neuropsychological assessment was conducted during 4 days due to the mentioned physical complaints.

**MRI Brain Scan**

According to the neuroradiologist, the MRI brain scan showed no abnormalities.

## RESULTS

As Table 2 shows, the patient scored below the cutoffs of the ASTM and WMT. With a total score of 58, her performance on the ASTM was far below the proposed cutoff of 85. The scores on the WMT SVT measures were also below the cutoff of 82.5% (IR = 55%, DR = 42.5%, and CNS = 72.5%). Her scores on the IR and DR measures were in the random response range. She also scored extremely low on the standard tests for attention and executive functioning (i.e., trail-making test) and memory (i.e., visual association test). With a total score of 12, her SIMS did not indicate deviancy. She scored relatively high on the SDE-BIDR, which suggests a tendency to deny psychological factors contributing to experienced symptoms. In accordance with this finding, the scores on a self-report measure for psychological distress (i.e., the symptom checklist-90) and on a self-report measure for depressive symptoms (i.e., the Beck depression inventory) were not heightened. The MMPI-2 detected a valid profile, allowing the clinical scales to be interpreted. The clinical scales 1 and 3 revealed dual elevations (i.e., the 1–3/3–1 configuration).

**Table 2**

*2012 Neuropsychological Test Results*

| Tests | Scores | Percentiles / Interpretation |
|---|---|---|
| **Symptom validity tests** | | |
| WMT | | |
| IR | 55% | Fail |
| DR | 42.5% | Fail |
| CNS | 72.5% | Fail |
| MC | 20% | Warning |
| PA | 15% | Warning |
| FR | 10% | Warning |
| ASTM | 58/90 | Below cut-off |
| SIMS | 12 | Normal |
| Amnesia | 5 | |
| Neurologic impairment | 2 | |
| Psychosis | 1 | |
| Affective disorder | 3 | |
| Low intelligence | 1 | |
| **Neuropsychological tests** | | |
| MoCA | 20/30 | Below cut-off |
| VAT | | |
| 1st trial | 3 | 4 |
| 1st + 2nd trial | 7 | 1 |
| Recall | 1 | 2 |
| TMT-A | 48 s | 31s |
| TMT-B | Aborted | - |
| Clock, Bike and House drawing | Incomplete | Deviant |
| **Questionnaires** | | |
| SCL-90 | 126 | Normal |
| BDI | 11 | Normal |
| SDE | 79 | Heightened |
| **Personality inventory** | | |
| MMPI-2 (*T* scores) | | |
| L | 50 | Normal |
| F | 51 | Normal |
| K | 68 | Heightened |
| VRIN | 42 | Normal |
| TRIN | 55F | Normal |
| Fb | 41 | Normal |
| 1 | 86 | High |
| 2 | 75 | Heightened |

| | | |
|---|---|---|
| 3 | 100 | High |
| 4 | 67 | Heightened |
| 5 | 51 | Normal |
| 6 | 52 | Normal |
| 7 | 65 | Normal |
| 8 | 79 | Heightened |
| 9 | 57 | Normal |
| 0 | 37 | Low |

*Note.* WMT = word memory test; IR = immediate recall; DR = delayed recall; CNS = consistency; MC = multiple choice; PA = paired associates; FR = free recall; ASTM = Amsterdam short-term memory test; SIMS = structured inventory of malingered symptoms; MoCA = Montreal cognitive assessment; VAT = Visual Association Test short version; TMT-A and B = Trail Making Test-A and B; SCL-90 = symptom check list; BDI = beck depression inventory; SDE-BIDR = self-deceptive enhancement subscale of the balanced inventory of desirable responding; MMPI-2 = Minnesota multiphasic personality inventory-2; L = lie; F = frequency; K = correction; VRIN = variable response inconsistency scale; TRIN = true response inconsistency scale; Fb = frequency back; 1 = hypochondriasis; 2 = depression; 3 = hysteria; 4 = psychopathic deviate; 5 = masculinity/femininity; 6 = paranoia; 7 = psychasthenia; 8 = schizophrenia; 9 = hypomania; 0 = social introversion.

## DISCUSSION

We presented a case study of a 59-year-old female patient who was reexamined by the departments of neurology (i.e., MRI brain scan) and medical psychology (i.e., neuropsychological assessment) at a general hospital roughly 10 years after her initial MCI diagnosis. The patient scored well below the proposed cutoffs of the ASTM and WMT, which clearly indicates cognitive underperformance. The total SIMS score was not deviant. Closer inspection of the SIMS subscales, however, did reveal an elevated score on the Amnesia subscale (5/15) but not on the Neurologic (i.e., "somatic") Impairment subscale (2/15). The elevated score on the Amnesia subscale and the patient's underperformance on cognitive tests suggests that she was selectively exaggerating cognitive impairment. Furthermore, a striking inconsistency was observed between the low cognitive test scores and both her level of daily functioning and the lack of cognitive problems observed during the interview. Based on the failing of two SVTs and the inconsistencies described, the test scores are considered to be the result of insufficient effort to perform well. Consequently, the low neuropsychological test scores cannot be validly interpreted as evidence of a genuine cognitive impairment. Therefore, we concluded that a diagnosis of cognitive disorder could not be made based on the available data, thereby compromising the MCI diagnosis. In other words, we removed the diagnosis of MCI because the second criterion for this diagnosis (i.e., objective evidence of impairment in one or more cognitive domains) was no longer fulfilled.

Although researchers agree that failing two well-validated SVTs indicates noncredible performance in all but the most extreme cognitively impaired patients (Slick et al., 1999), such failure does not identify the underlying causes of noncredibility. One possibility is that noncredibility is the result of a conscious attempt to perform poorly, motivated by an external incentive (i.e., malingering). The

patient scored in the random response range on the ASTM and on the IR and DR of the WMT, but not significantly below chance. Below-chance performance implies that the patient knew the correct answer on at least several items but chose to give the incorrect answer (Pankratz & Erickson, 1990). Because the patient did not perform below chance on the SVTs, it remains uncertain whether the patient deliberately chose to give the incorrect answer. The patient may have been motivated by the financial disability benefits for which she qualified through her MCI diagnosis. It is therefore possible that this external incentive prompted her poor performance. In fact, during the first reevaluation of the MCI diagnosis in 2009, she explicitly stated that the new examination might result in an extension of her disability benefits. Another possibility is that the self-reported fatigue and pain complaints led to low SVT performance because the effort required for the tests could not be sustained due to these symptoms. Indeed, the patient's medical file showed a long history of medically unexplained symptoms. Prior to the MCI diagnosis, CFS and FS diagnoses were made. However, although clinicians often appear to readily accept explanations such as pain, fatigue, and emotional factors (e.g., the ill-defined concept of "cry for help" or struggle for recognition), there is no evidence substantiating the assumption that SVT failure is caused solely by these symptoms and/or conditions (Boone, 2013; Johnson-Greene et al., 2013; Merten & Merckelbach, 2013).

The feedback session with our patient was based on recommendations of providing feedback on invalid test performance (Carone et al., 2010). In line with these recommendations, non-neurological factors that could interfere with the test results were addressed. Our differential diagnosis consisted of external gain (i.e., disability benefits) and psychological factors (i.e., somatoform disorder) as an explanation of her poor performance. In case of psychological problems, we would recommend psychotherapy. These considerations were addressed during the feedback session. The patient replied that she did not recognize any of these non-neurological explanations for her symptoms. She disagreed with our conclusions and did not give permission to inform the neurologist about our findings. In line with Article III.3.2.14 of the Code of Ethics of the Dutch Institute of Psychologists (Koene, 2007), the patient used her right to block the written report.

In this case, despite repeated indications of symptom distortion and noncredible cognitive test scores, consumption of medical services continued. Although contradictory evidence was abundant, the MCI diagnosis was made twice. Instead of restricting further medicalization, the health care sector provided additional medical examination and treatment, which may have intensified the patient's somatic fixation and has potentially had an iatrogenic and cost-ineffective effect. In the case of neuropsychological examination, attributing abnormal neuropsychological findings and subjective cognitive complaints too readily to cerebral impairment (i.e., by diagnosing the described patient with MCI) might cause this adverse side effect (van der Werf et al., 2000; van Hout et al., 2006).

Admittedly, this case is an extreme example of a patient presenting with noncredible symptoms. Nevertheless, it is important to acknowledge that even in less obvious symptom distortion scenarios, noncredible symptom reports inhibit the valid interpretation of neuropsychological test results, which can potentially lead to false-positive diagnosis. It is possible that professionals primarily engaged in diagnosing or ruling out severe age-related neurological conditions (e.g., dementia in a memory clinic) are less aware of symptom exaggeration. However, recent findings suggest that

noncredible test performance is not negligible in outpatient memory clinics. For example, Rienstra et al. (2013) found that 13% of 76 patients visiting a memory clinic who were younger than 65 years old failed the WMT.

Although noncredible performance during neuropsychological assessment is now recognized as a risk factor for erroneous diagnoses, SVTs are at this point in time not obliged and are not incorporated in the diagnostic make-up for mild cognitive disorders in all memory clinics in The Netherlands. Contradictory to the United States, guidelines for the use of SVTs are lacking for continental Europe (Merten et al., 2013). A recent survey conducted among neuropsychologists in continental Western Europe showed that SVTs were only administered in a minority of neuropsychological assessments and that inference about noncredible symptom reports in most cases is based on subjective judgment, which is known to be inadequate (Dandachi-FitzGerald et al., 2013). As illustrated in the current case study, clinicians' blindness toward noncredible symptom reports can be remediated by the standard incorporation of SVTs into the neuropsychological assessment.

## REFERENCES


Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., Gamst, A., Holtzman, D. M., Jagust, W. J., Petersen, R. C., Snyder, P. J., Carrillo, M. C., Thies, B., & Phelps, C. H. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia, 7*(3), 270–279. https://doi.org/10.1016/j.jalz.2011.03.008

Boone, K. B. (2007). A reconsideration of the Slick et al. (1999) criteria for malingered neurocognitive dysfunction. In K. B. Boone (Ed.), *Assessment of feigned cognitive impairment: A neuropsychological perspective,* p. 29-49. The Guilford Press.

Boone, K. B. (2013). *Clinical practice of forensic neuropsychology: An evidence-based approach*. The Guilford Press.

Bush, S. S., Ruff, R. M, Troster, A. I., Barth, J. T., Koffler, S. P., Pliskin, N. H., Reynolds, C. R., & Silver, C. H. (2005). NAN position paper. Symptom validity assessment: Practice issues and medical necessity. NAN Policy & Planning Committee. *Archives of Clinical Neuropsychology, 20,* 419–426

Carone, D. A., Green, P., & Drane, D. L. (2013). Word memory test profiles in two cases with surgical removal of the left anterior hippocampus and parahippocampal gyrus. *Applied Neuropsychology: Adult, 21,* 155–160. doi:10.1080/09084282.2012.755533

Carone, D. A., Iverson, G. L., Bush, S. S. (2010). A model to approaching and providing feedback to patients regarding invalid test performance in clinical neuropsychological evaluations. *The Clinical Neuropsychologist, 24,* 759–778. doi:10.1080/13854041003712951

Dandachi-FitzGerald, B., Ponds, R. W., & 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, 771–783*. doi:10.1093/arclin/act073

Faust, D. (1995). The Detection of Deception. Special issue: Malingering and conversion reactions. *Neurologic Clinics, 13,* 255-265.

Green, P. (2003). *Green's Word Memory Test for Microsoft Windows: User's manual*. Edmonton, Canada: Green's Publishing.

Han, J. W., Lee, S. B., Kim, T. H., Park, J. H., Lee, J. J., Huh, Y. S., Choi, E. A., Choe, J. Y., Do, Y. J., Lee, D. Y., & Kim, K. W. (2011). Functional impairment in the diagnosis of mild cognitive impairment. *Alzheimer disease and associated disorders*, *25*(3), 225–229. https://doi.org/10.1097/WAD.0b013e318209d517

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

Johnson-Greene, D., Brooks, L. & Ference, T. (2013). Relationship Between Performance Validity Testing, Disability Status, and Somatic Complaints in Patients With Fibromyalgia. *The Clinical Neuropsychologist, 27*, 148-158.

Koene, C. (Ed.). (2007, March 1). *Beroepscode voor Psychologen. [Professional code for psychologists].* Retrieved from http://www.psynip.nl/website-openbaar-documenten-nip-algemeen/beroepscode-voor-psychologen.pdf

Koepsell, T. D. & Monsell, S. E. (2012). Reversion from mild cognitive impairment to normal or near-normal cognition: Risk factors and prognosis. *Neurology, 79,* 1591-1598.

Lenahan, M. E., Klekociuk, S. Z. & Summers, M. J. (2012). Absence of a relationship between subjective memory complaint and objective memory impairment in mild cognitive impairment (MCI): is it time to abandon subjective memory complaint as an MCI diagnostic criterion? *International Psychogeriatrics, 24*(9), 1505-14

Merckelbach, H., Jelicic, M., & Pieters, M. (2011). The residual effect of feigning: How intentional faking may evolve into a less conscious form of symptom reporting. *Journal of Clinical and Experimental Neuropsychology, 33*, 131–139. doi:10.1080/13803395.2010.495055

Merckelbach, H., Koeyvoets, N., Cima, M. & Nijman, H. (2001). De Nederlandse versie van de SIMS [the Dutch version of the SIMS]. *De Psycholoog, 36,* 586-591.

Merckelbach, H., & Smith, G. P. (2003). Diagnostic Accuracy of the Structured Inventory of Malingered Symptomatology (SIMS) in detecting instructed malingering. *Archives of Clinical Neuropsychology, 18*, 145-152.

Merten, T. Dandachi-FitzGerald, B., Hall, V., Schmand, B., Santamaría, P. & González-Ordi, H. (2013). Symptom validity assessment in European countries: Development and state of the art . *Clínica y Salud, 24*, 129-138.

Merten, T. & Merckelbach, H. (2013). Symptom Validity Testing in Somatoform and Dissociative Disorders: A Critical Review. *Psychological Injury and Law, 6*, 122–137.

Mitchell, A. J. (2008). Is it time to separate subjective cognitive complaints from the diagnosis of mild cognitive impairment? *Age and Ageing, 37,* 497–499.

Pankratz, L., & Erickson, R. D. (1990). Two views of malingering. *The Clinical Neuropsychologist, 4,* 379–389.

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49–69). Lawrence Erlbaum Associates.

Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., & Kokmen, E. (1999). Mild cognitive impairment: Clinical characterization and outcome. *Archives of Neurology*, *56*, 303–308.

Rienstra, A., Groot, P. F., Spaan, P. E., Majoie, C. B., Nederveen, A. J., Walstra, G. J., de Jonghe, J. F., van Gool, W. A., Olabarriaga, S. D., Korkhov, V. V., & Schmand, B. (2013). Symptom validity testing in memory clinics: Hippocampal-memory associations and relevance for diagnosing mild cognitive impairment. *Journal of Clinical and Experimental Neuropsychology*, *35*(1), 59–70. https://doi.org/10.1080/13803395.2012.751361

Rienstra, A., Spaan, P. E. J., & Schmand, B. (2010). Validation of symptom validity tests using a 'child-model' of adult cognitive impairments. *Archives of Clinical Neuropsychology, 25*, 371–382. doi:10.1093/arclin/acq035

Schmand, B., de Sterke, S., & Lindeboom, J. (1999). *Amterdamse Korte Termijn Geheugen test: Handleiding [Amsterdam Short-Term Memory Test: A Manual]*. Lisse, The Netherlands: Swets & Zeitlinger.

Schmand, B., & Lindeboom, J. (2005). *Amsterdam Short-Term Memory Test. Manual.* Leiden, The Netherlands: PITS.

Slick, D. J., Sherman, E. M., & Iverson, G. L. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. *The Clinical Neuropsychologist, 13,* 545–561.

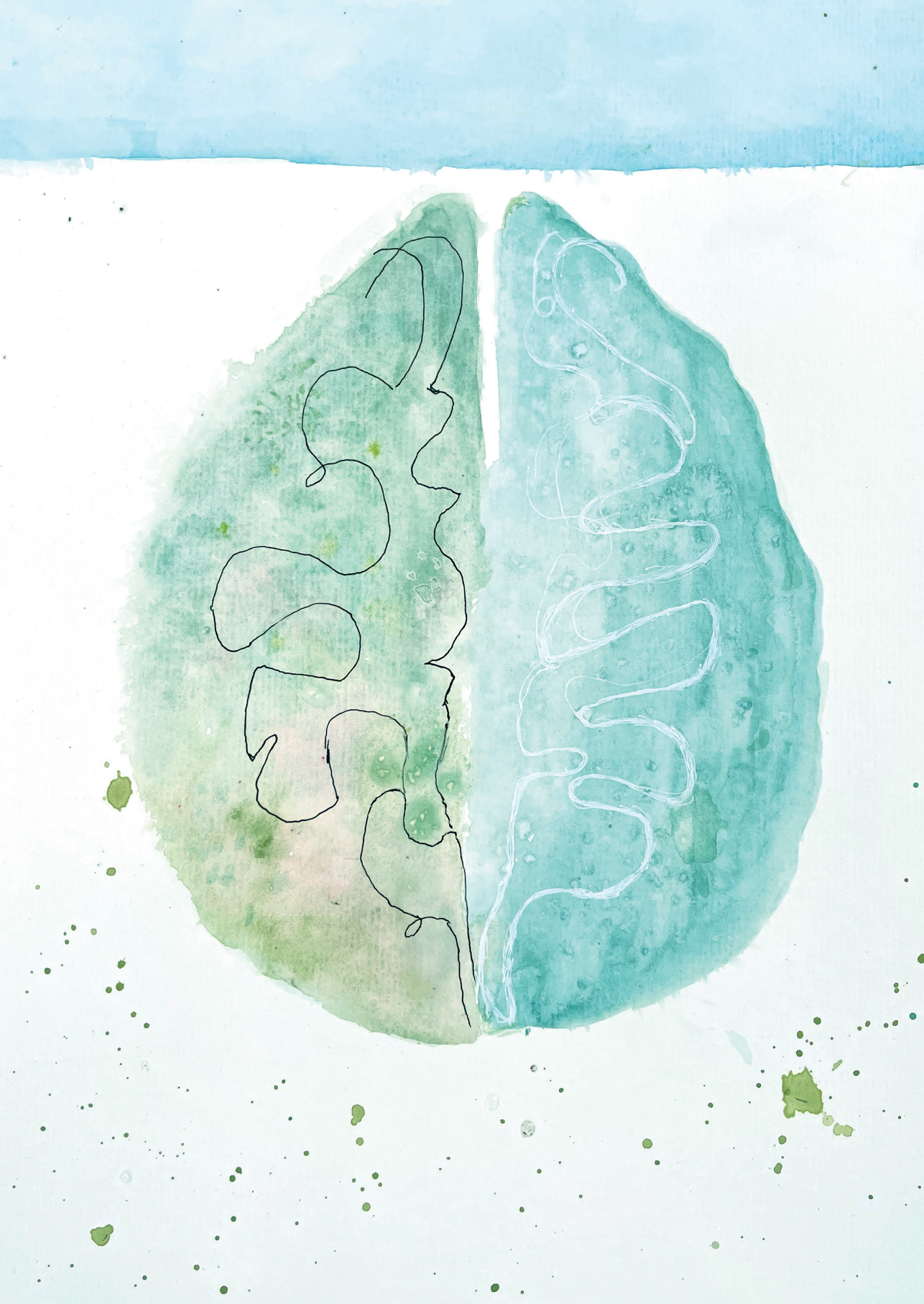
Van der Werf, S. P., Prins, J. B., Jongen, P. J., van der Meer, J. W. & Bleijenberg, G. (2000). Abnormal neuropsychological findings are not necesserely a sign of cerebral impairment: A matched comparison between chronic fatigue syndrome and multiple sclerosis. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology, 13*(3), 199-203.

Van Hout, M. S., Schmand, B., Wekking, E. M., & Deelman, B. G. (2006). Cognitive functioning in patients with suspected chronic toxic encephalopathy: Evidence for neuropsychological disturbances after controlling for insufficient effort. *Journal of Neurology, Neurosurgery and Psychiatry, 77,* 296-303.

Visser, P. J., Kester, A., Jolles, J., & Verhey, F. (2006). Ten year risk of dementia in subjects with mild cognitive impairment. *Neurology*, *67*, 1201–1207.

Visser, P. J. & Verhey, F. R. (2008). Mild cognitive impairment as predictor for Alzheimer's disease in clinical practice: Effect of age and diagnostic criteria. *Psychological Medicine, 38,* 113–122.

Vos, S. J., van Rossum, I. A., Verhey, F., Knol, D. L., Soininen, H., Wahlund, L. O., Hampel, H., Tsolaki, M., Minthon, L., Frisoni, G. B., Froelich, L., Nobili, F., van der Flier, W., Blennow, K., Wolz, R., Scheltens, P., & Visser, P. J. (2013). Prediction of Alzheimer disease in subjects with amnestic and nonamnestic MCI. *Neurology*, *80*(12), 1124–1132. https://doi.org/10.1212/WNL.0b013e318288690c



# CHAPTER 3

## Performance validity test failure in the clinical population: A systematic review and meta-analysis of prevalence rates

Roor, J. J., Peters, M. J., Dandachi-FitzGerald, B., & Ponds, R. W. (2023). Performance validity test failure in the clinical population: A systematic review and meta-analysis of prevalence rates. *Neuropsychology Review*.

## ABSTRACT

Performance validity tests (PVTs) are used to measure the validity of the obtained neuropsychological test data. However, when an individual fails a PVT, the likelihood that failure truly reflects invalid performance (i.e., the positive predictive value) depends on the base rate in the context in which the assessment takes place. Therefore, accurate base rate information is needed to guide interpretation of PVT performance. This systematic review and meta-analysis examined the base rate of PVT failure in the clinical population (PROSPERO number: CRD42020164128). PubMed/MEDLINE, Web of Science, and PsychINFO were searched to identify articles published up to November 5, 2021. Main eligibility criteria were a clinical evaluation context and utilization of stand-alone and well-validated PVTs. Of the 457 articles scrutinized for eligibility, 47 were selected for systematic review and meta-analyses. Pooled base rate of PVT failure for all included studies was 16%, 95% CI [14, 19]. High heterogeneity existed among these studies (Cochran's Q = 697.97, $p < .001$; $I^2 = 91\%$; $\tau^2 = 0.08$). Subgroup analysis indicated that pooled PVT failure rates varied across clinical context, presence of external incentives, clinical diagnosis, and utilized PVT. Our findings can be used for calculating clinically applied statistics (i.e., positive and negative predictive values, and likelihood ratios) to increase the diagnostic accuracy of performance validity determination in clinical evaluation. Future research is necessary with more detailed recruitment procedures and sample descriptions to further improve the accuracy of the base rate of PVT failure in clinical practice.

## INTRODUCTION

Neuropsychological assessment guides diagnostics and treatment in a wide range of clinical conditions (e.g., traumatic brain injury, epilepsy, functional neurological disorder, attention deficit hyperactivity disorder, multiple sclerosis, or mild cognitive impairment). Therefore, it is important that neuropsychological test results accurately represent a patients' actual cognitive abilities. However, personal factors such as a lack of task engagement or malingering can invalidate a patient's test performance (Schroeder & Martin, 2022). When invalid performance is not properly identified, clinicians risk attributing abnormally low scores to cognitive impairment, potentially leading to misdiagnosis and ineffective or even harmful treatments (e.g., Roor et al., 2016; van der Heide et al., 2020). Consequently, performance invalidity is not only relevant to diagnostics, but also extends to treatment efficacy (Roor et al., 2022).

Various tests are available for determining invalid performance on cognitive tests (for an overview, *see* Soble et al., 2022). Performance validity tests (PVTs) can be specifically designed to measure performance validity (i.e., stand-alone PVTs), or empirically derived from standard cognitive tests (i.e., embedded indicators). Overall, the psychometric properties of stand-alone PVTs have been found to be superior in comparison to embedded PVTs (Miele et al., 2012; Soble et al., 2022). Using well-researched stand-alone PVTs, the meta-analyses of Sollman and Berry (2011) found their aggregated mean specificity to be 0.90, with a mean sensitivity of 0.69. This finding is typical for stand-alone PVTs, for which empirical cutoff scores are chosen at a specificity of ≥ 0.90 to minimize the misclassification of a valid cognitive test performance as non-valid (i.e., a maximum 10% false-positive rate).

Importantly, sensitivity and specificity should never be interpreted in isolation from other clinical metrics like base rates (Lange & Lippa, 2017). To determine the positive and negative predictive value of a PVT score, the base rate of the condition (here: performance invalidity) needs to be considered (Richards et al., 2015). Using Bayes' rule, the likelihood that PVT failure is indeed indicative of performance invalidity can be calculated based upon: 1) the base rate of invalid performance in the specific population of that individual; 2) the score of a PVT; 3) the sensitivity; and 4) the specificity of the utilized PVT (Dandachi-FitzGerald & Martin, 2022; Tiemens et al., 2020). Ignoring Bayes' rule potentially leads to overdiagnosis of invalid performance when the base rate of invalidity is low and to underdiagnosis when the base rate is high. Therefore, it is essential that base rate information is available for each PVT in a specific clinical context and, ideally, for specific clinical patient groups (Schroeder et al., 2021a).

Early surveys amongst non-forensic clinical neuropsychologists reported an expectation that only 8% of general clinical referrals would produce invalid test results (Mittenberg et al., 2002). Over the last two decades, research on validity issues in clinical practice increased significantly, and neuropsychologists have become more aware of the need to identify invalid test performance (Merten & Dandachi-FitzGerald, 2022; Sweet et al., 2021). These factors probably contributed to the findings of a nearly double median reported base rate of 15% across clinical contexts and settings in a more recent survey (Martin & Schroeder, 2020). However, there has been a delay in research examining empirically derived bases rates of invalidity in clinical settings.

To address this issue, McWhirter et al. (2020) undertook a systematic review to examine PVT failure in clinical populations. Their main finding was that PVT failure rates were common, exceeding 25% for some PVTs and clinical groups. However, their study has been criticized on several aspects. First,

Kemp and Kapur (2020) mention that McWhirter et al. (2023) did not distinguish between stand-alone and (psychometrically inferior) embedded PVTs. Second, McWhirter et al. (2020) included studies that examined PVT failure in patients with dementia and intellectual disabilities, two groups in which PVTs are strongly discouraged due to unacceptable high false-positive rates when using the standard cutoffs (Larrabee et al., 2020; Lippa, 2018; Merten et al., 2007). Third, studies with ≥ 50% of the patient sample was involved in litigation or seeking welfare benefits were excluded, other types and lower rates of external gain incentives were not characterized. Therefore, external incentives that increase PVT failure rates in patients engaged in standard clinical evaluations (Schroeder et al., 2021b) may have contributed to their reported PVT failure rates. Importantly, McWhirter et al. (2020) summarized data on PVT failure based upon the literature search and data extraction performed by one author, without considering the quality of included studies or calculating a weighted average to get a more precise estimate.

The current meta-analysis is designed to address these gaps to improve the quality of reported PVT failure findings in clinical patient groups. The main aim of the present study is to provide comprehensive information regarding the base rate of PVT failure to facilitate its interpretation in clinical practice. We calculated pooled estimates of the base rate of PVT failure across the type of clinical context, distinct clinical patient groups, the potential for external incentives, and per PVT.

## METHODS

### Search Strategy

This meta-analysis was conducted in accordance with updated Preferred Reporting Items for Systematic Review and Meta-analyses guidelines (PRISMA; Page et al., 2021). A review protocol was registered at inception on PROSPERO (ID: CRD42020164128). The protocol was slightly modified to further improve the quality of included studies. Specifically, only stand-alone PVTs were included that met the restrictive selection criteria per Sollman and Berry (2011), and one additional database was searched. Electronic databases (PubMed/MEDLINE, Web of Science, and PsychINFO) were comprehensively searched using multiple terms for performance validity and neuropsychological assessment (*see* Online Resource 1 for detailed search strategies). Finally, we chose to focus on the base rate of PVT failure, without also addressing its impact on treatment outcome. The final search was conducted on November 5 2021.

### Study Selection

All studies in this systematic review and meta-analysis were performed in a clinical evaluation context of adult patients (18 + years of age), using standard/per manual administration procedure and cutoffs for the five stand-alone performance validity tests (PVTs) from Sollman and Berry (2011). These five PVTs are: the Word Memory Test (WMT; Green, 2003), the Medical Symptom Validity Test (MSVT; Green, 2004), the Test of Memory Malingering (TOMM; Tombaugh, 1996), the Victoria Symptom Validity Test (VSVT; Slick et al., 1997), and the Letter Memory Test (LMT; Inman et al., 1998). Based upon Grote et al. (2000), a higher cutoff was used for the hard items of the VSVT in patients with medically intractable epilepsy. All studies were original, peer-reviewed, and published in English. Studies were excluded if they

examined PVT failure rate in a non-clinical context (i.e., forensic/medico-legal context, data generated for research purposes). Studies that only addressed PVT failure in a sample already selected upon initially passing/failing a PVT were equally excluded (typically known-groups design). Studies performed on patients diagnosed with intellectual disability or dementia were excluded, as well as studies with a small (sub)sample size ($N < 20$). Finally, we chose to exclude studies of Veterans/military personnel since the distinction between clinical and forensic evaluations are difficult to make within the context of the Veterans Affairs (VA) system (Armistead-Jehle & Buican, 2012).

Unique patient samples were ensured by carefully screening for similar samples used in different studies. In case multiple studies examined the same patient sample, data with the largest sample size was included, or, when equal, the most recent paper.



**Data Collection and Extraction**

References resulting from the searches in PubMed/MEDLINE, Web of Science, and PsychINFO were imported into a reference manager (EndNote X8). After automatic duplicate removal, one of the investigators (JR) manually removed the remaining duplicate references. First, a single rater (JR) screened all titles and abstracts for broad suitability and eligibility. Doubtful references were addressed with a second rater (MP). If doubts remained, references were included for full-text scrutinization. Second, two independent raters (MP and JR) reviewed the remaining full-texts based on the mentioned inclusion and exclusion criteria, for which the online systematic review tool Rayyan (Ouzzani et al., 2016) was used. The interrater reliability was substantial (Cohen's $k = 0.63$), and agreement 89.83%. A sizable number of studies failed to clearly state information used for inclusion in the current study, which contributed to the suboptimal agreement between the two independent raters. Therefore, corresponding authors were contacted when additional information was required (e.g., regarding clinical context, utilized PVT cutoff, number of subjects that were provided and failed a PVT, or language/version of the utilized PVT). Non-responders were reminded twice, and if no author response was elicited, studies were excluded. Discrepancies were resolved by discussion with a third and fourth reviewer (BD and RP). Finally, one investigator (JR) extracted relevant information from the included full-text articles, such as setting, sample size, mean age, and utilized PVT(s) according to a standardized data collection form (*see* Online Resource 2).

**Statistical Analyses**

Statistical analysis was performed using MetaXL version 5.3 (www.epigear.com), a freely available add-in for meta-analysis in Microsoft Excel. Independence of effect sizes, a critical assumption in random-effects meta-analyses, was examined by checking if and how many studies used multiple, potentially inter-correlated PVTs from the same patient sample (Cheung, 2019). The frequency of PVT failure from the individual studies were pooled into the meta-analysis using a double-arcsine transformation. Back transformation was performed to report the pooled prevalence rates. We chose to use this transformation method to stabilize variance in the analysis. The double arcsine transformation has been shown to be preferential to logit transformation or no transformation usage in the calculation of pooled prevalence rates (Barendregt et al., 2013). All analyses were performed using the random-effects model since it

allows between study variation of PVT failure. Forest plots were used to visualize the pooled prevalence of PVT failure, with 95% confidence intervals [CIs]. Where possible, subgroup analyses were performed to examine whether the base rate of PVT failure was related to specific clinical contexts, distinct patient groups, utilized PVT, and the consideration of the presence of potential external gain. To further establish the generalizability of our study findings, the consistency across the included studies was assessed using the Cochran's Q-test (Higgins et al., 2003). For the Q-test, a $p$-value < 0.10 was considered to indicate statistically significant heterogeneity between studies. Because the number of included studies impacts the Q-test, we additionally evaluated the inconsistency index $I^2$ (Higgins & Thompson, 2002). An $I^2$ value over 75% would tentatively be classified as a "high" degree of between-study variance (Higgins et al., 2003). Since $I^2$ is a relative measure of heterogeneity and its value depends on the precision of included studies, we also calculated Tau squared ($\tau^2$). This measure quantifies the variance of the true effect sizes underlying our data, with larger values suggesting greater between-study variance (Borenstein et al., 2017).

**Study Quality**

An adapted version of the Prevalence Critical Appraisal Tool of the Joanna Briggs Institute (Munn et al., 2015) was used to rate the quality of all included studies. Amongst the currently available tools, it addresses the most important items related to the methodological quality when determining prevalence (Migliavaca et al., 2020). Three study quality domains were assessed: selection bias (items 1, 2, and 4), sample size/statistics (items 3 and 5), and attrition bias (item 6; *see* Online Resource 3 for a detailed description).

Doi plot and LFK index are relatively new graphical and quantitative methods that were used for detecting publication bias (Furuya-Kanamori et al., 2018). These analyses were also implemented using MetaXL. Contrary to the scatter plot of precision used in a more standard funnel plot to examine publication bias, the Doi plot uses a quantile plot providing a better visual representation of normality (Wilk & Gnanadesikan, 1968). A symmetric inverted funnel is created with a $Z$-score closest to zero at its tip if the trials are not affected by publication bias. The LKF index then quantifies the two areas under the Doi plot. The interpretation is based on the a-priori concern about positive or negative publication bias. Since we were concerned about possible positive publication bias, the LFK > 1 was used consistent with positive publication bias. Even in the case of limited included studies, the LKF index has a better sensitivity over the more standard Egger's test (Furuya-Kanamori et al., 2018).

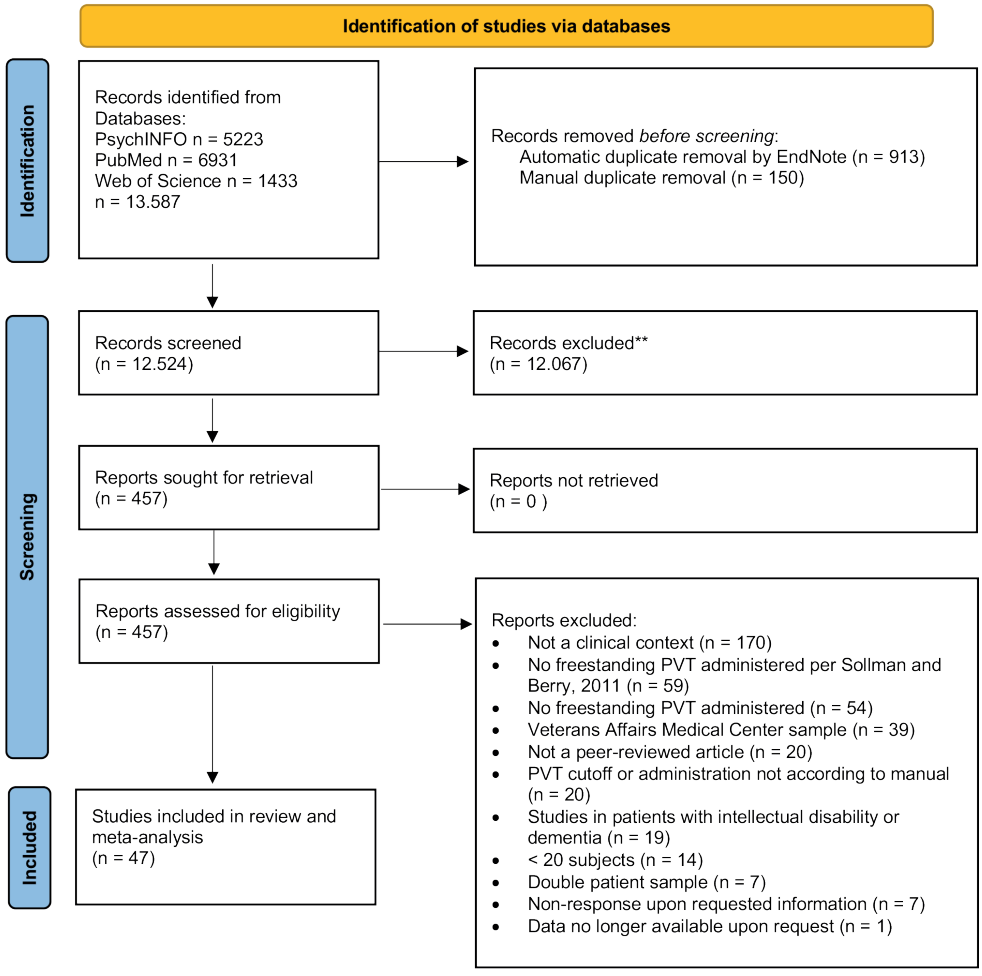
## RESULTS

### Literature Search

Figure 1 gives an overview of the search and selection process. Of the 13,587 identified abstracts, 457 (3.4%) were included for full-text scrutiny. We contacted the first author of 37 studies for additional information, and 30 authors responded. This resulted in 47 observational studies of PVT failure in the clinical context, with a total sample size of *n* = 6,484.

**Figure 1**

*PRISMA Flow Chart of Study Selection*




*Note.* ** No automation tool was used. All records were excluded manually.

**Characterization of Included Studies**

Table 1 reports study characteristics, including clinical context, clinical patient group, and sample size. Most studies were performed in a medical hospital ($k$ = 25), with others in an epilepsy clinic ($k$ = 7), psychiatric institute ($k$ = 6), rehabilitation clinic ($k$ = 4), and private practice ($k$ = 2). Three studies (6.2%) did not specify clinical context. In 15/47 (31.9%) of the studies, prevalence of PVT failure was reported for heterogeneous patient samples. The majority of the studies (32/47; 68.1%) reported PVT failure rates for one or multiple diagnostic subgroups. The diagnostic (sub)groups constituted of patients with traumatic brain injury (TBI) in most studies ($k$ = 10), followed by patients with epilepsy ($k$ = 9), patients with psychogenic non-epileptic seizures (PNES; $k$ = 5), patients that were seen for attention deficit hyperactivity disorder (ADHD) assessment ($k$ = 4), patients with mild cognitive impairment (MCI: $k$ = 4), patients with multiple sclerosis (MS; $k$ = 2), and patients with Parkinson's disease ($k$ = 2). Severity of TBI was not always specified or was poorly defined. The remaining diagnostic (sub)groups (i.e., sickle cell disease, Huntington's disease, patients with substance-use related disorders (SUD), inpatients with depression, memory complaints) were examined in single studies. In more than half of the included studies (25/47; 53.2%), the language (-proficiency) of the included patient sample was not reported. Potential external gain was not mentioned in 12/47 (25.5%) studies, and the remaining studies varied greatly in how they addressed its presence. Of the remaining 35 studies, only seven (i.e., 20%) specified how external gain was examined (Domen et al., 2020; Eichstaedt et al., 2014; Galioto et al., 2020; Grote et al., 2000; Rhoads et al., 2021a; Williamson et al., 2012; Wodushek & Domen, 2020). In most studies (28/35; 80%), the way the authors examined this variable (e.g., by checking the medical record of patient, querying patients about potential incentives being present during the assessment procedure) was not specified. Moreover, in only 4/35 (11.4%) studies, subjects were excluded when external gain incentives (e.g., workers compensation claim) were present (Dandachi-FitzGerald et al., 2020; Davis & Millis, 2014; Merten et al., 2007; Wodushek & Domen, 2020).

The TOMM was the most frequently administered PVT ($k$ = 18), followed by the WMT ($k$ = 17), the MSVT ($k$ = 9), the VSVT ($k$ = 6), or the LMT ($k$ = 1). Only 4/47 (8.5%) studies employed two PVTs (none used > 2 PVTs that fulfilled the inclusion/and exclusion criteria). The other 43/47 (91.5%) studies used one PVT. In two of the four studies reporting two PVTs, the same PVTs were not administered to all participants. Harrison et al. (2021) administered the MSVT to 648 patients and the WMT to 1810 patients, and Krishnan and Donders (2011) administered the TOMM to 39 patients and the WMT to 81 patients. Inclusion in these studies was – amongst others – based upon failing one PVT. Furthermore, these studies did not report the number of subjects that were provided with both PVTs. Therefore, it is unclear to what extend the reported PVT failure rates in these studies are influenced by potential dependence. In the two other studies reporting two PVTs (i.e., Cragar et al., 2006; Merten et al., 2007), all patients were administered both PVTs. The total number of subjects in these two studies that reported two likely dependent effect-sizes was $n$ = 76. This is 1.2% of the total of $n$ = 6487 patients from all 47 studies. We therefore argue that the reported effect sizes from the 47 included studies are (largely) independent.

**Table 1**

*Summary Details for Individual Studies that Reported the Prevalence of PVT Failure in Clinical Patients*

| Study | Type clinical context | Sample type | Sample (*n*) | Mean age (SD) | Mean education (SD) | Country | Language | External incentive | PVT | Administration | PVT failure N [%] * |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Cragar et al. (2006) | Epilepsy clinic | Epilepsy | 41 | 36.0 (9.6) | 11.9 (3.2) | United States | Not mentioned | '48% on or seeking disability'; | TOMM T2 or Retention LMT | Standard | 1 (2.4) |
| | | | | | | | | | | Computerized | 7 (17.1) |
| | | PNES | 21 | 40.8 (10.3) | 13.7 (6.1) | | | '35% on or seeking disability'. Table 1, p. 559 | TOMM T2 or Retention LMT | Standard | 3 (14.3) |
| | | | | | | | | | | Computerized | 5 (23.8) |
| Czornik et al. (2021) | Medical hospital | MCI | 28 | 66.8 (9.9)[a] | 12.5 (4.2)[a] | Austria | Not mentioned | 'Information about possible secondary gain was not available.' p. 272 | WMT IR, DR, or CNS | Computerized German version | 2 (7.1) |
| Dandachi-FitzGerald et al. (2020) | Medical hospital | MCI | 41 | 78.0 (7.2) | [70.7% medium education. Table 1, p. 317] | Netherlands | Dutch | 'Involvement in juridical procedures (e.g., litigation)' as exclusion criterion, p. 315 | TOMM T2 | Dutch version | 3/38 (7.9)[g] |
| | | Parkinson's disease | 41 | 63.7 (8.1) | [56.1% medium education. Table 1, p. 317] | | | | | | 1/40 (2.5)[g] |
| Davis & Millis (2014) | Medical hospital | Heterogeneous neurological | 87 | 42.9 (12.8)[a] | 14.0 (2.3)[a] | United States | English, with '19% of the sample reported history of English as a second language', p. 202 | 'Subjects with potential external incentives were excluded in subgroup analysis.' p. 204 | WMT IR, DR, or CNS | Standard | 12/58 (20.7)[h] |
| Deloria et al. (2021) | Private practice | Heterogeneous | 181 | 58.0 (15.7) | 13.7 (2.5) | United States | Not mentioned | '7.2% of the sample had indication of involvement in disability or litigation claims.' p. 3 | TOMM T2 or Retention | Standard | 7/38 (18.4)[i] |

**Table 1.** Continued

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Dodrill (2008) | Epilepsy clinic | Epilepsy | 65 | 35.2 (12.9) | 12.0 (3.0) | United States | English | Not mentioned | WMT IR, DR, or CNS | Oral version | 16 (25.0) |
| | | PNES | 32 | 42.2 (11.6) | 12.7 (2.4) | | | | | | 9 (28.0) |
| Domen et al. (2020) | Medical hospital | MS | 84[b] | 46.5 (12.9)[b] | 14.9 (2.6)[b] | United States | English | 'Of note, 16.67% of the analyzed sample endorsed currently applying for disability, and this information was unknown for another 29.76%.' p. 512 | MSVT IR, DR, or CNS | Standard | 13/108 (12.0) |
| Donders & Strong (2011) | Rehabilitation clinic | TBI | 100 | 37.5 (13.8)[a] | 13.3 (2.3)[a] | United states | English | '($n$ = 28) were involved in disputed financial compensation seeking at the time of the neuropsychological assessment'. p. 176 | WMT IR, DR, or CNS | Standard | 24 (24.0) |
| Dorociak et al. (2018) | Medical hospital | Sickle cell disease | 54 | 40.6 (12.3) | 13.1 (2.3) | United States | Not mentioned | 'None of the subjects was applying for disability or had other known financial incentives related to cognitive status'. p. 85 | TOMM T2 | Standard | 1 (1.9) |
| Drane et al. (2006) | Epilepsy clinic | Epilepsy | 41[a] | 36.9 (14.4) | 12.6 (2.3) | United States | English | Not mentioned | WMT IR, DR, or CNS | Oral version | 3/37 (8.1) |
| | | PNES | 43[a] | 40.6 (10.2) | 12.4 (2.6) | | | | | | 22 (51.2) |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Eichstaedt et al. (2014) | Medical hospital | Epilepsy | 26 | 37.8 (11.6) | 12.7 (2.6) | United states | English | 'Five participants with LTLE reported receiving disability benefits at the time of evaluation, and none failed the WMT.' p. 947 | WMT IR, DR, or CNS | Standard | 6 (23.1) |
| Erdodi et al. (2018) | Medical hospital | TBI | 104 | 38.8 (16.7) | 13.7 (2.6) | Not mentioned | Not mentioned | 'No data were available on litigation status'. p. 848 | WMT IR, DR, or CNS | Standard | 40 (38.5) |
| Galioto et al. (2020) | Medical hospital | MS | 102 | 47.2 (11.4) | 14.4 (2.6) | United States | English | For MS patients only: 27.9% not seeking disability, 38.5% seeking disability, 33.7% already receiving disability, Table 1, p. 1031 | VSVT hard items[f] | Standard | 15 (14.4) |
| | | Epilepsy | 102 | 47.2 (11.8) | 14.3 (2.5) | | | | | | 6 (5.8) |
| | | mTBI | 50 | 42.7 (13.5) | 14.3 (2.1) | | | | | | 10 (20.4) |
| Gorissen et al. (2005) | Mental healthcare institute | Schizophrenia spectrum | 64 | [Between 18-65. p. 201] | [Less than 6 years of education as exclusion criterion. p. 201] | Netherlands & Spain | Dutch & Spanish | Not mentioned | WMT IR, DR, or CNS | Dutch and Spanish oral versions | 46 (72.0) |
| | | Psychiatric (heterogeneous) | 63 | | | | | | | | 16 (25.0) |
| | | Neurological (heterogeneous) | 20 | | | | | | | | 2 (10.0) |

**Table 1.** Continued

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Grote et al. (2000) | Epilepsy clinic | Epilepsy | 30 | 33.4 (10.6) | 14.0 (2.6) | Not mentioned | Not mentioned | 'They were not seeking compensation at the time of their neuropsychological evaluation. 8 (26.7%) were receiving disability at the time of evaluation because of their seizure disorders.' p. 711 | VSVT hard items | Standard | 0 |
| Haber & Fichtenberg (2006) | Rehabilitation clinic | TBI | 22 | 36.4 (13.9) | 12.2 (1.4) | United States | Not mentioned | 'Subjects were not involved in litigation or workers' compensation cases.' p. 526 | TOMM T2 | Standard | 0 |
| Haggerty et al. (2007) | Medical hospital | Heterogeneous | 300 | 44.7 (13.0) | 13.8 (2.5) | United States | Not mentioned | 'Approximately 16% of the sample was involved in litigation and/or seeking compensation for an illness or injury (e.g., workers' compensation, disability) at the time of their evaluations.' p. 921 | VSVT hard items | Standard | 24 (8.0) |
| Harrison & Armstrong (2020) | Mental healthcare institute | ADHD | 245 | 20.4 (1.8) | [All participants were students. '57.1% in their first or second year.' p. 316] | Canada | Not mentioned | Not mentioned | MSVT IR, DR, or CNS | Standard | 49 (20.0) |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Harrison et al. (2021) | Mental healthcare institute | ADHD | 2463 | 21.8 (5.9) | ['All students were high school graduates or equivalent, with their college or university program in progress.' p. 2] | Canada | Not mentioned | 'All were seeking a diagnosis to allow access to disability supports and services'. p. 3 | MSVT IR, DR, or CNS | Standard | 57/648 (8.8) |
| | | | | | | | | | WMT IR, DR, or CNS | Standard | 206/1810 (11.4) |
| Hoskins et al. (2010) | Epilepsy clinic | Epilepsy | 31 | 38.5 (12.0) | 12.7 (2.7) | United States | English | Not mentioned | WMT IR, DR, or CNS | Oral version ($n = 30$); computerized version ($n = 31$) | 7 (22.6)[j] |
| | | PNES | 30 | | | | | | | | 11 (36.7)[j] |
| Jennette et al. (2021) | Medical hospital | Heterogeneous | 128 | 45.7 (16.4) | 14.0 (2.6) | United States | English,14% bilingual English. Table 1, p. 4 | 14% compensation seeking. Table 1, p. 4 | MSVT IR, DR, or CNS | Standard | 32 (25.0) |
| Keary et al. (2013) | Epilepsy clinic | Epilepsy | 404 | 38.5 (12.1) | 13.3 (2.1) | United States | Not mentioned | 'None of the patients in this clinically referred sample were known to be involved in litigation regarding their medical status or seeking financial compensation at the time of their neuropsychological evaluations.' p. 315 | VSVT hard items[f] | Standard | 22 (5.4) |
| Krishnan & Donders (2011) | Rehabilitation clinic | TBI | 115 | 40.7 (13.3) | 12.9 (2.2) | United States | Not mentioned | 32% seeking financial compensation. Table 1, p. 179 | TOMM T2 or Retention | Standard | 3/39 (8) |
| | | | | | | | | | WMT IR, DR, or CNS | Standard | 25/81 (31) |

**Table 1.** Continued

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Leppma et al. (2018) | Mental health care institute | ADHD | 350 | 22.6 | [22.9% Graduate Students. p. 213] | United States | Not mentioned | Not mentioned | NV-MSVT IR, DR, or CNS | Standard | 68 (21.1) |
| Locke et al. (2008) | Medical hospital | Acquired brain injury | 87 | 36.3 (12.2) | 13.4 (2.5) | United States | Not mentioned | '76% of the sample was on disability at the time of the evaluation.' p. 275 | TOMM T2 | Standard | 19 (21.8) |
| Loring et al. (2007) | Medical hospital | Neurological (heterogeneous) | 27 | 47.4 (13.2) | 13.9 (2.4) | United States | Not mentioned | 'No known external financial incentive'. p. 524 | VSVT hard items | Standard | 2 (7.0) |
| | | Memory complaints | 163 | 51.8 (13.0) | 13.8 (2.6) | | | | | | 16 (10.0) |
| | | TBI | 49 | 36.7 (10.8) | 12.7 (1.9) | | | | | | 6 (12.0) |
| Loring et al. (2005) | Medical hospital | Epilepsy | 120 | 34.5 (11.1) | 12.6 (2.2) | United States | Not mentioned | 'Not actively screened for compensation status.' p. 611 | VSVT hard items[f] | Standard | 14 (11.7) |
| Marshall et al. (2016) | Mental health care institute | ADHD | 428 | 26.4 (7.8) | 14.4 (1.9) | United States | Not mentioned | Not mentioned | WMT IR, DR, or CNS | Standard | 53/174 (30.4)[k] |
| Martins & Martins (2010) | Not specified | MCI | 21 | 71.2 (2.0) | [71.4% had less then 6 years of education. Table 1, p. 178] | Portugal | Portuguese | 'None of these patients had any identifiable secondary gain. All patients were retired and without ongoing legal processes.' p. 178 | WMT IR, DR, or CNS | Portuguese computerized version | 14 (67.0) |
| Merten et al. (2007) | Medical hospital | Heterogeneous | 48 | 56.4 (13.1)[a] | [Minimum of 8 years of formal schooling as inclusion criterion. p. 309][a] | Germany | German | 'Involvement in litigation' as exclusion criterion.' p. 309 | TOMM T2 or Retention | German version | 1/24 (4.2)[c] |
| | | | | | | | | | WMT IR, DR, or CNS | Oral, German version | 2/24 (8.3) |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Meyers et al. (2014) | Not specified | Heterogeneous | 255 | 34.5 (12.1) | 12.5 (1.8) | United states | Not mentioned | '76 [subjects] were in litigation' p. 225 | WMT, IR, DR, or CNS | Standard | 98 (38.4) |
| Moore & Donders (2004) | Rehabilitation clinic | TBI | 132 | 35.8 (14.2) | 12.3 (2.6) | United States | Not mentioned | 'Those seeking financial compensation (*n* = 26) were not excluded.' p. 977 | TOMM T2 | Standard | 11 (8.3) |
| Neale et al. (2022) | Medical hospital | Heterogeneous | 147 | 46.4 (14.5) | 13.2 (2.2) | United States | Not mentioned | 20/147 (13.6%) were seeking compensation. Table 2, p. 5 | MSVT IR, DR, or CNS | Standard | 30/145 (20.7)[l] |
| Rees et al. (2001) | Medical hospital | Inpatients with depression | 26 | 40.4 (11.2) | 4.9 (2.8) | Canada | English | Not mentioned | TOMM T2 or Retention | Standard | 0 |
| | | TBI | 24 | 40.4 (14) | 13.6 (2.7) | | | | | | 0 |
| Resch et al. (2021) | Not specified | Heterogeneous | 88 | 31.7 (10.2) | 15.4 (2.3) | United States | English | Not mentioned | TOMM T2 | Standard | 10/33 (30.0) |
| Rhoads et al. (2021a) | Medical hospital | Heterogeneous | 132 | 45.1 (16.3) | 14.0 (2.6) | United States | English | 'Finally, 15% of patients (*n* = 20) reported being concurrently compensation-seeking (e.g., disability) at the time of their clinical evaluation' p. 135-136 | MSVT IR, DR, or CNS | Standard | 28 (21.2)[m] |
| Rhoads et al. (2021b) | Medical hospital | Heterogeneous | 112 | 60.6 (15.9) | 8.1 (4.5) | United States | Spanish | *n* = 20 (17.9%) compensation seeking. Table 3, p. 272 | TOMM T2 | Spanish version | 20/86 (23.3) |
| Sabelli et al. (2021) | Private practice | mTBI | 326 | 39.5 (11.8) | 12.1 (2.6) | Canada | Not mentioned | Not mentioned | WMT IR, DR, or CNS | Standard | 104 (31.9) |

**Table 1.** Continued

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Schroeder et al. (2019) | Medical hospital | Heterogeneous | 162 | 46.4 (13.2) | 13.6 (2.3) | United States | Not mentioned | 'Roughly 65% of the sample had known or suspected secondary gain associated with the evaluation. The secondary gain was most commonly related to a pursuit of: disability, civil litigation, or workers compensation', p. 468 | TOMM T2 or Retention | Standard | 25 (15.0) |
| Sharland et al. (2018) | Medical hospital | Heterogeneous | 615 | 43.4 (12.8) | 12.6 (2.6) | United States | English | 'Additionally, while the sample was clinical in nature, it is possible a proportion of participants were also involved in litigation, applying for disability, or on workers compensation'. p.105 | TOMM T2 or Retention | Standard | 49 (7.9)[n] |
| Sieck et al. (2013) | Medical hospital | Huntington disease | 36 | 46.1 (12.6) | 13.6 (2.2) | United States | Not mentioned | Not mentioned | TOMM T2 | Standard | 3 (8.2) |
| Silverberg et al. (2017) | Medical hospital | mTBI | 80 | 40.8 (12.0) | [*n* = 37 (46.3%) with postsecondary degree. Table 1, p. 2143] | United States | English | *n* = 71 (86.4%) receiving or seeking injury compensation. Table 1, p. 2143 | MSVT IR, DR, or CNS | Standard | 20 (25.0) |
| Teichner & Wagner (2004) | Medical hospital | Cognitive impairment, not demented | 36 | 70.6 (8.1) | 14.2 (3.2) | United States | Not mentioned | Not mentioned | TOMM T2 or Retention | Standard | 3 (8.3) |
| | | Cognitively intact | 21 | 65.6 (8.6) | 14.2 (3.6) | | | | | | 0 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Vilar-López et al. (2021) | Mental healthcare institute | Substance-use related disorders (SUD) | 77 | 43.2 (8.2)[d] | [patients with primary schooling (*n* = 35) constituted the largest category. Table 1, p. 257] | Spain | Spanish | ' The second group, made up of SUD patients with compensation seeking (*n* = 36), completed a neuropsychological evaluation in order to apply for economic compensation due to their disability (according to their disability level, participants could obtain a monthly payment for life or a payment reviewable at 4 years).' p. 256-257 | TOMM T2 and Retention | Spanish version | 1 (1.3) |
| Walter et al. (2014) | Medical hospital | MCI | 31 | 66.0 (8.0) | 14.7 (2.1) | United states | Not mentioned | 'No participant was involved in litigation at the time of the evaluation or had a substantial external incentive to perform poorly.' p. 1200 | TOMM T2 | Standard | 3 (9.7) |
| Wodushek & Domen, (2020) | Medical hospital | Parkinson's disease | 55 | 65.2 (8.9) | 14.9 (2.8) | United States | English | 'Cases were excluded from analysis if the patient reported being involved in litigation, or if there was an obvious external or secondary gain issue, such as a disability application (*n* = 1).' p. 11 | MSVT IR, DR, or CNS | Standard | 5/51 (9.8) |

**Table 1.** Continued

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Williamson et al. (2012) | Epilepsy clinic | PNES | 90 | 39.0 (8.3)[e] | 13.2 (2.0)[e] | United States | English | 'The presence of financial incentives was determined on the basis of patient report. Patients were classified as having financial incentives if they were currently receiving or applying for disability benefits or other forms of financial compensation (e.g., worker's compensation)'. p. 591 | WMT IR, DR, or CNS | Oral version | 32 (35.5) |

*Note.* ADHD = attention deficit hyperactivity disorder; CNS = consistency score; DR = delayed recognition; IR = immediate recognition; LMT = letter memory test; MCI = mild cognitive impairment; MS = multiple sclerosis; (m)TBI = (mild) traumatic brain injury; (nv)MSVT = (non-verbal) medical symptom validity test; PNES = psychogenic non-epileptic seizures; T2 = trial 2; TOMM = test of memory malingering; VSVT = Victoria symptom validity test; WMT = word memory test.

[*] In case not all subjects in the patient sample received a given PVT, the proportion is shown in this column (.../...).

[a] The authors provided demographics for the total patient sample, not for the (sub)samples for which PVT failure rates were reported.

[b] The authors only provided demographics for the final sample, after excluding cases with missing data. The initial total sample that was provided with a PVT was larger than the sample size detailed in the article.

[c] Results of the "no clinically obvious cognitive impairment" subgroup ($n$ = 24).

[d] Demographics for the total sample ($n$ = 77) were not mentioned. Therefore, we choose to display the demographics of the non-compensation subgroup ($n$ = 41).

[e] Demographics for the total sample ($n$ = 90) were not mentioned. Therefore, we choose to display the demographics of the WMT fail subgroup ($n$ = 32).

[f] VSVT hard items cutoff per Grote et al., 2000 in epilepsy (sub)sample.

[g] B. Dandachi-Fitzgerald (personal communication, February 4, 2022)

[h] J. Davis (personal communication, February 26, 2021)

[i] A. Kivisto (personal communication, February 9, 2022)

[j] D. Drane (personal communication, June 30, 2021)

[k] P. Marshall (personal communication, February 26, 2021)

[l] A. Neale (personal communication, February 5, 2022)

[m] T. Rhoads (personal communication, December 4, 2021)

[n] M. Sharland (personal communication, February 4, 2022)

**Methodological Quality Assessment**

A summary of the methodological quality of the included studies for determining prevalence is provided in Online Resource 4. No study was rated as having high quality; all had limitations in at least one of the three prespecified domains (selection bias, attrition bias, and sample size/ statistical analyses). Most studies had a study sample that addressed the target population ($k$ = 41, 87.2%), whereas only a minority described relevant assessment and patient characteristics ($n$ = 15, 31.9%). The majority of included studies failed to clearly state how patients were recruited ($n$ = 27, 57.4%). Eleven studies (23.4%) had an inadequate response rate. The majority of the studies used appropriate statistical analyses ($n$ = 41, 87.2%), but also had inappropriate sample sizes ($n$ = 39, 83.0%).

The shape of the Doi plot showed slight asymmetry (*see* Online Resource 5), and the results of the LFK index (1.09) revealed minor asymmetry indicative of potential positive publication bias.

**Base Rate of PVT Failure in Clinical Patients**

The pooled prevalence of PVT failure of all ($n$ = 47) included studies was 16%, 95% CI [14, 19]. Significant between-study heterogeneity and high between-study variability existed (Cochran's Q = 697.97, $p$ < 0.001; $I^2$ = 91%; $\tau^2$ = 0.08) as revealed by the large 95% CIs (*see* Figure 2). The high $I^2$ statistic indicates that the variation in reported PVT failure is likely a result of true heterogeneity rather than chance.

**Subgroup Analyses Based upon Clinically Relevant Characteristics**

To facilitate the interpretation of PVT failure in clinical practice, subgroup analyses were performed for clinically relevant characteristics associated with performance validity (Table 2). It is important to emphasize that some of these findings are based upon relatively small numbers of studies (i.e., $k$ = 2 or 4), potentially impacting the stability if the reported estimates.

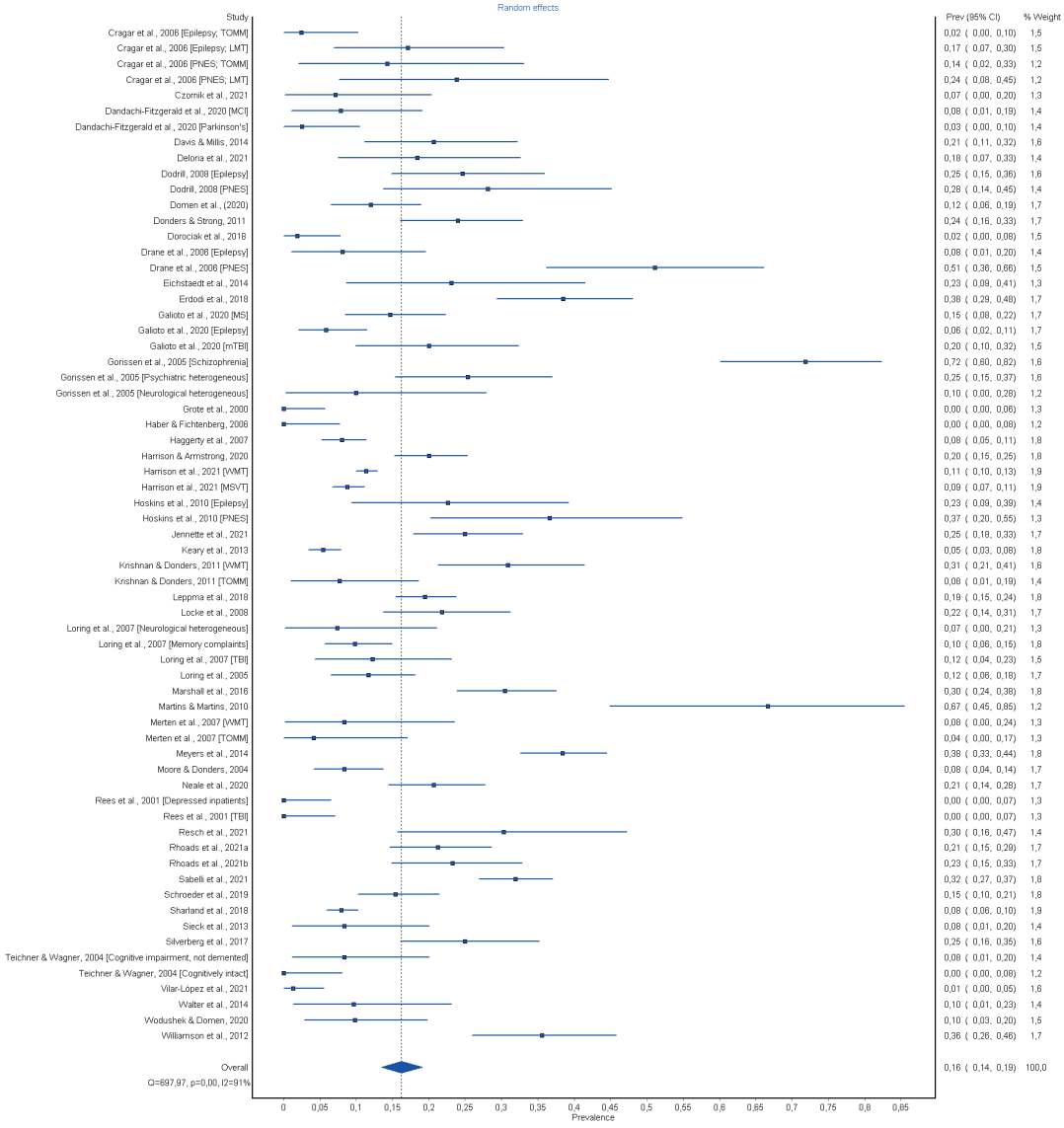
**False-Positive Scrutinization**

Although we excluded studies that examined PVT failure rates in patients with dementia or intellectual disability *a priori*, the included studies might still comprise patient samples with other conditions or combinations of characteristics that make them highly susceptible to false-positive PVT failure classification. Therefore, and in line with clinical guidelines (Sweet et al., 2021), we first examined included studies for the risk of unacceptably low specificity rates when applying standard PVT cutoffs, and two studies were identified. First, PVT performance in the subsample of severely ill schizophrenia spectrum and mostly inpatients from Gorissen et al. (2005) was significantly correlated with negative symptoms and general psychopathology. Second, the MCI subjects from Martins and Martins (2010) were of advanced age, Spanish speaking, and had the lowest formal schooling of all included studies (i.e., 71.4% had less than 6 years of formal education). These cultural/language factors in combination with low formal schooling are associated with unacceptably low specificity rates when applying standard PVT cutoffs (Robles et al., 2015; Ruiz et al., 2020). Exclusion of the subsample of patients with schizophrenia in the Gorissen et al. (2005 study) and of the Martins and Martins (2010) study led to a pooled prevalence of PVT failure of 15% (95% CI [13, 18]; Cochran's Q = 573.73, $p$ < 0.01; $I^2$ = 89%; $\tau^2$ = 0.07). However, after exclusion of these patient samples, between-study heterogeneity and between-study variability were

still high as indicated by a significant Cochran's Q statistic and high and $I^2$ statistic. Further subgroup analyses were performed in the remaining studies ($k = 46$; *see* Table 1).

**Figure 2**

*Forest Plot of the 47 Included Studies Estimating the Pooled Prevalence of PVT Failure in the Clinical Setting*




*Note.* CI = confidence interval; weights are from random effects analysis

**Clinical Context**

The pooled prevalence of PVT failure was the highest in the context of a private practice (27%, 95% CI [15, 40]; Cochran's Q = 2.98, $p = .08$, $I^2 = 66\%$; $\tau^2 = 0.03$), followed by the epilepsy clinic (19%, 95% CI [10, 29]; Cochran's Q = 128.07, $p < .001$, $I^2 = 91\%$; $\tau^2 = 0.17$), the mental healthcare institute (15%, 95% CI [10, 21]; Cochran's Q = 92.30, $p < .001$, $I^2 = 92\%$; $\tau^2 = 0.04$), the medical hospital (12%, 95% CI [10, 15]; Cochran's Q = 160.51, $p < .001$, $I^2 = 81\%$; $\tau^2 = 0.05$) and the rehabilitation clinic (13%, 95% CI [4, 25]; Cochran's Q = 31.07, $p < .001$, $I^2 = 88\%$; $\tau^2 = 0.10$). As can be seen, heterogeneity of pooled PVT failure rates was significant and between-study variability was moderately-high to high for all types of clinical context.

**Clinical Diagnoses**

The pooled prevalence of PVT failure was the highest for patients with PNES (33%, 95% CI [ 24, 43]; Cochran's Q = 10.65, $p = 0.06$; $I^2 = 53\%$; $\tau^2 = 0.03$), followed by subjects seen for ADHD assessment (17%, 95% CI [11, 23]; Cochran's Q = 68.80, $p < 0.01$; $I^2 = 94\%$; $\tau^2 = 0.03$), (m)TBI (17%, 95% CI [10, 25]; Cochran's Q = 89.57, $p < 0.01$; $I^2 = 89\%$; $\tau^2 = 0.09$), MS (13%, 95% CI [9, 18]; Cochran's Q = 0.32, $p = 0.57$; $I^2 = 0\%$; $\tau^2 = 0.00$), epilepsy (11%, 95% CI [6, 16]; Cochran's Q = 42.21, $p < 0.001$; $I^2 = 79\%$; $\tau^2 = 0.05$), MCI (9%, 95% CI [4, 16]; Cochran's Q = 0.11, $p = 0.95$; $I^2 = 0\%$; $\tau^2 = 0.00$), and Parkinson's disease (6%, 95% CI [1, 15]; Cochran's Q = 1.81, $p = 0.18$; $I^2 = 45\%$; $\tau^2 = 0.02$). Based upon Cochran's Q, heterogeneity of pooled PVT failure rates was significant in patients with PNES, (m)TBI, epilepsy, Parkinson's disease, and subjects seen for ADHD assessment. Non-significant heterogeneity in pooled PVT failure rates was found in patients with Parkinson's disease, MCI, and MS. Based upon the $I^2$ statistic, variability of base rate estimates of PVT failure was low in patients with MCI and MS, and (moderately) high for the other diagnostic patient groups. This suggests that for studies in patients with MCI and MS, the pooled PVT failure rates are more homogeneous. However, since these calculations are based upon small numbers of studies, these findings should be interpreted with caution (von Hippel, 2015).

**External Gain Incentives**

In the four studies where patients with potential external gain incentives were excluded from analysis, the pooled prevalence of PVT failure was as low as 10% (95% CI [5, 15]; Cochran's Q = 9.17, $p = 0.10$; $I^2 = 45\%$; $\tau^2 = 0.02$). For the 42 remaining studies that did not report to have actively excluded clinical patients with potential external gain incentives before reporting PVT failure, however, the pooled prevalence of reported PVT failure was 16% (95% CI [13, 19]; Cochran's Q = 560.93, $p < 0.001$; $I^2 = 90\%$; $\tau^2 = 0.07$). Although Cochran's Q statistic indicated that heterogeneity of pooled PVT failure rates in both groups was high, inconsistency was lower in the studies where patients with external gain were excluded from analysis.

**Table 2**

*Pooled Prevalence of PVT Failure in Clinical Patients, Stratified by False-Positive Scrutinization, Clinical Context, External Gain Incentives, Clinical Diagnosis, and PVT*

| Clinical characteristics | *n/k* | Pooled PVT failure rate (%) | 95% CI | $I^2$(%) | $\tau^2$ |
|---|---|---|---|---|---|
| **Overall** | 6,484/47 | 16 | 14-19 | 91 | 0.08 |
| **False-positive scrutinization** | | | | | |
| Probable risk of false positive PVT failure classification | 85/2 | 70 | 60-80 | 0 | 0.00 |
| Probable no risk of false positive PVT failure classification | 6,399/46 | 15 | 13-18 | 89 | 0.07 |
| **Clinical setting*** | | | | | |
| Private practice | 364/2 | 27 | 15-40 | 66 | 0.03 |
| Epilepsy clinic | 824/7 | 19 | 10-29 | 91 | 0.17 |
| Mental healthcare institute | 1,577/6 | 15 | 10-24 | 92 | 0.04 |
| Medical hospital | 3,057/25 | 12 | 10-15 | 81 | 0.05 |
| Rehabilitation clinic | 293/4 | 13 | 4-25 | 88 | 0.10 |
| **Subjects with potential external gain incentives excluded?*** | | | | | |
| Yes | 211/4 | 10 | 5-15 | 45 | 0.02 |
| No | 6,188/42 | 16 | 13-19 | 90 | 0.07 |
| **Clinical diagnosis*** | | | | | |
| PNES | 216/5 | 33 | 24-43 | 53 | 0.03 |
| ADHD | 1,417/4 | 17 | 11-23 | 94 | 0.03 |
| (m)TBI | 926/10 | 17 | 10-25 | 89 | 0.09 |
| MS | 210/2 | 13 | 9-18 | 0 | 0.00 |
| Epilepsy | 856/9 | 11 | 6-16 | 79 | 0.05 |
| MCI | 97/3 | 9 | 4-16 | 0 | 0.00 |
| Parkinson's disease | 91/2 | 6 | 1-15 | 45 | 0.02 |
| **PVT*** | | | | | |
| WMT | 1,482/13 | 25 | 19-32 | 93 | 0.10 |
| (nv)MSVT | 1,891/9 | 18 | 13-23 | 85 | 0.03 |
| TOMM | 1,759/18 | 9 | 6-12 | 80 | 0.05 |
| VSVT | 1,347/6 | 9 | 7-2 | 64 | 0.01 |

*Note.* ADHD = attention deficit hyperactivity disorder; CI = confidence interval; MCI = mild cognitive impairment; MS = multiple sclerosis; (m)TBI = (mild) traumatic brain injury; (nv)MSVT = (non-verbal) medical symptom validity test; PNES = psychogenic non-epileptic seizures; PVT = performance validity test; TOMM = test of memory malingering; VSVT = Victoria symptom validity test; WMT = word memory test. * = after exclusion of the subsamples of subjects with a probable risk of false-positive PVT failure classification (i.e., $k = 46$).

## PVT

The pooled prevalence of PVT failure was the highest for patients examined with the WMT (25%, 95% CI [19, 32]; Cochran's Q = 253.52, $p < 0.001$; $I^2 = 93\%$; $\tau^2 = 0.10$), followed by the (nv)MSVT (18%, 95% CI [13, 23]; Cochran's Q = 55.04, $p < 0.001$, $I^2 = 85$; $\tau^2 = 0.03$), the TOMM (9%, 95% CI [6, 12]; Cochran's Q =

103.35, $p < 0.001$, $I^2 = 80$; $\tau^2 = 0.05$), and the hard items of the VSVT (9%, 95% CI [7, 12]; Cochran's Q = 24.97, $p < 0.001$, $I^2 = 64$; $\tau^2 = 0.01$). Heterogeneity of pooled PVT failure rates was significant across studies examining the same PVT, whereas the between-study variability was moderately-high for studies using the VSVT and high in studies using other PVTs.

## DISCUSSION



This systematic review and meta-analysis examined the prevalence of PVT failure in the context of routine clinical care. Based on extracted data from all 47 studies involving 6,484 patients seen for clinical assessment, the pooled prevalence of PVT failure was 16%, 95% CI [14, 19]. Excluding two studies that likely represented patients where standard PVT cutoff application would probably lead to false positive classification, resulted in a pooled PVT failure of 15%, 95% CI [13, 18]. This number corresponds with the median estimated base rate of invalid performance in clinical settings reported in a recent survey amongst 178 adult-focused neuropsychologists (Martin & Schroeder, 2020). Our empirical findings confirm PVT failure in a sizeable minority of patients seen for clinical neuropsychological assessment.

Another key finding is that reported PVT failure rates vary significantly amongst the included studies (i.e., 0-52.2%). This variability is likely due to (1) sample characteristics, such as clinical setting, clinical diagnosis, and potential external incentives, and (2) the sensitivity and specificity of the PVT used. Pooled PVT failure was found to be highest (i.e., 27%, 95% CI [15, 40]) in patients seen in private practice. The pooled PVT failure rates for the other settings (i.e., epilepsy clinic, mental healthcare institute, medical hospital, and rehabilitation clinic) varied between 13-19%. The Sabelli et al. (2021) study had the largest private practice sample ($N = 326$), consisting of relatively young mTBI patients referred for neuropsychological evaluation. Since only 2/47 of the included studies were conducted in the private practice setting, the Sabelli et al. (2021) study with a PVT failure rate of 31.9%, was a major contributor to the higher pooled PVT failure rate in a private practice setting. Of interest, potential external incentives were not mentioned in that study. Therefore, potential external gain incentives may have been present and impacted the relatively high level of PVT failure –instead of assessment context per se. Unsurprisingly, but now clearly objectified, studies that excluded patients with potential external gain incentives had a significantly lower pooled PVT failure rate compared to studies where these subjects (potentially) remained in the analysis (i.e., 10%, 95% CI [ 5, 15] versus 16%, 95% CI [13, 19], respectively). However, although it is known that the presence of external incentive links directly to PVT failure in clinical assessments (e.g., Schroeder, Clark, & Martin, 2021), little over a quarter of the included studies failed to mention the presence of external gain incentives. Moreover, even when external gain incentives were known to be present, only a minority of studies excluded these subjects from further analyses. Pooled PVT failure rates were highest for patients diagnosed with PNES (i.e., 33%, 95% CI [24, 43]), patients seen for ADHD assessment (i.e., 17%, 95% CI [11, 23]), and (m)TBI (i.e., 17%, 95% CI [10, 25) with pooled PVT failure rates ranging between 6-13% for the other diagnostic groups (i.e., MS, epilepsy, MCI, and Parkinson's disease). These findings contradict the results of McWhirter et al. (2020) that PVT failure in subjects with functional neurological disorders (such as PNES) are no higher compared to MCI or

epilepsy. Likely, our strict in- and exclusion criteria, employment of only well-validated stand-alone PVTs, and meta-analysis application led to a more precise estimate of PVT failure across diagnostic groups. Our findings also indicate that pooled PVT failure rates for MCI, MS, and Parkinson's disease diagnostic groups are more homogeneous than those of PNES, (m)TBI, and patients seen for ADHD assessment. The higher levels of heterogeneity in these latter groups could indicate that other factors that likely impact PVT failure were present, such as external gain incentives, variation in diagnostic criteria, and bias in patient selection. Finally, pooled failure rates mainly varied across the utilized PVTs in line with their respective sensitivity/specificity ratios in correctly identifying invalid performance. The WMT is known for its relatively high sensitivity (Sollman & Berry, 2011), which likely resulted in the highest pooled failure rate amongst the examined stand-alone PVTs. The lowest pooled failure rate for the TOMM is probably related to its high specificity (Martin et al., 2020).

Our findings indicate that, in addition to PVT psychometric properties (i.e., sensitivity and specificity), the clinical setting, the presence of external gain incentives, and the clinical diagnosis impact pooled PVT failure rates. The clinician should therefore consider these factors when interpreting PVT results. Consider, for example, a well-researched stand-alone PVT with a sensitivity of .69 and specificity of .90, administered to two different clinical patients. The first patient is diagnosed with epilepsy and wants to get approved to return to work (i.e., no external gain incentives for invalid performance). If the mentioned PVT were failed in the context of this patient without external gain incentives (base rate PVT failure of 10%, *see* Table 2), the likelihood that PVT failure was indeed a true positive (i.e., positive predictive value, PPV) would be 43%. The second patient is also diagnosed with epilepsy but has a pending disability application because he/she doesn't consider him/herself to be able to return to work (i.e., potential external gain incentive for invalid performance). If the same PVT were failed in the context of this patient with potential external gain incentive (base rate PVT failure of 16%, *see* Table 2), PPV would be 57%.

Of importance, although a PPV increase of .43 to .57 is substantial, the latter is still not sufficient to determine performance validity. Therefore, in line with general consensus multiple, independent, validity tests should be employed (Sherman et al., 20920; Sweet et al., 2021). By chaining the positive likelihood-ratios (LRs) of multiple failed PVTs, the diagnostic probability of invalid performance (or PPV) is increased and the diagnostic error is decreased (for an explanation of how to chain likelihood ratios *see* Larrabee, 2014; Larrabee, 2022). Note that while considerable weight should be placed on the psychometric evaluation of performance validity, the clinician should also include other test and extra-test information (e.g., degree of PVT failure, (in)consistency of the clinical presentation) to draw conclusions about the validity of an individual patient's neuropsychological assessment (Dandachi-FitzGerald & Martin, 2022; Larabee, 2022; Sherman et al., 2020).

Strengths of the present study are its strict inclusion/exclusion criteria ensuring accurate PVT results. Unfortunately, none of the included studies fulfilled all components of the three pre-defined quality criteria selection bias, attrition bias, and adequate sample size/statistics for determining prevalence. Although, we excluded studies with < 20 subjects, most of the remaining studies still had limited sample sizes, increasing the likelihood of sampling bias and heterogeneity. Moreover, only 20/47 studies reported appropriate recruitment method (e.g., consecutive referrals of a good census)

necessary for determining the base rate of PVT failure. Additionally, diagnostic criteria varied across studies, limiting the generalizability of their calculated PVT failure base rates. Also, the way potential external gain incentives were examined and defined varied significantly. Surprisingly, in just over one quarter of included studies, potential external gain incentives were not mentioned at all, and potential external gain incentives may have been present. Finally, although language (-proficiency) and cultural factors relate to PVT failure (Robles et al., 2015; Ruiz et al., 2021), these factors were not mentioned in more than half of the included studies in our meta-analysis.



Additional empirical research is necessary to advance knowledge of performance validity test failure in clinical populations. An important first step in future research should be to provide comprehensive details regarding study design, such as recruitment procedure, clinical setting, and demographic/descriptive information (e.g., cultural factors, age, language and language proficiency, level of education). A second improvement would be to form comparable and homogeneous patient samples by specifying diagnostic criteria and providing a detailed specification of how external gain incentives were examined (e.g., querying the patient for potential external gain incentives, such as pending litigation or disability procedures; Schroeder et al., 2021a). Since administration of multiple PVTs is recommended (Sweet et al., 2021), future studies and specifically meta-analyses should consider using advanced statistical techniques (e.g., three-level meta-analyses) in handling non-independent effect sizes (Cheung, 2019).

In conclusion, the current meta-analysis demonstrates that PVT failure occurs in a substantial minority of patients seen for routine clinical care. Type of clinical context, patient characteristics, presence of external gain incentives, and psychometric properties of the utilized PVT are found to impact the rate of PVT failure. Our findings can be used for calculating clinically applied statistics (i.e., PPV/NPV, and LRs) in everyday practice to increase the diagnostic accuracy of performance validity determination. Future studies using detailed recruitment procedures and sample characteristics, such as external gain incentives and language (proficiency), are needed to further improve and refine knowledge about the base rates of PVT failure in clinical assessments.

## REFERENCES


Armistead-Jehle, P., & Buican, B. (2012). Evaluation context and symptom validity test performances in a U.S. military sample. *Archives of Clinical Neuropsychology*, *27*(8), 828–839. https://doiorg.mu.idm.oclc.org/10.1093/arclin/acs086

Barendregt, J. J., Doi, S. A., Lee, Y. Y., Norman, R. E., & Vos, T. (2013). Meta-analysis of prevalence. *Journal of Epidemiology and Community Health*, *67*(11), 974-978. https://doi.org/10.1136/jech-2013-203104

Borenstein, M., Higgins, J. P., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: $I^2$ is not an absolute measure of heterogeneity. *Research Synthesis Methods*, *8*(1), 5–18. https://doi.org/10.1002/jrsm.1230

Cheung M. W. (2019). A guide to conducting a meta-analysis with non-independent effect sizes. *Neuropsychology Review*, *29*(4), 387–396. https://doi.org/10.1007/s11065-019-09415-6

Cragar, D. E., Berry, D. T., Fakhoury, T. A., Cibula, J. E., & Schmitt, F. A. (2006). Performance of patients with epilepsy or psychogenic non-epileptic seizures on four measures of effort. *The Clinical Neuropsychologist*, *20*(3), 552–566. https://doi.org/10.1080/13854040590947380

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., Duits, A. A., Leentjens, A., Verhey, F., & Ponds, R. (2020). Performance and symptom validity assessment in patients with apathy and cognitive impairment. *Journal of the International Neuropsychological Society*, *26*(3), 314–321. https://doi.org/10.1017/S1355617719001139

Dandachi-FitzGerald, B. & Martin, P. K. (2022). Clinical Judgement and Clinically Applied Statistics: Description, Benefits, and Potential Dangers When Relying on Either One Individually in Clinical Practice. In: Schroeder, R. W. & Martin, P. K. (Eds.). *Validity Assessment in Clinical Neuropsychological Practice; Evaluating and Managing Noncredible Performance* (pp. 107-125). The Guilford Press.

Davis, J. J., & Millis, S. R. (2014). Examination of performance validity test failure in relation to number of tests administered. *The Clinical Neuropsychologist*, *28*(2), 199–214. https://doi.org/10.1080/13854046.2014.884633

Deloria, R., Kivisto, A. J., Swier-Vosnos, A., & Elwood, L. (2021). Optimal per test cutoff scores and combinations of failure on multiple embedded performance validity tests in detecting performance invalidity in a mixed clinical sample. *Applied neuropsychology: Adult*, 1–11. Advance online publication. https://doi.org/10.1080/23279095.2021.1973005

Dodrill C. B. (2008). Do patients with psychogenic nonepileptic seizures produce trustworthy findings on neuropsychological tests? *Epilepsia*, *49*(4), 691–695. https://doi.org/10.1111/j.1528-1167.2007.01457.x

Domen, C. H., Greher, M. R., Hosokawa, P. W., Barnes, S. L., Hoyt, B. D., & Wodushek, T. R. (2020). Are established embedded performance validity test cut-offs generalizable to patients with multiple sclerosis? *Archives of Clinical Neuropsychology*, *35*(5), 511–516. https://doi.org/10.1093/arclin/acaa016

Donders, J., & Strong, C. A. (2011). Embedded effort indicators on the California Verbal Learning Test - Second Edition (CVLT-II): an attempted cross-validation. *The Clinical Neuropsychologist*, *25*(1), 173–184. https://doi.org/10.1080/13854046.2010.536781

Dorociak, K. E., Schulze, E. T., Piper, L. E., Molokie, R. E., & Janecek, J. K. (2018). Performance validity testing in a clinical sample of adults with sickle cell disease. *The Clinical Neuropsychologist*, *32*(1), 81–97. https://doi.org/10.1080/13854046.2017.1339830

Drane, D. L., Williamson, D. J., Stroup, E. S., Holmes, M. D., Jung, M., Koerner, E., Chaytor, N., Wilensky, A. J., & Miller, J. W. (2006). Cognitive impairment is not equal in patients with epileptic and psychogenic nonepileptic seizures. *Epilepsia*, *47*(11), 1879–1886. https://doi.org/10.1111/j.1528-1167.2006.00611.x

Eichstaedt, K. E., Clifton, W. E., Vale, F. L., Benbadis, S. R., Bozorg, A. M., Rodgers-Neame, N. T., & Schoenberg, M. R. (2014). Sensitivity of Green's Word Memory Test genuine memory impairment profile to temporal pathology: a study in patients with temporal lobe epilepsy. *The Clinical Neuropsychologist*, *28*(6), 941–953. https://doi.org/10.1080/13854046.2014.942374

Erdodi, L. A., Abeare, C. A., Medoff, B., Seke, K. R., Sagar, S., & Kirsch, N. L. (2018). A single error is one too many: the forced choice recognition trial of the CVLT-II as a measure of performance validity in adults with TBI. *Archives of Clinical Neuropsychology*, *33*(7), 845–860. https://doi.org/10.1093/acn/acx110

Furuya-Kanamori, L., Barendregt, J. J., & Doi, S. (2018). A new improved graphical and quantitative method for detecting bias in meta-analysis. *International Journal of Evidence-based Healthcare*, *16*(4), 195–203. https://doi.org/10.1097/XEB.0000000000000141

Galioto, R., Dhima, K., Berenholz, O., & Busch, R. (2020). Performance validity testing in multiple sclerosis. *Journal of the International Neuropsychological Society*, *26*(10), 1028–1035. https://doi.org/10.1017/S1355617720000466

Gorissen, M., Sanz, J. C., & Schmand, B. (2005). Effort and cognition in schizophrenia patients. *Schizophrenia Research*, *78*(2-3), 199–208. https://doi.org/10.1016/j.schres.2005.02.016

Green, P. (2003). *Manual for the Word Memory Test.* Edmonton, Alberta, Canada: Green's Publishing.

Green P. (2004). *Manual for the Medical Symptom Validity Test*. Edmonton, Alberta, Canada: Green's Publishing.

Grote, C. L., Kooker, E. K., Garron, D. C., Nyenhuis, D. L., Smith, C. A., & Mattingly, M. L. (2000). Performance of compensation seeking and non-compensation seeking samples on the Victoria symptom validity test: cross-validation and extension of a standardization study. *Journal of Clinical and Experimental Neuropsychology*, *22*(6), 709–719. https://doi.org/10.1076/jcen.22.6.709.958

Haber, A. H., & Fichtenberg, N. L. (2006). Replication of the Test of Memory Malingering (TOMM) in a traumatic brain injury and head trauma sample. *The Clinical Neuropsychologist*, *20*(3), 524–532. https://doi.org/10.1080/13854040590967595

Haggerty, K. A., Frazier, T. W., Busch, R. M., & Naugle, R. I. (2007). Relationships among Victoria symptom validity test indices and personality assessment inventory validity scales in a large clinical sample. *The Clinical Neuropsychologist*, *21*(6), 917–928. https://doi.org/10.1080/13854040600899724

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

Harrison, A. G., Beal, A. L., & Armstrong, I. T. (2021). Predictive value of performance validity testing and symptom validity testing in psychoeducational assessment. *Applied Neuropsychology: Adult*, 1–15. Advance online publication. https://doi.org/10.1080/23279095.2021.1943396

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ (Clinical Research ed.)*, *327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Hoskins, L. L., Binder, L. M., Chaytor, N. S., Williamson, D. J., & Drane, D. L. (2010). Comparison of oral and computerized versions of the word memory test. *Archives of Clinical Neuropsychology*, *25*(7), 591–600. https://doi.org/10.1093/arclin/acq060

Inman, T. H., Vickery, C. D., Berry, D. T., Lamb, D. G., Edwards, C. L., & Smith, G. T. (1998). Development and initial validation of a new procedure for evaluating adequacy of effort given during neuropsychological testing: the letter memory test. *Psychological Assessment*, *10*(2), 128.

Jennette, K. J., Williams, C. P., Resch, Z. J., Ovsiew, G. P., Durkin, N. M., O'Rourke, J. J. F., Marceaux, J. C., Critchfield, E. A., & Soble, J. R. (2022). Assessment of differential neurocognitive performance based on the number of performance validity tests failures: A cross-validation study across multiple mixed clinical samples. *The Clinical Neuropsychologist*, *36*(7), 1915–1932. https://doi.org/10.1080/13854046.2021.1900398

Keary, T. A., Frazier, T. W., Belzile, C. J., Chapin, J. S., Naugle, R. I., Najm, I. M., & Busch, R. M. (2013). Working memory and intelligence are associated with victoria symptom validity test hard item performance in patients with intractable epilepsy. *Journal of the International Neuropsychological Society, 19*(3), 314–323. https://doi.org/10.1017/S1355617712001397

Kemp, S. & Kapur, N. (2020, July 29). Response to McWhirter et al. https://jnnp.bmj.com/content/91/9/945.responses#response-to-mcwhirter-et-al

Krishnan, M., & Donders, J. (2011). Embedded assessment of validity using the continuous visual memory test in patients with traumatic brain injury. *Archives of Clinical Neuropsychology, 26*(3), 176–183. https://doi.org/10.1093/arclin/acr010

Lange, R. T., & Lippa, S. M. (2017). Sensitivity and Specificity Should Never Be Interpreted in Isolation Without Consideration of Other Clinical Utility Metrics. *The Clinical Neuropsychologist*, *31*(6-7), 1015–1028. https://doi-org.mu.idm.oclc.org/10.1080/13854046.2017.1335438

Larrabee, G. J. (2014). Aggregating across multiple indicators improves the detection of malingering: relationship to likelihood-ratios. *The Clinical Neuropsychologist, 22*(4), 666-679, https://doi.org/10.1080/13854040701494987

Larrabee, G. J. (2022). Synthesizing data to reach clinical conclusion regarding validity status. In: Schroeder, R. W. & Martin, P. K. (Eds.). *Validity Assessment in Clinical Neuropsychological Practice; Evaluating and Managing Noncredible Performance* (pp. 193-210). The Guilford Press.

Larrabee, G. J., Boone, K. B., Bianchini, K. J., Rohling, M. L., & Sherman, E. M. (2020, July 29). Response to McWhirter et al (2020). https://jnnp.bmj.com/content/91/9/945.responses#response-to-mcwhirter-et-al

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

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

Locke, D. E., Smigielski, J. S., Powell, M. R., & Stevens, S. R. (2008). Effort issues in post-acute outpatient acquired brain injury rehabilitation seekers. *NeuroRehabilitation*, *23*(3), 273–281.

Loring, D. W., Larrabee, G. J., Lee, G. P., & Meador, K. J. (2007). Victoria symptom validity test performance in a heterogenous clinical sample. *The Clinical Neuropsychologist*, *21*(3), 522–531. https://doi.org/10.1080/13854040600611384

Loring, D. W., Lee, G. P., & Meador, K. J. (2005). Victoria symptom validity test performance in non-litigating epilepsy surgery candidates. *Journal of Clinical and Experimental Neuropsychology*, *27*(5), 610–617. https://doi.org/10.1080/13803390490918471

Marshall, P. S., Hoelzle, J. B., Heyerdahl, D., & Nelson, N. W. (2016). The impact of failing to identify suspect effort in patients undergoing adult attention-deficit/hyperactivity disorder (ADHD) assessment. *Psychological Assessment*, *28*(10), 1290–1302. https://doi.org/10.1037/pas0000247

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.mu.idm.oclc.org/10.1093/arclin/acaa017

Martin, P. K., Schroeder, R. W., Olsen, D. H., Maloy, H., Boettcher, A., Ernst, N. & Okut, H. (2020). A systematic review and meta-analysis of the Test of Memory Malingering in adults: Two decades of deception detection. *The Clinical Neuropsychologist, 34*(1), 88-119, https://doi-org.mu.idm.oclc.org/10.1080/13854046.2019.1637027

Martins, M., & Martins, I. P. (2010). Memory malingering: evaluating WMT criteria. *Applied Neuropsychology*, *17*(3), 177–182. https://doi.org/10.1080/09084281003715709

McWhirter, L., Ritchie, C. W., Stone, J., & Carson, A. (2020). Performance validity test failure in clinical populations-a systematic review. *Journal of Neurology, Neurosurgery, and Psychiatry*, *91*(9), 945–952. https://doi-org.mu.idm.oclc.org/10.1136/jnnp-2020-323776

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

Merten, T., & Dandachi-FitzGerald, B. (2022). Symptom and performance validity assessment: European trends in research and practice. *Psychological Injury and Law, 15*, 113–115. https://doi-org.mu.idm.oclc.org/10.1007/s12207-022-09454-0

Meyers, J. E., Miller, R. M., Thompson, L. M., Scalese, A. M., Allred, B. C., Rupp, Z. W., Dupaix, Z. P., & Junghyun Lee, A. (2014). Using likelihood ratios to detect invalid performance with performance validity measures. *Archives of Clinical Neuropsychology, 29*(3), 224–235. https://doi.org/10.1093/arclin/acu001

Miele, A. S., Gunner, J. H., Lynch, J. K., & McCaffrey, R. J. (2012). Are embedded validity indices equivalent to free-standing symptom validity tests? *Archives of clinical neuropsychology, 27*(1), 10–22. https://doi.org/10.1093/arclin/acr084

Migliavaca, C. B., Stein, C., Colpani, V., Munn, Z., Falavigna, M., & Prevalence Estimates Reviews – Systematic Review Methodology Group (PERSyst) (2020). Quality assessment of prevalence studies: a systematic review. *Journal of Clinical Epidemiology*, *127*, 59–68. https://doi.org/10.1016/j.jclinepi.2020.06.039

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

Moore, B. A., & Donders, J. (2004). Predictors of invalid neuropsychological test performance after traumatic brain injury. *Brain Injury*, *18*(10), 975–984. https://doi.org/10.1080/02699050410001672350

Munn, Z., Moola, S., Lisy, K., Riitano, D., & Tufanaru, C. (2015). Methodological guidance for systematic reviews of observational epidemiological studies reporting prevalence and cumulative incidence data. *International Journal of Evidence-Based Healthcare*, *13*(3), 147–153. https://doi.org/10.1097/XEB.0000000000000054

Neale, A. C., Ovsiew, G. P., Resch, Z. J., & Soble, J. R. (2022). Feigning or forgetfulness: The effect of memory impairment severity on word choice test performance. *The Clinical Neuropsychologist*, *36*(3), 584–599. https://doi.org/10.1080/13854046.2020.1799076

Ouzzani, M., Hammady, H., Fedorowicz, Z. *& Elmagarmid, A. (2016).* Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews, 5*(210). https://doi.org/10.1186/s13643-016-0384-4

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D. et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *The BMJ, 372*(71). doi:10.1136/bmj.n7

Rees, L. M., Tombaugh, T. N., & Boulay, L. (2001). Depression and the Test of Memory Malingering. *Archives of Clinical Neuropsychology*, *16*(5), 501–506.

Resch, Z. J., Soble, J. R., Ovsiew, G. P., Castillo, L. R., Saladino, K. F., DeDios-Stern, S., Schulze, E. T., Song, W., & Pliskin, N. H. (2021). Working memory, processing speed, and memory functioning are minimally predictive of victoria symptom validity test performance. *Assessment*, *28*(6), 1614–1623. https://doi.org/10.1177/1073191120911102

Rhoads, T., Resch, Z. J., Ovsiew, G. P., White, D. J., Abramson, D. A., & Soble, J. R. (2021)[a]. Every second counts: A comparison of four dot counting test scoring procedures for detecting invalid neuropsychological test performance. *Psychological Assessment*, *33*(2), 133–141. https://doi.org/10.1037/pas0000970

Rhoads, T., Leib, S. I., Resch, Z. J., Basurto, K. S., Castillo, L. R., Jennette, K. J., & Soble, J. R. (2021)[b]. Relative rates of invalidity for the test of memory malingering and the dot counting test among Spanish-speaking patients residing in the USA. *Psychological Injury and Law, 14(*4), 269-80. https://doi.org/10.1007/s12207-021-09423-z

Richards, P.M., Geiger, J.A. & Tussey, C.M. (2015). The Dirty Dozen: 12 Sources of Bias in Forensic Neuropsychology with Ways to Mitigate. *Psychological Injury and Law, 8,* 265–280. https://doi-org.ezproxy.ub.unimaas.nl/10.1007/s12207-015-9235-1

Robles, L., López, E., Salazar, X., Boone, K. B., & Glaser, D. F. (2015). Specificity data for the b test, dot counting test, Rey-15 item plus recognition, and Rey word recognition test in monolingual Spanish-speakers. *Journal of Clinical and Experimental Neuropsychology*, *37*(6), 614–621. https://doi.org/10.1080/13803395.2015.1039961

Roor, J. J., Dandachi-FitzGerald, B., & Ponds, R. W. (2016). A case of misdiagnosis of mild cognitive impairment: The utility of symptom validity testing in an outpatient memory clinic. *Applied Neuropsychology: Adult, 23*(3),172-8. doi: 10.1080/23279095.2015.1030018. Epub 2015 Oct 23. PMID: 26496437.

Roor, J. J., Dandachi-FitzGerald, B., Peters, M., Knoop, H., & Ponds, R. W. (2022). Performance validity and outcome of cognitive behavior therapy in patients with chronic fatigue syndrome. *Journal of the International Neuropsychological Society*, *28*(5), 473–482. https://doi-org.mu.idm.oclc.org/10.1017/S1355617721000643

Ruiz, I., Raugh, I. M., Bartolomeo, L. A., & Strauss, G. P. (2020). A meta-analysis of neuropsychological effort test performance in psychotic disorders. *Neuropsychology Review*, *30*(3), 407–424. https://doi.org/10.1007/s11065-020-09448-2

Sabelli, A. G., Messa, I., Giromini, L., Lichtenstein, J. D., May, N., & Erdodi, L. A. (2021). Symptom versus performance validity in patients with mild TBI: independent sources of non-credible responding. *Psychological Injury and Law, 14*(1), 17-36. https://doi.org/10.1007/s12207-021-09400-6

Schroeder, R. W., Boone, K. B., & Larrabee, G. J. (2021). Design Methods in Neuropsychological Performance Validity, Symptom Validity, and Malingering Research. In: Boone, K. B. (Ed.). *Assessment of Feigned Cognitive Impairment, second edition* (pp. 11-33). The 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*, 1–13. Advance online publication. https://doi-org.mu.idm.oclc.org/10.1080/13854046.2021.1895322

Schroeder, R. W., & Martin, P. K. (2022). Explanations of Performance Validity Test Failure in Clinical Settings. In: Schroeder, R. W. & Martin, P. K. (Eds.). *Validity Assessment in Clinical Neuropsychological Practice; Evaluating and Managing Noncredible Performance* (pp. 11-30). The Guilford Press.

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

Sherman, E., 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

Sharland, M. J., Waring, S. C., Johnson, B. P., Taran, A. M., Rusin, T. A., Pattock, A. M., & Palcher, J. A. (2018). Further examination of embedded performance validity indicators for the conners' continuous performance test and brief test of attention in a large outpatient clinical sample. *The Clinical Neuropsychologist*, *32*(1), 98–108. https://doi.org/10.1080/13854046.2017.1332240

Sieck, B. C., Smith, M. M., Duff, K., Paulsen, J. S., & Beglinger, L. J. (2013). Symptom validity test performance in the huntington disease clinic. *Archives of Clinical Neuropsychology*, *28*(2), 135–143. https://doi.org/10.1093/arclin/acs109

Silverberg, N. D., Iverson, G. L., & Panenka, W. (2017). Cogniphobia in mild traumatic brain injury. *Journal of Neurotrauma*, *34*(13), 2141–2146. https://doi.org/10.1089/neu.2016.4719

Slick, D. J., Hopp, G., Strauss, E., & Thompson, G. B. (1997). *Victoria Symptom Validity Test: Professional manual.* Psychological Assessment Resources.

Soble, J. R., Webber, T. A., & Bailey, K. C. (2022). An Overview of Common Performance Validity Tests for Practicing Clinicians. In: Schroeder, R. W. & Martin, P. K. (Eds.). *Validity Assessment in Clinical Neuropsychological Practice; Evaluating and Managing Noncredible Performance* (pp. 126-149). The Guilford Press.

Sollman, M. J., & Berry, D. T. (2011). Detection of inadequate effort on neuropsychological testing: a meta-analytic update and extension. *Archives of Clinical Neuropsychology*, *26*(8), 774–789. https://doi.org/10.1093/arclin/acr066

Sweet, J. J., Heilbronner, 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.mu.idm.oclc.org/10.1080/13854046.2021.1896036

Teichner, G., & Wagner, M. T. (2004). The test of memory malingering (TOMM): normative data from cognitively intact, cognitively impaired, and elderly patients with dementia. *Archives of Clinical Neuropsychology*, *19*(3), 455–464. https://doi.org/10.1016/S0887-6177(03)00078-7

Tiemens B., Wagenvoorde, R. & Witteman, C. (2020). Why Every Clinician Should Know Bayes' Rule. *Health Professions Education, 6*(3), 320-324.

Tombaugh, T. N. (1996). *Test of memory malingering*. Toronto, Canada: MultiHealth Systems.

van der Heide, D., Bošković, I., van Harten, P., & Merckelbach, H. (2020). Overlooking feigning behavior may result in potential harmful treatment interventions: two case reports of undetected malingering. *Journal of Forensic Science, 65*(4),1371-1375. doi: 10.1111/1556-4029.14320. Epub 2020 Mar 23. PMID: 32202670.

Vilar-López, R., Daugherty, J. C., Pérez-García, M., & Piñón-Blanco, A. (2021). A pilot study on the adequacy of the TOMM in detecting invalid performance in patients with substance use disorders. *Journal of Clinical and Experimental Neuropsychology*, *43*(3), 255–263. https://doi.org/10.1080/13803395.2021.1912298

von Hippel, P. T. (2015). The heterogeneity statistic I(2) can be biased in small meta-analyses. *BMC medical research methodology*, *15*, (35). https://doi.org/10.1186/s12874-015-0024-z

Walter, J., Morris, J., Swier-Vosnos, A., & Pliskin, N. (2014). Effects of severity of dementia on a symptom validity measure. *The Clinical Neuropsychologist*, *28*(7), 1197–1208. https://doi.org/10.1080/13854046.2014.960454

Wilk, M. B., & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika*, *55*(1), 1–17.

Williamson, D. J., Holsman, M., Chaytor, N., Miller, J. W., & Drane, D. L. (2012). Abuse, not financial incentive, predicts non-credible cognitive performance in patients with psychogenic non-epileptic seizures. *The Clinical Neuropsychologist*, *26*(4), 588–598. https://doi.org/10.1080/13854046.2012.670266

Wodushek, T. R., & Domen, C. H. (2020). Comparing two models of performance validity assessment in patients with Parkinson's disease who are candidates for deep brain stimulation surgery. *Applied Neuropsychology: Adult*, *27*(1), 9–21. https://doi.org/10.1080/23279095.2018.1473251

## ONLINE RESOURCE 1

**Detailed Search Strategies**

***Search Strategy in PubMed***

((((((((((((malingering[MeSH Terms] OR "symptom validity" OR svt OR effort OR underperformance OR "invalid performance" OR "performance validity" OR pvt))) AND (("neuropsychological tests"[MeSH Terms] OR "cognition disorders"[MeSH Terms] OR neuropsych*))) AND ((dutch[lang] OR german[lang] OR english[lang]))) AND humans[MeSH Terms]) AND adult[MeSH Terms]) NOT child[MeSH Terms])))))))

***Search Strategy in PsychINFO***

S1: malingering OR "symptom validity" OR svt OR effort OR underperformance OR "invalid performance" OR "performance validity" OR pvt

S2: "neuropsychological test*" OR cognition OR "neuropsychological assessment"

Limiters - Published Date: -20211131; Publication Type: Peer Reviewed Journal; Language: Dutch, English, German; Age Groups: Adulthood (18 yrs & older), Young Adulthood (18-29 yrs), Thirties (30-39 yrs), Middle Age (40-64 yrs), Aged (65 yrs & older), Very Old (85 yrs & older); Population Group: Human

***Search Strategy in Web of Science***

((TS=(malingering OR "symptom validity" OR svt OR underperformance OR "invalid performance" OR "performance validity" OR pvt )) AND TS=("neuropsychology*" OR "cognit*" )) NOT TS=("animal*")

[NB: Search term "effort" was dropped here due to too many hits.]

## ONLINE RESOURCE 2

**Data Collection Form**

**Version and date**: 4, 2020

| | |
|---|---|
| Study ID *(surname of first author and year first full report of study was published e.g. Smith 2001)* | |
| Abstract ID *(from endnote library)* | |

**Study Characteristics // Methods**

| | |
|---|---|
| **Type of study** | case-control cross-sectional retrospective cohort |
| **Participants** (N, mean age/SD, level of education, language, country) | |
| **Population description** ***(from which study participants are drawn)*** | |
| **Diagnoses (specify when heterogeneous)** | |
| **Clinical setting** **(evaluation context)** | Medical hospital Mental Health Care Inst.<br>Specialized clinic (eg. epilepsy) Private practice<br>Other |
| **Specify setting 'other'** | |
| **Inclusion criteria** | |
| **Exclusion criteria** | |
| **Confounding variables low IQ <u>and/or</u> severe cognitive impairment mentioned?** | YES NO |

| | |
|---|---|
| **External gain known?** | YES NO |
| **How is external gain known?** | Assumed based on context of assessment (i.e., not in litigation etc.)<br><br>Other |
| **Specify external gain 'other'.** | |

| | |
|---|---|
| **PVT(s)** ***(specify when > 1 PVT used)*** | |
| **Utilized cut-off's** | |
| **Administered in line with manual?** | YES NO (= possible exclusion!)<br><br>Not stated |
| **% PVT failure** ***(specify for every diagnostic and external gain group when possible)*** | |
| **PVTs mean/SD/score-range** | |

| | |
|---|---|
| **Correspondence for further study information** **(fill out when necessary)** | |

## ONLINE RESOURCE 3

**Adapted version of the Joanna Briggs Institute (JBI) Critical Appraisal Checklist for studies Reporting Prevalence Data to rate Study Quality**

1. Was the sample frame appropriate to address the target population?
   The in- and exclusion criteria ensured that the sample frame was appropriate to address the target population (i.e., patients seen for routine care in a clinical context). In case additional diagnostic (sub) groups were examined, the following diagnostic criteria were used to examine adequate diagnostic (sub)group allocation.

   For MCI, we used the diagnostic criteria from Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., & Kokmen, E. (1999). Mild cognitive impairment: Clinical characterization and outcome. *Archives of Neurology, 56*, 303–308.

   For (m)TBI, we used presence of one or more of the following DSM-V criteria: (1) loss of consciousness, (2) posttraumatic amnesia (PTA), (3) disorientation and confusion, and (4) neurological signs (e.g., visual field cuts, or new onset of seizures) (American Psychiatric Association, 2013. Neurocognitive disorders. In *Diagnostic and statistical manual of mental disorders,* 5th ed.

   For epilepsy diagnosis, seizures needed to be confirmed by EEG.

   For PNES, clinical manifestations (e.g., shaking or unresponsiveness) in the absence of EEG abnormalities was used as diagnostic criterion.

   For Parkinson's disease, the Queens Square Brain Bank criteria were used as diagnostic criteria (Lees, A. J., Hardy, J. & Revesz, T., (2009). Parkinson's disease, *The Lancet*, 373(9680), 2055-2066).

2. Were study participants recruited in an appropriate way?
   Sampling has to be clearly stated. Reporting on all data (e.g., consecutive referrals) from a good census will identify everybody and is considered appropriate.

3. Was the sample size adequate?
   Sample size was calculated using this formula: $n = Z^2 P(1\text{-}P)/d^2$, where $n$ = sample size, Z = statistic for 95% level of confidence (1.96), $P$ = expectation prevalence of PVT failure (15%, based upon Martin and Schroeder, 2021), and $d$ = precision (5%). This resulted in a minimal sample size of 196 subjects who were administered a PVT (per subgroup, in case results are displayed per subgroup).

4. Were the study subjects and setting described in detail?
   Since PVT scores are potentially influenced by external gain incentives and language (-proficiency), these clinical variables are considered relevant in describing a target population

5. Was there appropriate statistical analysis?
   The numerator (i.e., number subjects who failed a PVT) and denominator (i.e., total sample size of subjects provided a PVT) should be clearly reported.
   *[Note: This is item 8 of the original format]*

6. Was the response rate adequate, and if not, was the response rate managed appropriately?
   We defined response rate as the number of subjects who were initially described as eligible for inclusion in the study versus the final number of subjects in analyses. In case the initial sample was not fully included in the neuropsychological assessment beyond factors that clearly hamper cognitive functioning (e.g., postictal discharge), the dropouts must be compared with the included sample on relevant variables for PVT moderation (level of education, language-proficiency, and external gain incentives). In case there are no group-differences, the response rate was deemed appropriate.
   [Note: In this item, items 5 and 9 of the original format are combined]

## ONLINE RESOURCE 4

**Table 3**

*Study Quality Using the Joanna Briggs Institute's Critical Appraisal Checklist for Studies Reporting Prevalence Data*

| Question | Cragar, 2006 | Czornik, 2021 | Dandachi-Fitzgerald, 2020 | Davis, 2014 | Deloria, 2021 | Dodrill, 2008 | Domen, 2020 | Donders, 2011 | Dorociak, 2018 | Drane, 2006 | Echstaedt, 2014 | Erdodi, 2018 | Galioto, 2020 | Gorissen, 2005 | Grote, 2000 | Haber, 2006 | Haggerty, 2007 | Harrison, 2020 | Harrison, 2021 | Hoskins, 2010 | Jeannette, 2021 | Keary, 2013 | Krishnan, 2011 | Leppma, 2018 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. Was the sample frame appropriate to address the target population? | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | N | Y | Y | Y | Y | Y | Y | Y | Y | N | Y | Y |
| 2. Were study participants sampled in an appropriate way? | Y | U | U | Y | N | Y | N | Y | Y | Y | N | N | N | N | N | Y | Y | Y | Y | Y | N | Y | Y | N |
| 3. Was the sample size adequate? | N | N | N | N | N | N | N | N | N | N | N | N | N | N | N | N | Y | Y | Y | N | N | Y | N | Y |
| 4. Were the study subjects and the setting described in detail? | N | N | Y | Y | N | N | Y | Y | N | N | Y | N | N | N | N | N | N | N | N | N | Y | N | N | N |
| 5. Was there appropriate statistical analysis? | Y | Y | N | N | N | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y |
| 6. Was the response rate adequate, and if not, was the low response rate managed appropriately? | Y | N | N | N | N | Y | N | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | N | Y | Y | Y | Y | Y |

| Question | Locke, 2008 | Loring, 2007 | Loring, 2005 | Marshall, 2016 | Martins, 2010 | Merten, 2007 | Meyers, 2014 | Moore, 2005 | Neale, 2020 | Rees, 2001 | Resch, 2021 | Rhoads, 2021[a] | Rhoads, 2021[b] | Sabelli, 2021 | Schroeder, 2019 | Sharland, 2018 | Sieck, 2013 | Silverberg, 2017 | Techner, 2004 | Vilar-Lopez, 2021 | Walter, 2014 | Wodushek, 2021 | Williamson, 2012 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. Was the sample frame appropriate to address the target population? | Y | N | N | Y | N | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | Y | U | Y |
| 2. Were study participants sampled in an appropriate way? | Y | N | N | Y | U | U | U | Y | N | N | N | N | N | N | N | Y | U | Y | Y | N | N | N | Y |
| 3. Was the sample size adequate? | N | N | N | N | N | N | Y | N | N | N | N | N | N | Y | N | Y | N | N | N | N | N | N | N |
| 4. Were the study subjects and the setting described in detail? | N | N | N | N | Y | Y | N | N | N | N | N | Y | Y | N | N | Y | N | Y | N | Y | N | Y | Y |
| 5. Was there appropriate statistical analysis? | Y | Y | Y | N | Y | Y | Y | Y | Y | Y | Y | Y | N | Y | Y | N | Y | Y | Y | Y | Y | Y | Y |
| 6. Was the response rate adequate, and if not, was the low response rate managed appropriately? | Y | Y | Y | U | Y | Y | Y | Y | Y | Y | U | Y | N | Y | Y | Y | Y | N | Y | Y | Y | N | Y |

*Note:* N = No; U = Unclear; Y = Yes; item 5 = item 8 of the original format; item 6 = item 9 + item 5 of the original format.

## ONLINE RESOURCE 5

### Results of Publication Bias Analyses

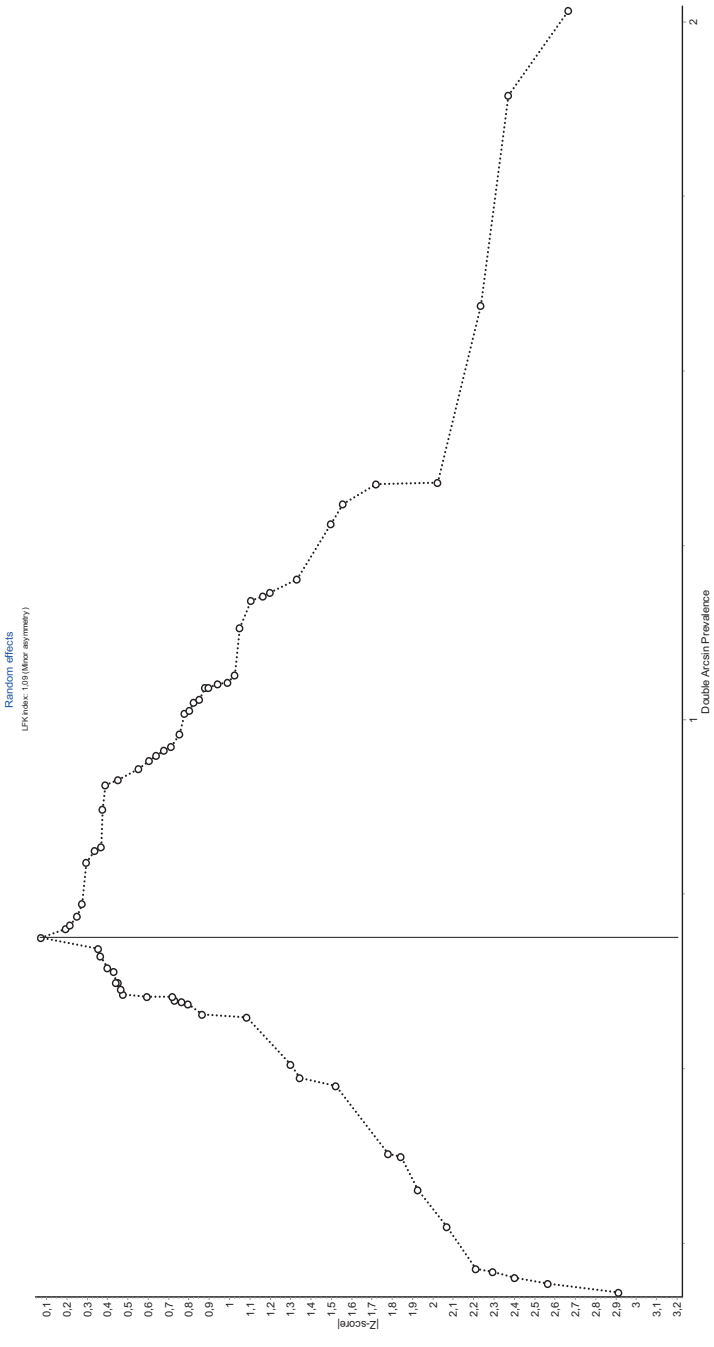
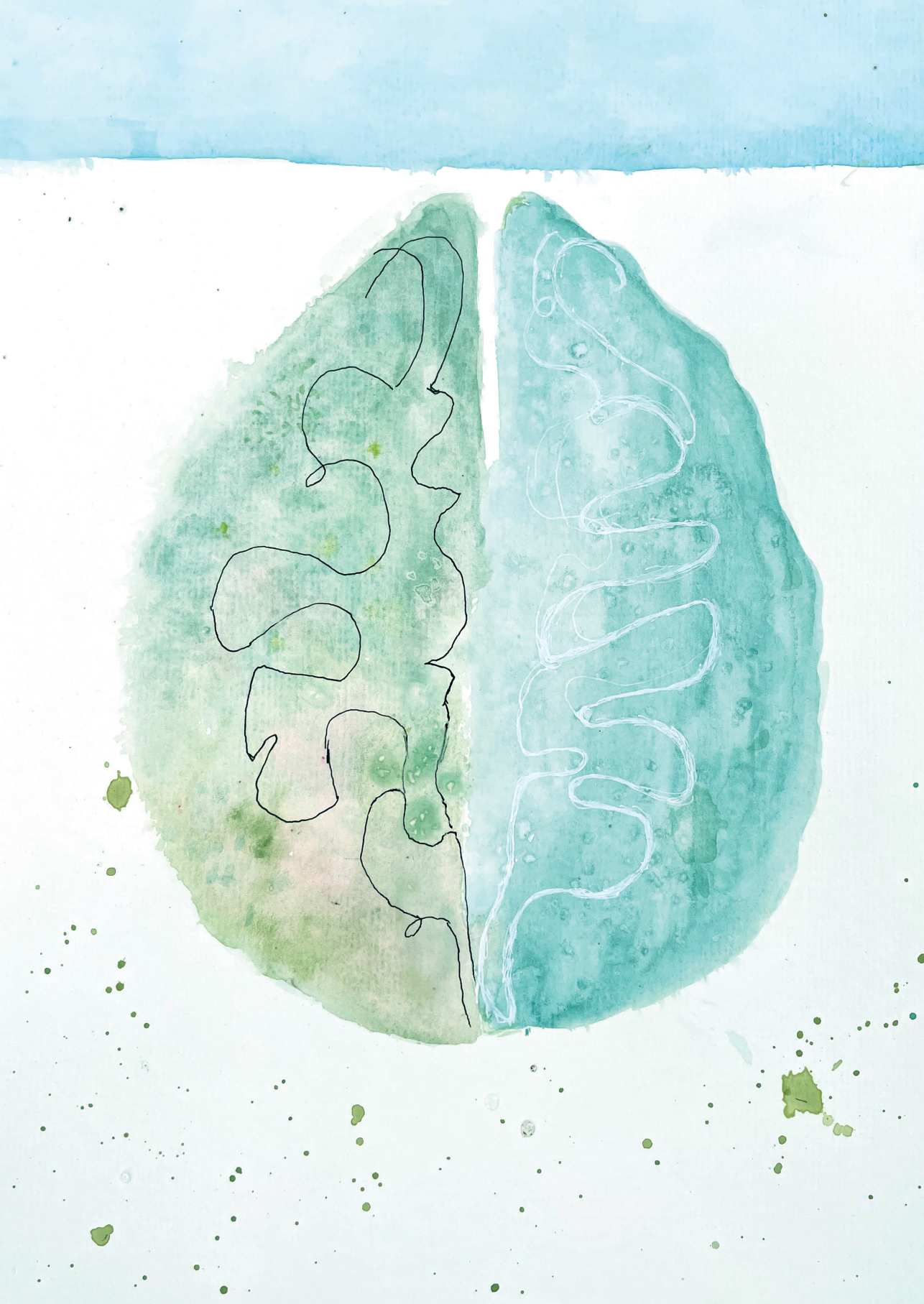



**Figure 1**
*DOI plot and LKF-index*



# CHAPTER 4

## Feedback on underperformance in patients with chronic fatigue syndrome: The impact on subsequent neuropsychological test performance

Roor, J. J., Knoop., H. Dandachi-FitzGerald, B., Peters, M. J., Bleijenberg, G. & Ponds, R. W. (2020). Feedback on underperformance in patients with chronic fatigue syndrome: The impact on subsequent neuropsychological test performance. *Applied Neuropsychology: Adult, 27*(2), 188-196

## ABSTRACT

Performance validity tests (PVTs) are used to measure the credibility of neuropsychological test results. Until now, however, a minimal amount is known about the effects of feedback upon noncredible results (i.e., underperformance) on subsequent neuropsychological test performance. The purpose of this study was to investigate the effects of feedback on underperformance in chronic fatigue syndrome (CFS) patients. A subset of these patients received feedback on Amsterdam Short-Term Memory test (ASTM) failure (i.e., feedback [FB] group). After matching, the final sample consisted of two comparable groups (i.e., FB and No FB; both $n = 33$). At baseline and follow-up assessment, the patients completed the ASTM and two measurements of information processing speed (Complex Reaction Time [CRT] and Symbol Digit Test [SDT]). Results indicated that the patients in the FB group improved significantly on the CRT, compared to the No FB group. Although not significant, a comparable trend-like effect was observed for the SDT. Independent of the feedback intervention there was a substantial improvement on ASTM performance at re-administration. A limited feedback intervention upon underperformance in CFS patients may result in improvement on information processing speed performance. This implies that such an intervention might be clinically relevant, since it maximizes the potential of examining the patients' actual level of cognitive abilities.

## INTRODUCTION

Chronic fatigue syndrome (CFS) is defined by a severe and medically unexplained fatigue persisting for six months or more and leading to a substantial reduction in activities. Concentration and memory complaints are among the eight additional symptoms (Fukada, et al., 1994; Reeves et al., 2003). Subjective cognitive complaints are reported by up to 89% of patients with CFS (Jason et al., 1999). These complaints are significantly related to social and occupational dysfunction (Christodoulou et al., 1998) and to the level of fatigue (Capuron et al., 2006).

Although subjective cognitive complaints are highly prevalent, only a subgroup of CFS patients shows impaired performance on neuropsychological testing (Cockshell & Mathias, 2014; Knoop et al., 2007). In their meta-analysis, Cockshell and Mathias (2010) reported inconsistent and even contradictory findings about cognitive performance in CFS patients. These authors found most evidence for cognitive impairments in the domains of information processing speed, and attention. Additionally, fatigue and depressive symptoms did not entirely account for the variance in cognitive test performance (Cockshell & Mathias, 2010). Until now, the exact extent and nature of reduced neuropsychological test performance in CFS patients is unclear (DeLuca et al., 2010; Cockshell & Mathias, 2014).

One factor that could partially explain low cognitive test scores in CFS is underperformance (Goedendorp et al., 2013; Van der Werf et al., 2000). Underperformance is conceptualized as the extent to which a person's test performance is not an accurate reflection of his or her actual level of cognitive abilities (Larrabee, 2012). Performance validity tests (PVTs) were developed to detect underperformance due to various causes, such as deliberate uncooperativeness for external gain (e.g., obtaining financial benefits) or more psychological causes (e.g., a patient's fear that symptoms are not being recognized) (Boone, 2007; Carone et al., 2010).

Determining the validity of cognitive test results in CFS patients is important for the diagnostic process. More specific, underperformance causes noise in the neuropsychological data. This noise, in turn, potentially clouds the expected brain–behavior relationship that underlies neuropsychological test interpretation (Fox, 2011). As a result, incorrect conclusions about cognitive impairment can be drawn based on invalid data due to underperformance, leading to inadequate diagnoses and treatment (e.g., Roor et al., 2016). For this reason, the standard usage of PVTs to assess for underperformance has been advised by professional organizations (Bush et al., 2005; Heilbronner et al., 2009).

Underperformance in clinical assessments is non-negligible. For example, the prevalence of PVT failure in patients with psychiatric disorder (Dandachi-FitzGerald et al., 2011), ADHD (Marshall et al., 2010), or traumatic brain injury (Krishnan & Donders, 2011) was found to range between 21% and 31%. To date, most studies have shown that investigating underperformance is also relevant to understanding cognitive functioning in patients with CFS. Three studies have found that between 16% and 30% of CFS patients obtained scores indicative of underperformance on a PVT (Goedendorp et al., 2013; Van der Werf et al., 2000; Van der Werf et al., 2002). Other studies have found that the PVT failure rate was low (i.e., 6%; Cockshell & Mathias, 2012) or even zero (Busichio et al., 2004) in this patient group. These inconsistent findings on the prevalence of underperformance in CFS patients could be explained by methodological

differences between studies. That is, the different PVTs that were used and the heterogeneity of the CFS patient samples.

Although the literature has provided guidelines for the determination and classification of underperformance (Bush et al., 2005), a minimal amount of knowledge exists regarding when and how underperformance can best be communicated to the patient. To the best of our knowledge, the study by Suchy and colleagues (2012) is the only study that investigated whether providing patients with feedback on underperformance had an effect on subsequent neuropsychological test performance. These authors conducted a retrospective study in which two groups of patients with multiple sclerosis (MS) were compared. In one group, feedback was provided about PVT failure, while the other group did not receive feedback. The feedback intervention resulted in significantly improved PVT (i.e., the Victoria Symptom Validity Test, VSVT) scores upon re-administration and better performance on a general memory test (i.e., the Wechsler Memory Scale-III, WMS-III) post-feedback. Although these results are promising, it remains unclear whether these results generalize to other clinical samples.

The aim of this study was to investigate whether the favorable effect of providing feedback on underperformance found by Suchy and colleagues (2012) could also be found in CFS patients.

## METHOD

### Participants

Patients were consecutively referred to the Expert Center for Chronic Fatigue of the Radboud University Medical Center, a tertiary treatment facility for chronic fatigue. Consultants at the outpatient clinic of the Department of Internal Medicine assessed the patients' medical status to decide whether the patients had been sufficiently examined to exclude a medical explanation for their fatigue. If their medical evaluation was deemed insufficient, the patients were seen again for anamnesis, full physical examination, case history evaluation, and laboratory tests, following the national CFS guidelines (Centraal Begeleidings Orgaan [CBO], 2013), which are in accordance with the US Centers for Disease Control (CDC) guidelines (Fukuda et al, 1994; Reeves et al., 2003). The CDC criteria for CFS were used: fatigue had to be present for six months or more, accompanied with four out of the following eight symptoms: sore throat, tender lymph nodes, muscle and joint pain, headaches, sleep disturbance, post-exertional malaise lasting more than 24 hours, and cognitive dysfunction (Fukuda et al, 1994; Reeves et al., 2003). If patients met the CDC criteria for CFS, they were referred to the Expert Centre for Chronic Fatigue. The patients were seen routinely for CFS management purposes (i.e., not as part of a separate research project). All of the patients in this sample sought treatment for CFS (cognitive behavioral therapy, CBT). If patients were engaged in a disability claim, they could not participate in the treatment program.

Patients were included in this study if they scored 35 or higher on the fatigue severity subscale of the Checklist Individual Strength (CIS; Worm-Smeitink et al., 2017), and had a weighted total score ≥ 700 on the Sickness Impact Profile 8 (SIP8; Jacobs et al., 1990). For this study, additional inclusion criteria included: (a) underperformance (i.e., score ≤ 85) on the Amsterdam Short-Term Memory test (ASTM; Schmand & Lindeboom, 2005); (b) Dutch language proficiency; (c) repeated neuropsychological

assessment; and (d) being 18 years or older. Patients who showed psychiatric comorbidity during a clinical interview that could explain the fatigue were excluded. The initial number of patients in the database was 1,382. A total of 331 CFS patients (23.9%) underperformed (i.e., ASTM ≤ 85). This percentage of underperformance was in accordance with the prevalence found in previous studies of CFS patients (e.g., Van der Werf et al., 2002). The data were collected between July 2004 and July 2012. During the inclusion period, the policy was changed in that since July 2007 feedback was given on PVT failure. Of the patients who fulfilled the mentioned criteria, 103 were provided with feedback (FB group), and 33 were not provided with feedback (No FB group). After matching (*see* Procedures section) the final sample consisted of two comparable groups (i.e., FB and No FB; both $n = 33$). Table 1 displays the demographical and clinical characteristics of the FB and No FB groups.

The medical ethics committee of the Radboud University Medical Center approved this study.



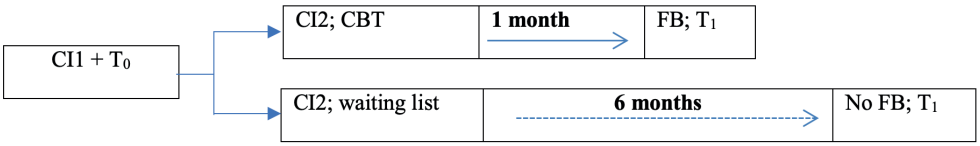
**Procedures**

The psychological assessment part of routine clinical care was conducted before starting with CBT. The following actions were conducted successively (*see* Figure 1). First, one of the psychologists met with the patient for a clinical interview. Then, tests and questionnaires (*see* Instruments) were administered by a trained test assistant. A PVT (i.e., ASTM) was administered at the beginning of this test session, followed by the Symbol Digit Test (SDT) and the Complex Reaction Time task (CRT). During most of the inclusion period of this study, the policy was that all patients with indications for CBT and who failed the ASTM received a feedback intervention before the second assessment, which occurred approximately one week before the start of CBT. The goal of this intervention was to try to attempt to positively influence subsequent psychological assessments and treatment. During this feedback session, the psychologist: (a) addressed that the CFS symptoms of the patient were difficult to evaluate because of the lower-than-expected test performance on a previously administered test; (b) emphasized exerting the patient's best effort and that improvement was expected and (c), explained that therefore tests needed to be repeated (FB group). The remaining subjects in this sample did not receive feedback on the initial PVT failure (No FB group). Therefore, group assignment was conducted on the basis of naturalistic changes in clinical procedures, which were not dependent on patient characteristics. At the second assessment, the ASTM, SDT and CRT were re-administered (i.e., ASTM 2, SDT 2 and CRT 2). The two groups, FB group and No FB group, differed in the meantime interval between the first and second administration. On average, this was one month for the FB group and six months for the No FB group. The patients in the No FB group were placed on a waiting list due to limited treatment capacity for CBT, hence the longer time before re-assessment occurred and treatment was started.

At baseline, we found that the two groups only differed in mean ASTM score (Mann-Whitney *U* test, $p = .003$). Therefore, 33 participants were selected from the FB group ($n = 103$) by matching them on their ASTM score at baseline with the No FB group - additionally to age, sex, and level of education - blind for other scores at baseline and re-assessment (i.e., neuropsychological test and questionnaire scores at T0 and T1).

**Figure 1**

*Timeline Study Procedure*




*Note.* CI1 = clinical interview 1; $T_0$ = neuropsychological test administration baseline; CI2 = clinical interview 2 (randomization to CBT or waiting list); $T_1$ = neuropsychological test administration follow-up.

## Instruments

The Amsterdam Short-Term Memory test (ASTM; Schmand & Lindeboom, 2005) was used to measure underperformance. The ASTM is a forced-choice verbal recognition task, presented to subjects as a memory test. Please refer to the ASTM manual for additional details regarding materials and test procedures (Schmand & Lindeboom, 2005). The ASTM has been thoroughly validated in 17 normative ($n$ = 222) and clinical patient groups ($n$ = 1281) (Schmand & Lindeboom, 2005). Based upon these studies, its internal consistency is excellent (Cronbach's α = 0.91). The test-retest effect of the ASTM was examined in patients with: (1) documented brain damage/disease (e.g., Korsakoff's syndrome, traumatic brain injury, etc.); and (2) no external incentives. Stability in test behavior and hence in performance in these patients was expected. After a time interval of one to three days, Pearson's correlation between the first and second administrations of the ASTM was .91 (Schmand & Lindeboom, 2005). In the validation studies, a cutoff score of 86 was associated with a specificity of 83% and a sensitivity of 92% (Schmand & Lindeboom, 2005).

The Complex Reaction Time task (CRT; Vercoulen et al., 1998) is a reaction time test described in detail in previous studies of CFS patients (e.g., Goedendorp et al., 2013). It was used as a measurement of information processing speed. This test is comprised of three consecutive tasks consisting of 30 trials each. For the purposes of this study, a mean reaction time compound score of the three consecutive tasks was calculated and used for further analysis.

The second performance test was the Symbol Digit Test (SDT) of the Dutch version of the Wechsler Adult Intelligence Scale (WAIS; Stinissen, Willems, Coetsier, & Hulsman, 1970). Here, symbols are presented and must be decoded based on a key translating nine symbols into nine corresponding digits within a 90 second timeframe. This test taps mainly into information processing speed but also attention. Previous research has shown that CFS patients had significantly slower information processing speed compared to healthy controls, based upon the SDT and CRT (Vercoulen et al., 1998).

The revised version of the Symptom Checklist (SCL-90-R; Derogatis, 1994) was designed to measure psychological and somatic symptoms over the previous seven days. In this study, the total score was used as a general measurement of psychological and somatic symptoms. The Dutch version of the SCL-90-R provides extensive normative data (Arrindell & Ettema, 2005).

The fatigue subscale of the Checklist Individual Strength (CIS Fatigue) was used to indicate the level of fatigue over the previous two weeks. The subscale consists of eight items. The total scores range

between 8, indicating no fatigue, and 56, indicating severe fatigue. Scores of ≥ 35 on the CIS Fatigue are indicative of severe fatigue. The CIS has been extensively validated for the assessment of fatigue (Worm-Smeitink et al., 2017), and it is sensitive for detecting changes in fatigue severity (Knoop et al., 2008).

Impairments in daily functioning were assessed with the Sickness Impact Profile 8 total score (SIP8 Total; Jacobs et al., 1990). The eight subscales of the SIP are totaled for a weighed total score, with higher scores indicating more functional impairments (range of 0-5799). The mean SIP8 total score of a healthy group of 78 women was 65.5 (SD 137.8) (Servaes, et al., 2002). The SIP8 has good reliability (Bergner et al., 1981) and was validated for the Dutch population (Jacobs et al., 1990).

**Data analysis**

Data were analyzed using Statistical Package for the Social Sciences software (SPSS), version 23.0, with an alpha of $p < .05$ (two tailed).

Raw test scores were checked for outliers, score distributions and missing data. There was one extreme outlier for SDT 1, one outlier for SDT 2, one outlier for SCL-90-R, two outliers for CRT 1 compound score, one for CRT 2 compound score. These scores were replaced by the sample mean plus three standard deviations, as described by Field (2013). The scores on the CRT compound score and ASTM were not normally distributed and these variables were therefore log transformed. There were no missing data.

First, descriptive statistics were calculated. The FB and No FB groups were compared with the Mann-Whitney *U* test (ASTM), independent samples *t*-test (age; SDT; SCL-90-R; SIP8 Total; CIS Fatigue; log transformed CRT compound score), or the Chi-square test (education; sex). Additionally, the patients from the FB group that were selected in the matching procedure were compared with the non-selected patients on ASTM 2, SDT 2, and log transformed CRT 2 compound scores.

To determine the effects of feedback on subsequent performance on the ASTM, repeated measures analysis of variance (ANOVA) was conducted. The log transformed ASTM total score, SDT score and log transformed CRT score were used as dependent variables, time (baseline and follow-up) as a within-subjects factor, and feedback group (FB and No FB) as a between-subjects factor. Within-group effect sizes were calculated to evaluate the effects of repeating the neuropsychological tests on the FB and No FB groups.

To evaluate the clinical relevance of the feedback intervention, the SDT scores at baseline and follow-up were compared between the two groups (i.e., FB and No FB). Raw test scores on the SDT were compared with demographically adjusted SDT *T*-scores of the published norms ($n = 201$) in the manual of the WAIS (Stinissen et al., 1970). An SDT *T*-score of one standard deviation (i.e., SD) less than the mean was considered below "normal". No norm scores were available for the CRT. Therefore, the procedure of comparing individual scores with demographically adjusted norm scores could not be executed for the CRT.

## RESULTS

### Descriptive Statistics

After matching on their respective baseline measures, selected patients from the FB group ($n = 33$) were comparable to non-selected patients from the FB group ($n = 70$) on the ASTM 2 mean score (Mann-Whitney $U$ test, $p = .754$), SDT 2 score ($t = -1.68$, $df = 101$, $p = .095$) and log transformed CRT 2 compound score ($t = .946$, $df = 101$, $p = .346$). In the final sample, each group (i.e., FB and No FB) consisted of 33 patients.

As can been seen in Table 1, the patients in the two groups (i.e., FB and No FB) had comparable demographic and clinical measures at baseline. The mean level of education in both groups was medium vocational training. The sample consisted of primarily female subjects in their late thirties. The ASTM score ranges showed a strong ceiling effect. In this sample, 81.8% of the ASTM baseline scores varied between 82 and 85, within a range of 64-85. On the second administration of the ASTM, 81.7% of the scores varied between 82 and 90, within a range of 75-90.

### Effect of Feedback on Underperformance

Table 2 depicts the group (i.e., FB vs. No FB) differences on the repeated measure of underperformance (i.e., ASTM) and measures of information processing speed (i.e., SDT and CRT). Repeated measures analysis of variance (ANOVA), with the log transformed ASTM total score at baseline and follow-up as a within-subjects factor, and feedback group (FB and No FB) as a between-subjects factor, found no significant interaction between time and feedback group: $F(1, 64) = 1.63$, $p = .20$. There was a main effect of time, $F(1, 64) = 25.25$, $p < .001$, $\eta_p^2 = .28$, showing an overall improvement on re-administration. There was no main effect for feedback group, $F(1, 64) = 0.32$, $p = .57$, $\eta_p^2 = .005$. Together, these results demonstrated that the ASTM scores improved on re-administration in both the FB and No FB patient groups.

### Effect of Feedback on Information Processing Speed Performance

First, for the SDT, repeated-measures analysis of variance (ANOVA), with the SDT total score at baseline and follow-up as a within-subjects factor, and feedback group (FB and No FB) as a between-subjects factor the interaction between time and feedback group failed to reach significance, $F(1, 64) = 4.29$, $p = .058$, $\eta_p^2 = .05$. There was a main effect of time, $F(1, 64) = 9.58$, $p = .003$, $\eta_p^2 = .13$, indicating that the SDT scores increased between baseline and re-administration. There was a main effect of feedback group, $F(1, 64) = 4.67$, $p = .034$, $\eta_p^2 = .06$, indicating that there was a difference between the FB and No FB CFS patients groups on SDT scores. This difference seems more outspoken at follow-up assessment, as is illustrated by a larger within-group effect size of -.69 in the FB group compared to -.14 in the No FB group.

**Table 1**

*Baseline Means, Median Scores (for the ASTM), Standard Deviations, and Ranges for Demographical and Clinical Characteristics*

| | FB (*n* = 33) | No FB (*n* = 33) | *p*-value |
|---|---|---|---|
| Age (years) | 37.23 (9.39) | 39.90 (10.07) | .271 |
| | 20-55 | 22-59 | |
| Education, *n* (%) | | | .580 |
| Low | 9 (27.3) | 8 (24.2) | |
| Medium | 14 (42.4) | 18 (54.5) | |
| High | 10 (30.3) | 7 (21.2) | |
| Female sex, *n* (%) | 27 (81.8) | 25 (75.7) | .547 |
| SCL-90-R (total raw score) | 169.18 (42.41) | 188.73 (48.82) | .087 |
| | 112-297 | 123-317 | |
| ASTM (total raw score) | 83 (4.24) | 83 (4.37) | .704 |
| | 64-85 | 64-85 | |
| SDT (total raw score) | 51.39 (14.88) | 47.45 (13.04) | .257 |
| | 24-89 | 6-65 | |
| CRT compound (msec.) | 415.74 (103.95) | 453.73 (119.17) | .167 |
| | 292.00-792.15 | 299.00-792.15 | |
| CIS Fatigue (total raw score) | 51.69 (4.77) | 50.18 (5.18) | .221 |
| | 40-56 | 37-56 | |
| SIP8 (total raw score) | 1692.90 (540.50) | 1789.81(576.98) | .484 |
| | 829-2944 | 849-3382 | |

*Note.* Standard deviations are presented in parentheses; range scores are presented below the mean test scores; level of education was assessed by classifying formal schooling on an 8-point scale often used in the Netherlands (De Bie, 1987). Three groups of education levels were identified: Low (those with a primary education at most), Medium (those with junior vocational training at most), and High (those with senior vocational or academic training). FB = Feedback group; No FB = No Feedback group; ASTM = Amsterdam Short-Term Memory test; SDT = Symbol Digit Test; SCL-90-R = Symptom Checklist-Revised; CRT compound = Complex Reaction Time task compound score; CIS Fatigue = Checklist Individual Strength, Fatigue subscale; SIP8 Total = Sickness Impact Profile 8.

At baseline, an equal number of participants scored less than "normal" (i.e., < 1 *SD* below the mean) on the SDT in both groups (i.e., 4/33 in the FB and No FB groups). At follow-up, significantly more subjects improved to a "normal" score on the SDT in the FB group, compared to the No FB group ($\chi^2 = 5.41$, $p = .020$). During follow-up, all of the subjects in the FB group scored in the "normal" range (i.e., > 1 *SD* below the mean). In the No FB group, 5 of 33 (i.e., 15%) produced less than average scores at follow-up.

Second, for the CRT log transformed compound score, repeated-measures analysis of variance (ANOVA), with the CRT log transformed compound score at baseline and follow-up as a within-subjects factor, and feedback group (FB and No FB) as a between-subjects factor, found an interaction effect for time and feedback group, $F(1, 64) = 13.27$, $p = .001$, $\eta_p^2 = .17$ indicating that patients in the FB group improved significantly more on the CRT compared to patients in the No FB group (*see* Table 2). There was

a main effect of time, $F(1, 64) = 5.17$, $p = .02$, $\eta_p^2 = .07$, showing that the CRT scores improved between baseline and re-administration. There was a main effect for feedback group, $F(1, 64) = 8.31$, $p = .005$, $\eta_p^2 = .11$, indicating that patients in the FB group scored significantly higher on the CRT compared to the No FB group. Additionally, the within-group effect size was -.71 in the FB group and .20 in the No FB group. In summary, these results demonstrated that the FB group showed significant greater improvement on the CRT than the No FB group.

**Table 2**

*Baseline and Follow-Up Means (SDT & CRT), Median Scores (ASTM), Percentages Failing the ASTM, and Percentages of Subjects Who Score Below Average on the SDT*

| | **FB (n = 33)** | | **No FB (n = 33)** | |
|---|---|---|---|---|
| | **Time 1** | **Time 2** | **Time 1** | **Time 2** |
| ASTM (total raw score) | 83 (4.24) | 85 (3.23) | 83 (4.37) | 85 (4.02) |
| *% below cut-off* | 100 | 52 | 100 | 61 |
| SDT (total raw score) | 51.39 (14.88) | 58.12 (13.66) | 47.45 (13.04) | 48.94 (11.58) |
| *% < 1 SD below mean* | 12 | 0 | 12 | 15 |
| CRT compound (msec.) | 415.74 (103.95) | 362.03 (55.35) | 453.73 (119.17) | 472.01 (151.38) |

*Note.* Standard deviations are presented in parentheses.

## DISCUSSION

We examined the effect of providing feedback on underperformance on the subsequent neuropsychological test performance of CFS patients. To our knowledge, this is the second study to examine the effect of feedback on underperformance in clinical patients.

Our main findings are that (a) underperformance occurred in a substantial number of CFS patients referred for treatment (i.e., 23.9%), (b) underperformance was not stable between assessments, and (c) the feedback intervention had no effect upon underperformance on the re-administered PVT (i.e., ASTM), and was associated with mixed findings on measurements of information processing speed during follow-up.

Before discussing the study findings in detail, we want to address one major limitation of this study: the difference in time interval between baseline and follow-up of the FB group (i.e., one month) and No FB group (i.e., six months). The difference in the time interval was caused by limited treatment capacity, and not by a systematic flaw in the study design. Also, the two groups were matched on baseline measures, and analyses showed that this matching was done without influencing outcome measurements. Nonetheless, the difference in time interval might have influenced our results. First to mention is the difference in practice effect (Heilbronner et al., 2010) in the two conditions. It could be argued that the shorter between-session period in the FB group resulted in higher scores at re-assessment compared to the No FB group. Unfortunately, the practice effects of the utilized neuropsychological

tests in this study (i.e., SDT and CRT) for the two different time intervals in CFS patients are unknown. Additionally, the longer waitlist period in the No FB group may have resulted in changes in health status (i.e., worsening of symptoms) compared to the FB group, which could have affected the results.

The large proportion of patients in both groups (i.e., 39% and 48%) that performed within normal limits on the ASTM during the repeated neuropsychological assessment, underscores the idea that underperformance is not a static trait. That is, given the high test-retest reliability of the ASTM, a change in test score at follow-up assessment, was probably not due to error variance but due to a change in underperformance. Our findings are comparable to those of van Valen et al. (2015), who found that, in a sample of 323 clinical patients with chronic solvent-induced encephalopathy (CSE), 42% reverted from an invalid to valid score on a PVT at re-assessment after one year. The practical implication of these findings is that performance validity should be checked for in every repeated neuropsychological assessment.



The lack of a group difference on the ASTM in combination with improvements on measures of information processing speed in the FB group could intuitively seem inconsistent, and raises questions. However, it could be argued that - because the CRT is likely more sensitive in detecting change (i.e., measured in msec.) compared to the SDT -, a group difference was found on this measure. Moreover, the SDT showed a comparable trend - albeit borderline significant - and also showed more "clinical" improvement at re-assessment in the FB group. Therefore, despite the mentioned limitations, these findings may tentatively suggest a positive feedback effect.

An important strength of the current study is that during follow-up all tests (i.e., ASTM, SDT, and CRT) were re-administered in both groups (i.e., FB and No FB). Suchy and colleagues (2012) only repeated the PVT (i.e., VSVT) in the FB group, and found that 68% of the patients reverted to a valid score range upon re-administration. Consequently, they concluded that a feedback intervention upon PVT failure effectively decreased underperformance. In the current study, we found that 48% of the patients in the FB group scored in the valid score range at follow-up. However, because 39% in the No FB group also showed improvement on the ASTM at re-administration, we found no significant effect of the feedback intervention upon subsequent PVT performance. This result suggests that the conclusion of a positive effect of a feedback intervention upon underperformance reported by Suchy and colleagues (2012) was likely premature.

Although our study has its merits in advancing the understanding of the effect of feedback upon underperformance, its retrospective and naturalistic design is a limitation. An experimental design with random group allocation (i.e., FB and No FB) is necessary to further determine the effect of feedback on underperformance. Additionally, future studies could use theoretical frameworks associated with feedback responsiveness in cases of non-credible test results, such as deterrence theory (Horner et al., 2017) and cognitive dissonance theory (Merckelbach et al., 2015). This results in a theory-driven intervention, which could lead to better understanding and reproducibility of study findings.

How to interpret retest data after feedback upon underperformance clinically? First, patients who continue to underperform during follow-up are likely capable of better performance. Their neuropsychological test results are invalid. Second, improvement on a PVT (i.e., performance in the valid score range at re-administration) after a feedback intervention does not automatically indicate that the

patient did not underperform and that the neuropsychological test scores are thus valid. Nonetheless, even if a normal score reflects an underestimation, this score still excludes cognitive deficits. This fact is important since the primary purpose of a neuropsychological assessment is, as stated by Boone (2007, p. 37), "to determine whether patients have objectively (i.e., credible) verified cognitive dysfunction". Although some researchers argued that warning subjects that measures of underperformance will be administered reduces subsequent response bias (i.e., Schenk & Sullivan, 2010; Johnson & Lesniak-Karpiak, 1997), Youngjohn et al., (1999) suggest that an improvement in performance on a re-administered PVT after such a warning intervention might be the result of a more sophisticated form of response bias, in which case the question about the validity of the test results remains. This result is not only a limitation of this study but of the entire field of neuropsychological assessment: a gold standard to measure performance validity is lacking. Therefore, clinicians should be cautious in interpreting neuropsychological test results in the presence of PVT failure, even after improved performance with repeated administration. In general, the credibility of the symptoms and test performance of an individual patient is determined by the clinician relative to all of the available information (i.e., information from the interview with the patient, his/her behavioral presentation during the assessment, and the scientific knowledge of patterns and severity of cognitive disorders associated with the clinical condition). This determination of the credibility of the neuropsychological test results is improved by the employment and consideration of validity tests, compared to clinical judgment alone (Dandachi-FitzGerald et al., 2017).

In conclusion, our findings suggest that (a) underperformance occurs in a substantial number of CFS patients referred for treatment, (b) performance validity is not a stable characteristic but fluctuates between assessments, and (c) a limited feedback intervention may result in improvement on information processing speed performance. These findings imply that that performance validity should be assessed in every repeated neuropsychological assessment. Similarly, research studies on cognitive deficits in CFS patients need to take performance validity into account. Finally, the possible positive effect of a feedback intervention warrants further research. Engaging in underperformance might reinforce the patient's experience of symptoms and illness behavior. Therefore, it is important that research into strategies that might prevent or alter this behavior is conducted.

## REFERENCES


Arrindell, W. A., & Ettema, J. H. (2005). Symptom Checklist. Handleiding bij een multidimensionale psychopathologie indicator [Symptom Checklist. Manual of a multidimensional psychopathology indicator]. Amsterdam: Harcourt Test Publishers.

Bergner, M., Bobbitt, R. A., Carter, W. B., & Gilson, B. S. (1981). The Sickness Impact Profile - development and final revision of a health-status measure. *Medical Care, 19,* 787-805.

Boone, K. B. (2007). *Assessment of feigned cognitive impairment: A neuropsychological perspective.* New York: Guilford Press.

Bush, S. S., Ruff, R. M., Tröster, A. I., Barth, J. T., Koffler, S. P., Pliskin, N. H., Reynolds, C. R., & Silver, C. H. (2005). Symptom validity assessment: practice issues and medical necessity NAN policy & planning committee. *Archives of Clinical Neuropsychology*, *20*(4), 419–426. https://doi.org/10.1016/j.acn.2005.02.002

Busichio, K., Tiersky, L. A., DeLuca, J., & Natelson, B. H. (2004). Neuropsychological deficits in patients with chronic fatigue syndrome. *Journal of the International Neuropsychological Society, 10,* 278–285.

Capuron, L., Welberg, L., Heim, C., Wagner, D., Solomon, L., Papanicolaou, D. A., Craddock, R. C., Miller, A. H., & Reeves, W. C. (2006). Cognitive dysfunction relates to subjective report of mental fatigue in patients with chronic fatigue syndrome. *Neuropsychopharmacology*, *31*(8), 1777–1784. https://doi.org/10.1038/sj.npp.1301005

Carone, D. A., Iverson, G. L., Bush, S. S. (2010). A model to approaching and providing feedback to patients regarding invalid test performance in clinical neuropsychological evaluations. *The Clinical Neuropsychologist, 24,* 759–778.

Centraal Begeleidings Orgaan. (2013). *Richtlijn Diagnose, behandeling, begeleiding en beoordeling van patiënten met het chronisch vermoeidheidssyndroom (CVS) [Guideline: Diagnosis, treatment, coaching and evaluation of patients suffering chronic fatigue syndrome (CFS)].* Retrieved from *https:// www.diliguide.nl/document/3435/file/pdf/* 2013.

Christodoulou, C., DeLuca, J., Lange, G., Johnson, S. K., Sisto, S. A., Korn, L., & Natelson, B. H. (1998). Relation between neuropsychological impairment and functional disability in patients with chronic fatigue syndrome. *Journal of Neurology, Neurosurgery & Psychiatry, 64*, 431–434.

Cockshell, S. J., & Mathias, J. L. (2010). Cognitive functioning in chronic fatigue syndrome: A meta-analysis. *Psychological Medicine, 40,* 1253–67.

Cockshell, S. J., & Mathias, J. L. (2012). Test effort in persons with chronic fatigue syndrome when assessed using the Validity Indicator Profile. *Journal of Clinical and Experimental Neuropsychology, 34*, 679-687.

Cockshell, S. J., & Mathias, J. L. (2014). Cognitive functioning in people with chronic fatigue syndrome: A comparison between subjective and objective measures. *Neuropsychology, 28,* 394-405.

Dandachi-FitzGerald, B., Merckelbach, H. & Ponds, R. W. (2017). Neuropsychologists' ability to predict distorted symptom presentation. *Journal of Clinical and Experimental Neuropsychology, 39,* 257-264.

Dandachi-FitzGerald, B., Ponds, R. W., Peters, M. J., & Merckelbach, H. (2011). Cognitive underperformance and symptom over-reporting in a mixed psychiatric sample. *The Clinical Neuropsychologist*, *25*(5), 812–828. https://doi.org/10.1080/13854046.2011.583280

De Bie, S. (1987). Standaardvragen 1987: Voorstellen voor uniformering van vraagstellingen naar achtergrondkenmerken en interviews [Standard questions 1987: Proposal for uniformization of questions regarding background variables and interviews]. Leiden, the Netherlands: Leiden University Press.

DeLuca J., Johnson, S., Ellis, S. P., & Natelson, B. H. (1997). Cognitive functioning is impaired in patients with chronic fatigue syndrome devoid of psychiatric disease. *Journal of Neurology, Neurosurgery & Psychiatry, 62,* 151–155.

Derogatis, L. R. (1994). SCL-90-R: Administration, scoring and procedures manual (3rd ed.). Minneapolis, MN: Nation Computer Systems.

Field, A. P. (2013). *Discovering statistics using IBM SPSS Statistics: and sex and drugs and rock 'n' roll (fourth edition)*. London: Sage publications.

Fox, D. D. (2011). Symptom validity test failure indicates invalidity of neuropsychological tests. *The Clinical Neuropsychologist, 25,* 488-498.

Fukuda, K., Straus, S. E., Hickie, I., Sharpe, M. C., Dobbins, J.G., & Komaroff, A. (1994). The chronic fatigue syndrome: A comprehensive approach to its definition and study. International Chronic Fatigue Syndrome Study Group. *Annals of Internal Medicine, 121*, 953–959.

Goedendorp, M., van der Werf, S., Bleijenberg, G., Tummers, M., & Knoop, H. (2013). Does neuropsychological test performance predict outcome of cognitive behavior therapy for chronic fatigue syndrome and what is the role of underperformance? *Journal of Psychosomatic Research, 75*, 242–248.

Heilbronner, R., Sweet, J., Attix, D., Krull, K., Henry, G., & Heart, R. (2010). Official position of the American Academy of Clinical Neuropsychology on serial neuropsychological assessments: The utility and challenges of repeat test administration in clinical and forensic contexts. *The Clinical Neuropsychologist, 24,* 1267-1278.

Heilbronner, R., Sweet, J., Morgan, J., Larrabee, G., Millis, S. & Conference Participants (2009). American Academy of Clinical Neuropsychology consensus conference statement on the neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist, 23,* 1093–1129.

Horner, M. D., Turner, T. H., VanKirk, K. K., & Denning, J. H. (2017). An intervention to decrease the occurrence of invalid data on neuropsychological evaluation. *Archives of Clinical Neuropsychology, 32,* 228-237.

Jacobs, H. M., Luttik, A., Touw-Otten, F. W., & de Melker, R. A. (1990). The Sickness Impact Profile; results of an evaluation study of the Dutch version (De 'sickness impact profile'; resultaten van een valideringsonderzoek van de Nederlandse versie). *Nederlands Tijdschrift voor Geneeskunde, 134,* 1950–1954.

Jason, L. A., Richman, J. A., Rademaker, A. W., Jordan, K. M., Plioplys, A. V., Taylor, R. R., McCready, W., Huang, C. F., & Plioplys, S. (1999). A community-based study of chronic fatigue syndrome. *Archives of Internal Medicine*, *159*(18), 2129–2137. https://doi.org/10.1001/archinte.159.18.2129

Johnson, J. L., & Lesniak-Karpiak, K. (1997). The effect of warning on malingering on memory and motor tasks in college samples. *Archives of Clinical Neuropsychology, 12*, 231–238.

Knoop, H., Prins, J., Stulemeijer, M., van der Meer, J. W., & Bleijenberg, G. (2007). The effect of cognitive behaviour therapy for chronic fatigue syndrome on self-reported cognitive impairments and neuropsychological test performance. *Journal of Neurology, Neurosurgery and Psychiatry, 78,* 434–436.

Knoop, H., van der Meer, J. W., & Bleijenberg, G. (2008). Guided self-instructions for people with chronic fatigue syndrome: randomised controlled trial. *The British Journal of Psychiatry, 193*, 340-341.

Krishnan, M. & Donders, J. (2011). Embedded assessment of validity using the Continuous Visual Memory Test in patients with traumatic brain injury. *Archives of Clinical Neuropsychology, 26,* 176-183.

Larrabee, G. J. (2012). Performance validity and symptom validity in neuropsychological assessment. *Journal of the International Neuropsychological Society, 18,* 625–631.

Marshall, P., Schroeder, P., O'Brien, J., Fischer, R., Ries, A., Blesi, B., & Barker, J. (2010). Effectiveness of symptom validity measures in identifying cognitive and behavioral symptom exaggeration in adult attention deficit hyperactivity disorder. *The Clinical Neuropsychologist, 24*, 1204-1237.

Merckelbach, H., Dandachi-FitzGerald, B., van Mulken, P., Ponds, R. W., & Niesten, E. (2015). Exaggerating psychopathology produces residual effects that are resistant to corrective feedback: An experimental demonstration. *Applied Neuropsychology: Adult, 22,* 16-22.

Reeves, W. C., Lloyd, A., Vernon, S. D., Klimas, N., Jason, L. A., Bleijenberg, G., Evengard, B., White, P. D., Nisenbaum, R., Unger, E. R., & International Chronic Fatigue Syndrome Study Group (2003). Identification of ambiguities in the 1994 chronic fatigue syndrome research case definition and recommendations for resolution. *BMC health services research*, *3*(1), 25. https://doi.org/10.1186/1472-6963-3-25

Roor, J. J., Dandachi-FitzGerald, B., & Ponds, R. W. (2016). A case of misdiagnosis of mild cognitive impairment: The utility of symptom validity testing in the outpatient memory clinic. *Applied Neuropsychology: Adult, 23*, 172-178.

Schenk, K. & Sullivan, K. A. (2010). Do warnings deter rather than produce more sophisticated malingering? *Journal of Clinical and Experimental Neuropsychology, 32,* 752–762.

Schmand, B., & Lindeboom, J. (2005). Amsterdam Short-Term Memory Test: Manual. Leiden, The Netherlands: Psychologische Instrumenten, Tests en Services.

Servaes, P., Verhagen, C., & Bleijenberg, G. (2002). Determinants of chronic fatigue in disease-free breast cancer patients, a cross-sectional study. *Annals of Oncology, 13,* 589-598.

Stinissen, J., Willems, P. J., Coetsier, P., & Hulsman, W. L. (1970). Handleiding bij de Nederlandse bewerking van de WAIS [Manual of the Dutch edition of the WAIS]. Amsterdam: Swets and Zeitlinger.

Suchy, Y., Chelune, G., Franchow, E. I., & Thorgussen, S. R. (2012). Confronting patients about insufficient effort: The impact on subsequent symptom validity and memory performance. *The Clinical Neuropsychologist, 26,* 1296–1311.

Van der Werf, S., Prins, J., Jongen, P., van der Meer, J., & Bleijenberg G. (2000). Abnormal neuropsychological findings are not necessarily a sign of cerebral impairment: A matched comparison between chronic fatigue syndrome and multiple sclerosis. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology, 13,* 199–203.

Van der Werf, S., de Vree, B., van der Meer, J., & Bleijenberg, G. (2002). The relations among body consciousness, somatic symptom report, and information processing speed in chronic fatigue syndrome. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology, 15,* 2–9.

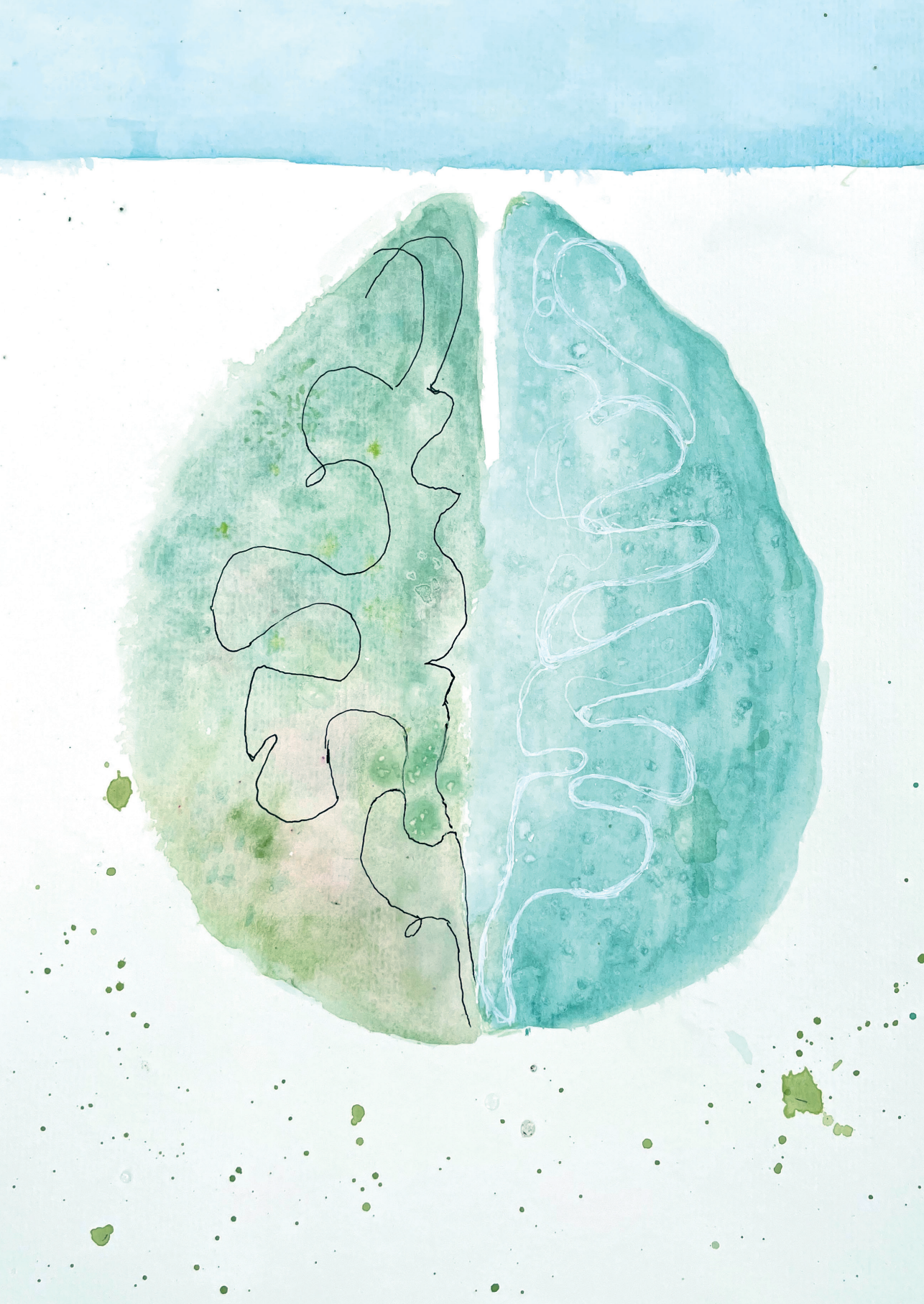
Van Valen, E., van Hout, M., Heutink, M., Wekking, E., van der Laan, G., Hageman, G., van Dijk, F., Sprangers, M., & Schmand, B. (2015, June). Performance validity in chronic solvent-induced encephalopathy. Poster session presented at the annual meeting of the Developmental Neurotoxicology Society, Montréal, Canada.

Vercoulen, J. H., Alberts, M., & Bleijenberg, G. (1999). De Checklist Individual Strength (CIS). *Gedragstherapie, 32*, 131–136.

Vercoulen, J. H., Bazelmans, E., Swanink, C. M., Galama, J. M., Fennis, J. F., van der Meer, J. W., & Bleijenberg, G. (1998). Evaluating neuropsychological impairment in chronic fatigue syndrome. *Journal of Clinical and Experimental Neuropsychology, 20*, 144–56.

Worm-Smeitink, M., Gielissen, M., Bloot, L., van Laarhoven, H. W., van Engelen, B. G., van Riel, P., Bleijenberg, G., Nikolaus, S., & Knoop, H. (2017). The assessment of fatigue: Psychometric qualities and norms for the Checklist individual strength. *Journal of Psychosomatic Research, 98,* 40-46.

Youngjohn, J. R., Lees-Haley, P. R., & Binder, L. M. (1999). Comment: Warning malingerers produces more sophisticated malingering. *Archives of Clinical Neuropsychology, 14,* 511–515.



# CHAPTER 5

## No impact of a clinical feedback intervention when patients invalidate testing: A multi-site, single-blind randomized controlled trial

Roor, J. J., Dandachi-FitzGerald, B., Peters, M. J., & Ponds, R. W.
No impact of a clinical feedback intervention when patients invalidate testing:
A multi-site, single-blind randomized controlled trial.
*(under review)*

## ABSTRACT

### Objective

Performance below the actual abilities of the examinee can be measured using performance validity tests (PVTs). PVT failure negatively impacts the quality of the neuropsychological assessment. In our study, we addressed this issue by providing feedback to improve test-taking behavior.

### Method

This study is a multisite single-blind randomized controlled trial in a consecutive sample of clinically referred adult patients ($N$ = 196) in a general hospital setting. Patients who failed a PVT ($n$ = 71) were randomly allocated to feedback condition (FB; $n$ = 39) or no-feedback condition (NO-FB; $n$ = 32). Only the FB group received immediate feedback upon failing a PVT, in which lower than expected performance was addressed. Both groups (FB and NO-FB) were provided with the same subsequently repeated and newly administered tests.

### Results

There were no group (FB vs. NO-FB) differences on both the repeated and single-administered PVTs and standard cognitive tests. In fact, independent of feedback condition, the vast majority of patients continued to fail at least one PVT.

### Conclusions

Our study found that providing immediate feedback to address PVT failure and improve test-taking behavior did not improve consequent test performance. These results suggest limited value of using feedback as an intervention to address PVT failure. It highlights the need for more research to identify more effective approaches that can enhance patients test-taking behavior. Ultimately, such efforts are critical in ensuring accurate diagnosis and effective treatment recommendations for patients.

## INTRODUCTION

Over the past two decades, there has been increased attention to performance validity testing during neuropsychological assessment in routine clinical care, leading to updated consensus statements by major professional associations (Moore et al., 2021; Sweet et al., 2021). Nowadays, proactively assessing performance validity during clinical assessments is considered crucial for confidently interpreting neuropsychological test performance (Sweet et al., 2021). Moreover, this approach is justifiable since recent meta-analyses findings have confirmed the substantial presence of invalid performance in adult clinical patients (Roor et al., 2023). Survey results indicate that most clinicians acknowledge the aforementioned consensus statements and claim to adhere to their validity testing recommendations (Hirst et al., 2017). However, these professional guidelines primarily focus on diagnostic features and research practices, providing little guidance on how and when to communicate invalid test results to patients.

As Postal and Armstrong (2013) discussed in their book, "providing direct feedback to patients should be considered a key element of a neuropsychologist's professional scope" (p. 5). Recently, empirical findings showed that a feedback intervention upon neuropsychological assessment in patients with multiple sclerosis (MS) has the potential to improve outcomes at one-month follow-up (Longley et al., 2022). However, feedback upon neuropsychological assessment also comes with specific challenges, especially in the context of performance validity test (PVT) failure. As stated by Carone and colleagues (2010), one aspect may lie in the nature of the feedback (e.g., that the patient did not try his/her best) that "creates the potential for significant interpersonal conflict with the patient" (p. 760). Clinicians may feel uncomfortable in situations where patients do not provide accurate information, demonstrating a potential unwillingness or inability to fully collaborate during the neuropsychological assessment. For example, some clinicians may decide to omit feedback on invalid performance entirely out of fear that the patient will file a complaint of professional misconduct to regulatory agencies. Others may expect that the patient will react quarrelsome when the issue of noncredible performance is addressed, and choose to downplay these findings and use more comfortable but inaccurate explanations of PVT failure (e.g., pain, fatigue, or the ill-defined concept "cry for help"). However, such feedback strategies are undesirable as they may lead to patient harm such as iatrogenesis (Martin & Schroeder, 2022). Instead, and in line with professional ethical principles, accurate communication of evidence-based performance validity assessment should be based upon scientific and professional knowledge (American Psychological Association, 2017).

Whereas the vast majority of clinicians indicate that they do discuss PVT failure directly with the patient, there is minimal consensus regarding a specific feedback approach (Martin & Schroeder, 2021). Carone et al. (2010) provided a first practice-based model for addressing feedback on PVT failure. They suggest – amongst others – a non-adversarial approach and 'good-news bad-news' approach (i.e., 'bad news' are the low test scores, 'good news' is that the test results show that the patient is capable of much better performance). Recently, Martin and Schroeder (2022) published their 'firm-beneficent' feedback approach for addressing invalid performance. A major difference with the model of Carone et al. (2010), is that it does not attempt to uncover reasons for invalid presentations. According to these authors,

"exploring the exact reasons for underperformance is often unproductive, and that doing so may detract from an opportunity to shift the focus of the feedback session into a more therapeutic direction" (p. 61). Consequently, they argue to only briefly focus on invalid data and recommend to switch to discussing therapy goals "believed to be truly present and in need of treatment" (p. 63).

It is important to note that both proposed feedback-models to improve test performance specifically focus on a separate feedback session *after* the neuropsychological assessment is completed. Survey results, however, found that approximately half of the working neuropsychologists reported giving immediate feedback upon indication of invalid performance *during* the assessment, ranging from "sometimes" to "always" (Dandachi-Fitzgerald et al., 2013; Hirst et al., 2017). Yet, guidelines on when and how to address invalid performance *during* the neuropsychological assessment are still scarce, as are empirical data on the effects of addressing indications of invalid performance on subsequent test-taking attitude.

To the best of our knowledge, only two studies examined the impact of a feedback intervention when adult clinical patients failed a PVT. Suchy et al. (2012) examined archival data of patients with multiple sclerosis (MS) in which a subset of patients was given feedback upon PVT-failure. This feedback was given *during* the test session by the neuropsychologist consisting of (a) stressing the importance of the patient putting forth their best effort, followed by (b) the information to the patient that performance on a previously administered measure (i.e., Victoria Symptom Validity Test, VSVT) was questionable and needed to be repeated. The majority of subjects who received immediate feedback on invalid performance reverted to a VSVT-score in the valid range during re-assessment. Moreover, the authors found that the group of patients who received immediate feedback performed significantly higher on a standard memory test (i.e., Wechsler Memory Scale-III, WMS-III) than the group who did not receive feedback. One methodological drawback of the study of Suchy et al. (2012) is that practice effects possibly impacted their findings, as they did not use a control condition in which tests were repeated without an intervention upon PVT-failure. In the second study, Roor et al. (2020) examined archival data of patients diagnosed with chronic fatigue syndrome (CFS). Due to naturalistic changes in clinical procedures, a subset of these patients was provided with feedback upon failing an earlier administered PVT. This was done *after* the initial neuropsychological assessment was completed. During the feedback intervention, the psychologist (a) addressed that the CFS symptoms of the patient were difficult to evaluate because of the lower-than-expected performance on a previously administered test, (b) emphasized exerting the patient's best effort and that improvement was expected, and (c) explained that therefore tests needed to be repeated. After the feedback intervention, the PVT and two neuropsychological tests (Complex Reaction Time task, CRT; Symbol Digit Test, SDT) were repeated. This group was compared to a matched group of CFS patients who were not provided with feedback upon PVT-failure. The authors found that performance on a PVT equally improved in both groups upon re-assessment, but the feedback group did show significant improvement on one out of two measures of information processing speed (i.e., CRT) compared to the no-feedback group. A major limitation in the study from Roor et al. (2020), however, was the considerable difference in the average time intervals between the first and second assessment for the feedback group (one month) and no-feedback group

(six months). Consequently, although a control condition was used in this study, their results may be impacted by the mentioned group differences in time-interval.

In the current study, we took these methodological problems into account by using an experimental design in which adult patients in a clinical setting were randomly assigned to a feedback condition directly after failing a PVT. We examined whether this feedback intervention - that was based upon the mentioned studies from Suchy et al. (2012) and Roor et al. (2020) - impacted subsequent performance of a repeated PVT and repeated standard cognitive tests. Additionally, group differences (feedback vs. no-feedback) were examined for a subsequently single-administered (i.e., non-repeated) PVT and standard cognitive tests. We expected that subjects in the feedback condition would show improvement in their efforts to perform at the best of capabilities, leading to improved PVT results and consequently improved performance on standard cognitive tests, compared to the subjects in the no-feedback condition where invalid performance was left unaddressed.



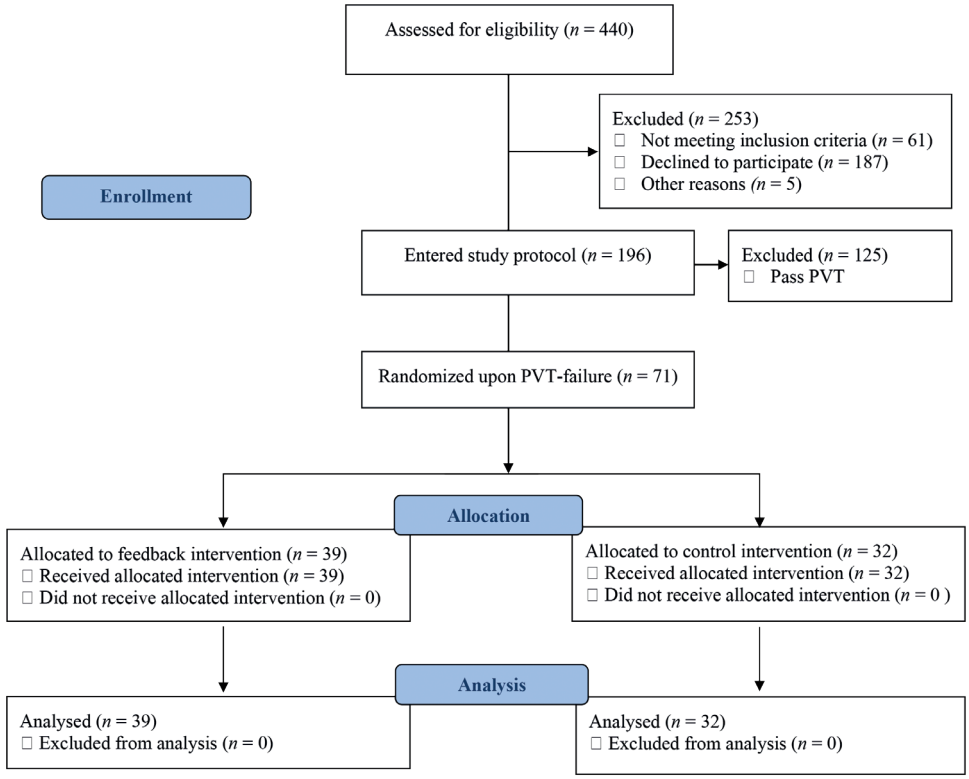
## METHOD

### Participants

Patients were consecutively referred for neuropsychological assessment to the Medical Psychology departments of seven medical hospitals in the Netherlands between January 2016 and October 2018. These hospitals were: VieCuri Medical Center, Venlo; Maastricht University Medical Center (MUMC+); Isala Clinics, Zwolle; ZiekenhuisGroep Twente (ZGT), Almelo; Elizabeth-TweeSteden Hospital (ETZ), Tilburg; St. Jans Gasthuis (SJG), Weert; and Máxima Medical Center (MMC), Veldhoven. All referrals were outpatients seen for neuropsychological assessment in the context of routine clinical care. While some patients were concurrently involved in legal proceedings, it is crucial to note that none of the evaluations conducted in this study were medico-legal assessments. The PVT used for group allocation in this study is not deemed valid in case of severe cognitive impairments, as seen in patients with Alzheimer's disease, Korsakoff, acute psychosis, and traumatic brain injury patients still in a state of post-traumatic amnesia. Accordingly, in adherence to the test manual, patients exhibiting clinically obvious symptoms were excluded. The operational definition of the Amsterdam Short-Term Memory test (ASTM) was used: "Clinically obvious symptoms are symptoms that are obvious during informal contact with the patient or during history taking, without there being any necessity to use formal cognitive tests to provoke these symptoms (e.g., repeating the same information, not knowing recent personal facts, or failing to refer to an earlier subject of conversation". (Schmand & Lindeboom, 2005, p. 4). Additionally, for the same reason of preventing false-positive PVT classification, subjects with intellectual disability or less than eight years of formal schooling were excluded from participation as standard PVT cutoffs in these patient groups have shown unacceptable low specificity rates (e.g., Lippa, 2018). Furthermore, patients had to be proficient of the Dutch language, be 18 years or older, and had to be mentally competent to consent for participation. The number of eligible patients who gave written informed consent during the study period constituted the final sample. Figure 1 delineates the flow of the participants through the trial based on these criteria.

**Figure 1**

*Consort Diagram*




## Study Design

This study was designed as a multicenter single-blind randomized-controlled trial. Patients who failed a PVT were allocated to either the feedback intervention (i.e., feedback-group; FB), or were asked to repeat several tests without specifically addressing the issue of performance invalidity (i.e., no-feedback group; NO-FB). The research protocol was reviewed and approved by the standing ethical committee of Maastricht University. Local, independent ethics committees of the participating seven hospitals reviewed and approved the study protocol. The study was conducted in accordance with the International Conference on Harmonization Tripartite Guidelines on Good Clinical Practice (ICH-GCP) and in compliance with the Declaration of Helsinki 1964, as modified in October 2000. To ensure study protocol adherence, all site personnel received schooling by the first author (JR). Additionally, the study was independently monitored by the Clinical Trail Center Maastricht on all study sites to verify protocol adherence and that trial data are accurate, complete, and verifiable from source documents.

## Measures

### *Performance Validity Tests (PVTs)*

Performance validity was measured with the Amsterdam Short-Term Memory test (ASTM; Schmand et al., 1999). The ASTM is presented as a memory test and involves a 30-trial forced-choice word recognition procedure. The total calculated score is used as a cutoff for invalid performance. As stated by Dandachi-FitzGerald and Martin (2022, p. 114), "misclassifying authentic cognitive impairments as noncredible can have serious adverse impact on a patient's future health care. Hence, the general consensus is that validity tests require a specificity of at least .90". Therefore, instead of the standard cutoff of ≤ 84, a one-point lower cut score (i.e., ≤ 83) was used as this cut score is associated with the generally accepted specificity > 90% (i.e., 93.2%; Schmand & Lindeboom, 2005) – thereby minimizing the potential for false-positive findings on this test. Administration followed instructions specified in the test manual. The ASTM versions used at baseline and follow-up were identical. Performance validity was also examined using the Word Memory Test (WMT; Green, 2005). This is one of the most studied, validated, and used measures of performance validity (Eichsteadt, et al., 2014; Martin et al., 2015). We used the primary effort indicators Immediate Recognition (IR), Delayed Recognition (DR), and Consistency (CNS) subtests of the computerized Dutch-language version of the WMT, assessing the recognition of an earlier presented word list. Administration and interpretation were according to the test manual.



### *Standard Cognitive Tests*

The Digit Span subtest of the Wechsler Adult Intelligence Scale-IV (WAIS-IV; Wechsler, 2008) is a test of verbal attention and working memory. To evaluate semantic-based word retrieval, Category Fluency subtest of the Groninger Intelligence Test-II (GIT-II; Luteijn & Barelds, 2004; van der Elst et al., 2006a) was used. The Dutch language version of the Stroop Color Word Test (SCWT; Hammes, 1971; van der Elst et al., 2006b) was used for measuring executive functioning. Episodic memory was examined with the Dutch version of the Verbal Learning Test (VLT; Brand & Jolles, 1985; Van der Elst et al., 2005).

### *Clinician's Checklist*

Clinicians (i.e., neuropsychologist involved with the neuropsychological assessment) completed a checklist adapted from Dandachi-FitzGerald et al. (2017) after the clinical interview and having seen the patient files, but before testing took place. The checklist addressed the following patient variables: age, sex, level of education, diagnostic category, employment status, type of income, (partial) disability or sickness benefits, and being involved in legal proceedings because of their medical condition (e.g., pending disability claim).

## Procedures

### *Recruitment and Informed Consent*

Recruitment for the study occurred by offering eligible patients a study information brochure following the clinical interview. This brochure detailed the study procedure, such as an additional 50 minutes of testing on top of the tests and questionnaires that would be administered as care-as-usual. In line with professional guidelines, participants were naïve about the true study objective that feedback on

notions of invalid performance was to be examined –to ensure PVT effectiveness (Sweet et al., 2021, p. 17). All participants completed a written informed consent form, and were not provided with any form of incentive for participating in the study. During this consent procedure, conducted just after the clinical interview, the clinician emphasized to participants the importance of putting forth their best efforts during the planned neuropsychological assessment in order to obtain valid data on their cognitive status.

***Neuropsychological Assessment***

The neuropsychological assessment followed the study protocol, in which an PVT, the ASTM, was administered first, followed by the Digit Span and the Category Fluency respectively. These two other tests were administered along with the ASTM in the first part of the assessment to avoid identification of the ASTM as a PVT. In case the ASTM was failed, the participant was randomly allocated to a specific feedback intervention (FB) or not (NO-FB). At the start of the study, concealed envelopes containing equivalent numbers of FB and NO-FB conditions were randomly distributed over the seven participating hospitals. After the allocated feedback was communicated to the participant, the ASTM, Digit Span, and Category Fluency were repeated. Then, all patients completed the VLT, SCWT, and the IR and DR trials of the WMT. After completion of the study protocol, the neuropsychological assessment was continued as care-as-usual. It is important to note that, just before and throughout the assessment, the technician would, when deemed appropriate, stress the importance of the participant performing at best capabilities to obtain valid data.

***Feedback Intervention***

The technician communicated the allocated condition (i.e., FB or NO-FB) verbally to the participants using a neutral tone. In the FB condition, the following wording was used: "In the first part of the assessment, you performed lower than expected. That is why I will repeat some of the tests I have just administered". In the NO-FB condition, the following instructions were given: "As part of the assessment, I will repeat some of the tests I have just administered". It is important to note that the feedback intervention followed after the participant had already received immediate feedback about their performance on each item of the ASTM. As a standard test-taking procedure (*see* ASTM manual), subjects were provided with oral feedback from the technician after each of the 30 trials. This varied from 'You're doing wonderful! All three correct!' using enthusiastic tone, to 'One word correct' using a neutral tone of voice.

To summarize, before the allocated feedback intervention (FB or NO-FB) was given, a context was created where (1) the neuropsychologist emphasized the importance of performing at their best capabilities to obtain valid data on the patients' cognitive status during the informed consent procedure, (2) the technician reiterated the importance of performing at their best capabilities for obtaining valid data just before and throughout the assessment, as deemed appropriate, and (3) the technician, following the ASTM manual, provided real-time verbal feedback on performance on the ASTM for each trial. Therefore, we expected that the allocated feedback intervention (FB vs. NO-FB) would be grounded in these contextual aspects, alerting the participant about the importance of performing at their best capabilities. Moreover, as the vast majority of surveyed working neuropsychologists reported to

encourage examinees to try their best and only a small minority reported to confront examinees upon indications of noncredible performance (Martin, Schroeder, & Odland, 2015), our feedback approach followed common practice. Because little is known about the impact of feedback upon noncredible performance, and tentative statements in validity research literature suggest that feedback should be based upon objective data/facts, instead of striking a confrontational or accusatory tone (Carone et al., 2020; Martin & Schroeder, 2022; Merckelbach et al., 2015), a more direct feedback approach to PVT failure was therefore recognized as likely inappropriate.

**Data Reduction and Analysis**

Raw test scores of the standard cognitive tests were converted to *Z*-scores. These *Z*-scores of specific cognitive tasks were clustered in five cognitive domains as compound scores; *memory*, *category fluency*, *speed of information processing*, *working memory*, and *interference* (Dandachi-FitzGerald et al., 2011; van Boxtel et al., 1996; van der Elst et al., 2006b). Compound scores of related cognitive tasks were used to reduce the number of tests - and consequently reducing type I error associated with multiple comparisons - while improving the robustness of the underlying cognitive construct (Lezak, 1995). The memory compound score was calculated by averaging the *Z*-transformed scores immediate recognition on the VLT (IR_VLT) and the delayed recognition on the VLT (DR_VLT). A working memory compound score was calculated by averaging the three *Z*-transformed digit span scores (i.e., Digit Span forwards, backwards, and sequencing). A speed of information processing compound score was calculated by averaging the *Z*-transformed scores of SCWT-1 and SCWT-2. A category fluency compound score was calculated by averaging the *Z*-transformed scores on the subtests Animals and Professions. Lastly, an interference compound score was calculated by subtracting the mean transformed *Z*-score SCWT-3 from the mean of SCWT-1 and SCWT-2. The sign of the speed scores were inverted, so that positive scores reflect above average performance and negative scores reflect below average performance.[1]

The score distributions of the *Z*-transformed compound scores were used to spot outliers. We used a widely used approach (e.g., Field, 2013) for handling outliers. Instead of excluding patients with extreme scores (*Z*-score > 3.29) - since subjects who show invalid responding might also show extreme performance on the other measures used in our study -, we replaced outliers by the next highest (or lowest) score that is *not* an outlier. Two patients (1.0%) had an extreme outlier on one of the compound scores.

Data were analyzed in several steps. First, to examine group (FB vs. NO-FB) differences in demographics, clinical characteristic, and baseline test scores, independent *t*-testing, or chi-squared testing were employed. Second, Fisher's exact tests were employed to examine the proportion of participants who failed one or both of the PVTs in the FB group versus NO-FB group. Third, to explore a potential interaction effect between group (FB vs. NO-FB) and subsequent performance on standard

1 The following formulas were used:
Memory = (ZVLT_IR + ZVLT_DR) / 2
Speed of information processing = - (ZSCWT1 + ZSCWT2) / 2
Working Memory: (ZDSForward + ZDSBackward +ZDSequencing) / 3
Category fluency: (ZAnimals + ZProfessions) / 2
Interference score: - (ZSCWT-3 – (ZSCWT-1 + ZSCWT-2) / 2)

cognitive tests, we conducted two repeated measures ANOVAs (RM-ANOVAs). The first RM-ANOVA examined the re-administered cognitive tests and included a 2 (FB vs. NO-FB) x 2 (baseline vs. follow-up) x 2 (working memory and category fluency) analysis. The second RM-ANOVA examined the single-administered cognitive tests and involved a 2 (FB vs. NO-FB) x 3 (memory, speed of information processing, and interference) RM-ANOVA. Cases with missing data were excluded.

Analyses were performed with SPSS version 27.0 for Mac. Alpha level was set at $p < .05$ (two-tailed).

## RESULTS

Of the 71 participants who failed the ASTM at baseline, 39 were randomly allocated to the feedback condition (FB) and 32 to the no-feedback (NO-FB) condition. Table 1 provides an overview of the demographics, clinical characteristics, and test results at baseline for these two groups. There were no statistically significant differences between the FB group vs. NO-FB group for any of these measures. There were missing data: For one participant in the FB group, the ASTM was not repeated and, in both groups two participants were not provided with the WMT (due to temporary computer malfunctioning).

**Table 1**

*Demographics, clinical Characteristics, and Test Results for the FB Group and NO-FB Group at Baseline*

| | **Randomized upon PVT failure** | | |
|---|---|---|---|
| **Variable Name** | **FB** | **NO-FB** | ***p*-value** |
| | **(*n* = 39)** | **(*n* = 32)** | |
| **Age in years, mean (SD)** | 53.26 (12.06) | 54.25 (11.64) | .727[a] |
| **Education, *n* (%)[#]** | | | .233[b] |
| Low | - | - | |
| Medium | 29 (74.36) | 28 (87.50) | |
| High | 10 (25.64) | 4 (12.50) | |
| **Female, *n* (%)** | 14 (35.89) | 14 (43.75) | .626[b] |
| **Employment status, *n* (%)** | | | .261[b] |
| Paid employment | 12 (30.77) | 5 (15.62) | |
| Sick leave | 17 (43.59) | 17 (53.12) | |
| Incapacitated for work (full or partial) | 3 (7.70) | 3 (9.38) | |
| Unemployed | 1 (2.56) | 2 (6.25) | |
| Pension | 6 (15.38) | 3 (9.38) | |
| Missing | - | 2 (6.25) | |
| **Receiving benefits (sickness or disability), *n* (%)** | | | .477[b] |
| No | 21 (53.85) | 14 (43.75) | |
| Yes (full or partial) | 18 (46.15) | 18 (56.25) | |
| **Currently involved in legal proceedings, *n* (%)** | | | .257[b] |
| No | 25 (64.10) | 24 (75.00) | |
| Yes | 12 (30.77) | 5 (15.62) | |
| Missing | 2 (5.13) | 3 (9.38) | |

| | | | |
|---|---|---|---|
| **Diagnostic categories, *n*** | | | |
| MCI | 1 | 1 | |
| CVA | 5 | 5 | |
| TBI (moderate-severe) | 4 | 1 | |
| Neurodegenerative disease (e.g., MS, Parkinson's disease) | 1 | 2 | |
| Medically unexplained symptoms (e.g., fibromyalgia, chronic fatigue syndrome) | 1 | - | |
| Cognitive complaints not further specified | 9 | 9 | |
| Psychological condition | 1 | 3 | |
| Other (unspecified) | 5 | 7 | |
| Missing | 12 | 4 | |
| Heterogeneous "neurological condition" (i.e., MCI, CVA, moderate-severe TBI, neurodegeneration), *n* (%) | 11 (28.20) | 9 (28.12) | .781[b] |
| Heterogeneous "without (known) medical condition" (i.e., MUS, cognitive complaints, psychological condition), *n* (%) | 11 (28.20) | 12 (37.50) | .100[b] |
| **Baseline test results** | | | |
| ASTM, median (IQR) | 80.00 (6.00) | 79.00 (7.75) | .349[c] |
| Digit Span forward, mean (SD) | 6.87 (1.64) | 7.06 (1.98) | .659[a] |
| Digit Span backward, mean (SD) | 6.85 (1.88) | 6.53 (1.39) | .435[a] |
| Digit Span sequencing, mean (SD) | 6.33 (2.14) | 5.75 (2.20) | .263[a] |
| Fluency animals, mean (SD) | 18.51 (4.95) | 16.88 (5.19) | .180[a] |
| Fluency professions, mean (SD) | 13.08 (3.64) | 13.31 (4.67) | .812[a] |

*Note*. *p*-values were derived from independent samples *t*-tests[a], Fisher's exact tests[b] (Fisher's exact tests were also employed to examine (a) paid employment status yes/no and (b) currently involved in legal proceedings yes/no, instead of differences between all categories of these variables), and Mann-Whitney *U* tests[c]. [#] Education was quantified with an 8-point scale that ranges from primary school (1; fewer than six years of education) to university degree (8; 16 years of education or more) (De Bie, 1987). Three groups of education levels were identified: Low (those with a primary education at most), Medium (those with junior vocational training at most), and High (those with senior vocational or academic training). Abbreviations: ASTM = Amsterdam short-term memory test total score; CVA = cerebrovascular accident; FB = feedback condition; MCI = mild cognitive impairment; MS = multiple sclerosis; MUS = medically unexplained symptoms; NO-FB = no feedback condition; TBI = traumatic brain injury; IQR = interquartile range.



### Impact of Feedback on PVT Performance

In total, 50 of the 70 participants (71.43%), failed the ASTM upon re-administration (i.e., scored below cutoff), with no significant differences between the proportion of participants in the FB group (*n* = 27/38; 71.05%) and NO-FB group (*n* = 23/32; 71.88%), (Fisher's exact test, *p* = 1.00). The subsequently administered WMT was failed by more than half of all participants (*n* = 39/67; 58.21%), with no significant differences between the proportion of participants in the FB group (*n* = 19/37; 51.35%) and the NO-FB group (*n* = 20/30; 66.67%), (Fisher's exact test, *p* = .225). The vast majority (*n* = 56/70; 80.00%) of all participants failed one of the two PVTs, with no significant differences between the proportion of participants in the FB group (*n* = 30/38; 78.9%) and NO-FB group (*n* = 26/32; 81.25%), (Fisher's exact test,

$p = 1.00$). Moreover, almost half of all participants ($n = 30/67$; 44.78%) failed both the repeated ASTM and WMT, with comparable failure rates in the FB group ($n = 15/37$; 40.54%) and NO-FB group ($n = 15/30$; 50.00%), (Fisher's exact test, $p = .469$). Finally, a minority ($n = 14/67$; 20.89%) of all participants passed both the repeated ASTM and WMT, with comparable passing rates in the FB group ($n = 10/37$; 27.03%) and NO-FB group ($n = 4/30$, 13.33%), (Fisher's exact test, $p = .231$).

**Impact of Feedback on Standard Cognitive Test Performance**

A repeated measures ANOVA with group (FB vs. NO-FB) as between-subjects factor and time (baseline vs. follow-up) as within-subjects factor for the compound scores category fluency and working memory found no statistically significant interaction effect ($F(1, 68) = 1.47$, $p = .238$, $\eta_p^2 = .041$). There were also no main effects for group ($F(1, 68) = .547$, $p = .581$, $\eta_p^2 = .016$) and time ($F(1, 68) = 1.660$, $p = .198$, $\eta_p^2 = .047$). These results indicate that performance on the repeated compound scores were not affected by the feedback intervention nor time. The latter indicates that there is no significant improved performance between the first and second administration for both the category fluency compound score and working memory compound score.

A repeated measures ANOVA with group (FB vs NO-FB) as between-subjects factor and compound scores memory, speed of information processing, and interference (based upon the clinical measures administered during follow-up) as repeated measures found no statistically significant interaction effect ($F(2, 68) = .116$, $p = .891$ $\eta_p^2 = .003$). There also was no statistically significant main effect of group, ($F(1, 69) = 1.633$, $p = .206$, $\eta_p^2 = .023$). There was a main effect of compound scores ($F(2, 68) = 3.788$, $p = .028$, $\eta_p^2 = .100$). These results indicate that there were no statistically significant differences between the FB group and NO-FB group on these compound scores.

Repeated measures ANOVAs were also performed after excluding those participants involved in legal proceedings and found comparable results (data not shown). This indicates that involvement in legal proceedings is not related to the mentioned null findings.

## DISCUSSION

This study investigated whether an immediately feedback intervention upon PVT failure would increase performance during the neuropsychological assessment in adult clinical patients using an experimental design. Results were not in line with our main hypotheses, as performance on a repeated PVT and repeated standard cognitive tests did not differ between participants who were allocated to a feedback intervention (i.e., FB group) and those for whom invalid performance was left unaddressed (i.e., NO-FB group). We also found no significant group (FB vs. NO-FB) differences on a subsequently (single-administered) PVT and standard cognitive tests. In fact, PVT failure was found to persist during the neuropsychological assessment, as the vast majority of participants failed at least one subsequently administered PVT –independent of feedback.

These findings challenge the seemingly (partly) positive feedback intervention results on a memory test and a PVT in adult clinical patients with MS (Suchy et al., 2012) and on a measure of

information processing in adult clinical patients with CFS (Roor et al., 2020). These promising results may be explained by omissions in their study designs. First, Suchy et al. (2012) failed to use a control group where the same tests were repeated but without proving feedback on invalid performance. Therefore, instead of an intervention effect, improvements in test performance may also be explained by the repeated administration itself. And second, although Roor et al. (2020) used a control group where the same tests were repeated, there was a significant difference in time-interval between the feedback group compared to the no-feedback group that may have impact their findings -independent of the feedback intervention.

Although performing below the best of capabilities does not overlap with non-credible symptom reporting, both concepts of symptom- and performance validity are related to the broader construct of symptom exaggeration (e.g., Dandachi-FitzGerald et al., 2011). Therefore, the impact of feedback interventions targeting non-credible symptom reporting may also provide insights that are potentially relevant for patients who perform below best of capabilities. To the best of our knowledge, only Merckelbach et al. (2015) experimentally examined the effects of corrective feedback, but only for subsequent symptom reporting. In their study, participants read a case vignette and were given the option to exaggerate symptoms, and subsequently received a self-report symptom validity test (SVT). One group was provided with feedback about their symptom exaggeration on the SVT in a neutral manner, whereas the other group was given feedback in a sympathetic way. The authors found no effects of both types of corrective feedback on the level of symptom reporting, as it remained significantly higher compared to controls. Their findings are in line with the current study null findings, in which we also used an experimental design and a control condition, but in a routine clinical context using test performance as an outcome instead.

How can the persistence of PVT failure in this sample of clinical patients without obvious cognitive impairment and with medium to high education levels who were provided with immediate feedback to increase their performance be explained? First, the non-directive/neutral feedback intervention itself may have been too weak to evoke increased efforts in the feedback group. Second, potential explanations of PVT failure itself may also be considered to help understand the absence of a feedback intervention effect. Whereas the underlying mechanisms for PVT failure are not yet fully understood (but *see* Dandachi-FitzGerald et al., 2022 for proposed explanatory levels of poor symptom validity), empirical research (Roor et al., 2023) and survey results (Martin & Schroeder, 2021) indicate that incentives are associated with PVT failure in a significant number of clinical cases. In the current study sample, almost half of all participants were receiving sickness or disability benefits, and a substantial minority was currently involved in legal proceedings. Therefore, it could be argued that these incentives prohibited them from benefitting from a feedback intervention, as improving on cognitive tests would not be in their interest of demonstrating cognitive symptoms for, for example, medico-legal settlement or disability payments. However, no significant differences were found in the presence of potential external gain incentives between the proportion of participants who failed at least one subsequently administered PVT and those who passed both subsequently administered PVTs. Moreover, after excluding those participants involved in legal proceedings, comparable results were found showing no group (FB vs. NO-FB) differences for single- and repeated administered tests. However, this does not rule out the

option that participants who are afraid of not getting proper treatment (i.e., another, intrinsic form of incentive), would not improve their performance upon the feedback intervention. Besides, there are indications that invalid responding is not passive nor an end state. Merckelbach et al. (2011) found that subjects who were initially instructed to feign symptoms but at re-testing are told to respond honestly, continued to endorse higher symptom levels compared to subjects who were initially instructed to respond honestly. This finding led these authors to the suggestion that initial symptom overreporting produces a residual effect. Moreover, as already mentioned, Merckelbach et al. (2015) found subjects to persist with increased symptom reporting, despite corrective feedback. In light of these studies, our null findings may not be the result of weaknesses in the feedback intervention per se, but could (also) be impacted by the undermining and persistent effects of engaging in sub-optimal test-taking behavior. This is also supported by the absence of any improvement on the repeated standard cognitive tests. Even practice effects were absent, whereas on category fluency practice effects are well-documented (e.g., Wilson et al., 2000).

Strengths of this study are its experimental design, strict in- and exclusion criteria to prevent false-positive PVT results, and the clinical procedure used. The study procedure and approach to increasing the patients' effort to perform at true capabilities was specifically chosen to stay close to common clinical practice and recommendations for dealing with PVT failure in routine clinical care. First, although this procedure was not systematically registered in this study, in line with common practice (Martin & Schroeder, 2015), most patients were by default encouraged to do their best by the clinician (neuropsychologist) and/or technician before starting the assessment. Second, recommendations to continue testing after PVT failure (Suhr, 2012) and to use continuous sampling of performance validity during the neuropsychological examination (Boone, 2009) were followed. Finally, we chose to address the topic of invalid performance evidenced from the ongoing assessment (i.e., PVT failure) immediately to the patient, in line with common clinical practice (Dandachi-FitzGerald et al., 2013; Hirst et al., 2017).

Study limitations should also be noted. First, we chose not to specifically mention that PVTs were to be administered during the neuropsychological assessment. Although this approach is in line with current practices and recommendations, such a direct statement (as employed by Suchy et al., 2012) may alert patients about their necessity to put forth their best of efforts. Relatedly, in our study the technician delivered the immediate feedback intervention. This may have had less impact than when a separate immediate intervention was provided by the neuropsychologist. For example, in the study by Suchy et al (2012), the neuropsychologist was called into the examination room by the technician and provided the feedback. Calling the neuropsychologist in the room directly following PVT failure, can be considered an intervention in itself, irrespective of the subsequent verbal instructions and may thus be examined in future studies. Second, as forementioned, we used a brief non-accusatory/neutral feedback approach. This approach using non-adversarial language likely fits best with our current knowledge (Martin & Schroeder, 2022; Moore et al. 2021; Sweet et al., 2021) and practice (Martin & Schroeder, 2021) on how to communicate invalid performance to clinical patients. Although this feedback approach may not have been clear enough in itself to elicit improvements in test-taking behavior, it is important to stress that this intervention was communicated to the participants after the neuropsychologist and technician emphasized to perform at best of their capabilities. However, a more direct immediate

feedback approach to PVT failure is worth exploring in future studies. For example, by clearly stating that test results indicate that the patient is not performing to its full potential and improvements are expected (*see* Roor et al., 2020). Third, there is a likely difference between improvement in performance and normalization of performance with feedback. For example, Suchy et al (2012) found that, after corrective feedback upon indications of invalid performance, improvements on a standard memory test (i.e., WMS-III) showed no complete normalization for all subjects (i.e., scored in the same range as patients without indications of invalid performance). Therefore, on the one hand, feedback may inadvertently be creating a more sophisticated underperformer (rather than normalizing their performance overall). On the other hand, feedback may improve test performance to the point that frank impairment can be ruled out – which is the primary goal of neuropsychological assessment –, whereas low test scores may still be largely uninterpretable. However, such post-feedback increased test performance to "broadly normal ranges" may still represent an underestimation of true cognitive capabilities due to persistent (but somewhat decreased) performance below best of capabilities. Fourth, although the focus on limiting false-positives (i.e., by using a lower PVT cutoff with specificity > 90%, excluding patients with intellectual disability, low formal schooling, or 'obvious cognitive symptoms') is justifiable given that the main purpose of the study was to examine immediate feedback in patients with a high likelihood of invalid performance, this comes with a cost. Due to their inverse relationships, an increase in specificity by definition leads to in decrease in sensitivity (i.e., increase of false-negatives). As a logical consequence, patients who performed invalidly but were left undetected by the initial ASTM, are not represented in the current clinical sample. Therefore, our study findings cannot be generalized to all clinical patients. Lastly, instead of determining invalid performance based on failing one PVT, the usage of multiple, ideally non-correlated PVTs may be used in future studies to further reduce the risk of false-positive classifications (Sweet et al., 2021).

Improving the quality of neuropsychological test performance is a crucial and important step for neuropsychologists. Therefore, future research may want to examine additional ways that can support them in preventing and/or managing invalid performance. A pre-assessment intervention on general test-taking attitude and patient expectations of the assessment and its potential outcomes may be worthwhile to test. In addition, contrasting various and more extreme types of immediate feedback (sympathetic, neutral, and confrontational) in experimental studies are needed because of the little existing research.

It is important to acknowledge that these suggested venues for future research assume that invalid performance can potentially be re-directed through the right intervention. However, when performance below actual capabilities is persistent (Merckelbach et al., 2011), and feedback interventions have little to no clinical impact (Merckelbach et al., 2015), we should also look into alternatives. A radically different approach for dealing with invalid performance, is to discontinue the assessment and start with treatment instead. For example, Jurick et al. (2020) found that veterans who failed a PVT exhibited a clinically meaningful reduction in symptoms of PTSD, depression, and post-concussive symptoms after treatment. Strikingly, their PVT performance also increased after treatment was completed, leading to significantly less subjects who failed a PVT. Therefore, these authors state that 'veterans with invalid PVTs

should be enrolled in trauma-focused treatment and may benefit from neuropsychological assessment after, rather than before, treatment' (p. 108).

In conclusion, this is the first multisite single-blind RCT to examine the potential feedback effect on subsequent neuropsychological test performance in clinical patients who evidenced invalid performance. Our findings expand on two previous observational studies by showing that a brief, neutral feedback intervention following proposed clinical recommendations and common practices had no effects on subsequently PVT performance and standard cognitive tests. In fact, the vast majority of the participants continued to fail at least one subsequently administered PVT -independent of feedback. These results suggest that there might be limitations to using immediate feedback upon indications of invalid performance for increasing patients' efforts to perform at the best of their capabilities. Therefore, more research is needed to examine various forms of feedback and other approaches to manage patients' test-taking efforts before, during, or after the neuropsychological assessment. This is important, as it can contribute to increasing the overall quality of neuropsychological outcomes and therewith improve appropriate diagnostic conclusions and recommendations for treatment.

## REFERENCES


Alverson, A. W., O'Rourke, J. J. & Soble, J. R. (2019). The Word Memory Test genuine memory impairment profile discriminates genuine memory impairment from invalid performance in a mixed clinical sample with cognitive impairment. *The Clinical Neuropsychologist, 33*(8), 1420-1435, https://doi:10.1080/13854046.2019.1599071

American Psychological Association. (2017). *Ethical principles of psychologists and code of conduct* (2002, amended effective June 1, 2010, and January 1, 2017). https://www.apa.org/ethics/code/

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

Brand, N. & Jolles, J. (1985). Learning and retrieval rate of words presented auditory and visually. *Journal of General Psychology*, *112*(2), 201–210, https://doi: 10.1080/00221309.1985.9711004

Carone, D. A., Iverson, G. L., & Bush, S. S. (2010). A model to approaching and providing feedback to patients regarding invalid test performance in clinical neuropsychological evaluations. *The Clinical Neuropsychologist, 24*(5), 759–778, https://doi: 10.1080/13854041003712951

Dandachi-FitzGerald, B., Merckelbach, H., & Merten, T. (2022). Cry for help as a root cause of poor symptom validity: A critical note. *Applied neuropsychology. Adult*, 1–6 [Advance online publication], https://doi.org/10.1080/23279095.2022.2040025

Dandachi-FitzGerald, B., Merckelbach, H., & Ponds, R. W. H. M. (2017). Neuropsychologists' ability to predict distorted symptom presentation. *Journal of Clinical and Experimental Neuropsychology, 39*(3), 257–264, https://doi.org/10.1080/13803395.2016.1223278

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.mu.idm.oclc.org/10.1093/arclin/act073

Dandachi-FitzGerald, B., Ponds, R. W. H. M., Peters, M. J. V., & Merckelbach, H. (2011). Cognitive underperformance and symptom over-reporting in a mixed psychiatric sample. *The Clinical Neuropsychologist, 25*(5), 812–828, https://doi.org/10.1080/13854046.2011.583280

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). SAGE Publications.

Green, P. (2003). *Manual for the Word Memory Test.* Edmonton, Alberta, Canada: Green's Publishing.

Greve, K. W., Ord, J., Curtis, K. L., Bianchini, K. J. & Brennan, A. (2008). Detecting malingering in traumatic brain injury and chronic pain: A comparison of three forced-choice symptom validity tests. *The Clinical Neuropsychologist, 22*(5), 896-918, https://doi.org/10.1080/13854040701565208

Hammes, J. (1973). *De Stroop Kleur-woord Test: Handleiding [The Stroop Color-Word Test: Manual].* Amsterdam: Swets & Zeitlinger.

Hirst, R. B., Han, C. S., Teague, A. M., Rosen, A. S., Gretler, J., & Quittner, Z. (2017). Adherence to validity testing recommendations in neuropsychological assessment: A survey of INS and NAN members. *Archives of Clinical Neuropsychology, 32*(4), 456–471, https://doi.org/10.1093/arclin/acx009

Jurick, S. M., Crocker, L. D., Merritt, V. C., Hoffman, S. N., Keller, A. V., Eglit, G. M. L., Thomas, K. R., Norman, S. B., Schiehser, D. M., Rodgers, C. S., Twamley, E. W., & Jak, A. J. (2020). Psychological symptoms and rates of performance validity improve following trauma-focused treatment in veterans with PTSD and history of mild-to-moderate TBI. *Journal of the International Neuropsychological Society, 26*(1), 108–118. https://doi.org/10.1017/S1355617719000997

Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). Oxford University Press.

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

Longley, W. A., Tate, R. L., & Brown, R. F. (2022). The psychological benefits of neuropsychological assessment feedback as a psycho-educational therapeutic intervention: A randomized-controlled trial with cross-over in multiple sclerosis. *Neuropsychological rehabilitation*, 1–30 [Advance online publication], https://doi.org/10.1080/09602011.2022.2047734

Luteijn, F., & Barelds, D. P. (2004). Revisie van de Groninger Intelligentie Test (GIT) [revision of the Groningen Intelligence Test]. *Diagnostiek-wijzer*, *3*, 114 - 120.

Martin P. K. & Schroeder, R. W. (2021). Feedback with patients who produce invalid testing: Professional values and reported practices. *The Clinical Neuropsychologist, 35*(6), 1134-1153, https://doi:10.1080/13854046.2020.1722243

Martin**,** P. K., & Schroeder, R. W. (2022). A framework for providing clinical feedback when patients invalidate testing. In R. W. Schroeder, & P. K. Martin (Eds), *Validity assessment in clinical neuropsychological practice: Evaluating and managing noncredible performance* (pp. 47-69). The Guilford Press.

Martin, P. K., Schroeder, R. W., & Odland, A. P. (2015). Neuropsychologists' validity testing beliefs and practices: A survey of north American professionals. *The Clinical Neuropsychologist, 29*(6), 741-776, https://doi:10.1080/13854046.2015.1087597

Merckelbach, H., Dandachi-FitzGerald, B., van Mulken, P., Ponds, R., W. H. M. & Niesten, E. (2015). Exaggerating psychopathology produces residual effects that are resistant to corrective feedback: an experimental demonstration. *Applied neuropsychology. Adult*, *22*(1), 16–22, https://doi.org/10.1080/23279095.2013.816850

Merckelbach, H., Jelicic, M., & Pieters, M. (2011). The residual effect of feigning: how intentional faking may evolve into a less conscious form of symptom reporting. *Journal of Clinical and Experimental Neuropsychology*, *33*(1), 131–139, https://doi-org.mu.idm.oclc.org/10.1080/13803395.2010.495055

Moore, P., Bunnage, M., Kemp, S., Dorris, L., & Baker, G. (2021). *Guidance on the assessment of performance validity in neuropsychological assessment*. The British Psychological Society.

Postal, K., & Armstrong, K. (2013). *Feedback that sticks: The art of effectively communicating neuropsychological assessment results.* Oxford University Press.

Roor, J. J., Dandachi-FitzGerald, B., Peters, M. J. V., & Ponds, R. W. H. M. (2023). Performance validity test failure in the clinical population: A systematic review and meta-analysis of prevalence rates. *Neuropsychology Review* [Advance online publication], https://doi: 10.1007/s11065-023-09582-7

Roor, J. J., Knoop, H., Dandachi-FitzGerald, B., Peters, M. J. V., Bleijenberg, G. & Ponds, R. W. H. M. (2020). Feedback on underperformance in patients with chronic fatigue syndrome: The impact on subsequent neuropsychological test performance. *Applied Neuropsychology: Adult, 27*(2), 188-196, https://doi:10.1080/23279095.2018.1519509

Schmand, B. & Lindeboom, J. (2005). *Amsterdam Short-Term Memory Test: Manual*. Leiden, The Netherlands: Psychologische Instrumenten, Tests en Services.

Schmand, B., de Sterke, S., & Lindeboom, J. (1999). *Amsterdamse Korte Termijn Geheugen test [Amsterdam Short-Term Memory Test*]. Lisse, NL: Swets & Zeitlinger

Stevens, A. & Licha, C. (2019). The Word Memory Test in medicolegal assessment: A measure of effort and malingering? *Journal of Forensic Psychiatry and Psychology, 30*(2), 220-249, https://doi.org/10.1080/14789949.2018.1539509

Suchy, Y., Chelune, G., Franchow, E. I., & Thorgusen, S. R. (2012). Confronting patients about insufficient effort: The impact on subsequent symptom validity and memory performance. *The Clinical Neuropsychologist, 26*(8), 1296–1311, https://doi: 10.1080/13854046.2012.722230

Suhr, J. A. (2012, February 15-18). *Strategies for addressing noncredible performance in assessment of young adults*. [Conference presentation]. Annual meeting of the International Neuropsychological Society, Montréal, Québec, Canada.

Sweet, J. J., Heilbronner, R. L., Morgan, J. E., Larrabee, G. J., Rohling, M. L., Boone, K. B., … 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

Valentijn, S. A., van Boxtel, M. P., Van Hooren, S. A., Bosma, H., Beckers, H. J., Ponds, R. W., & Jolles, J. (2005). Change in sensory functioning predicts change in cognitive functioning: Results from a 6-year follow-up in the Maastricht Aging Study. *Journal of the American Geriatrics Society, 53*(3), 374-380, https://doi.org/10.1111/j.1532-5415.2005.53152.x

Van Boxtel, M. P., Langerak, K., Houx, P. J., & Jolles, J. (1996). Self-reported physical activity, subjective health, and cognitive performance in older adults. *Experimental Aging Research*, *22*(4), 363–379, https://doi.org/10.1080/03610739608254017

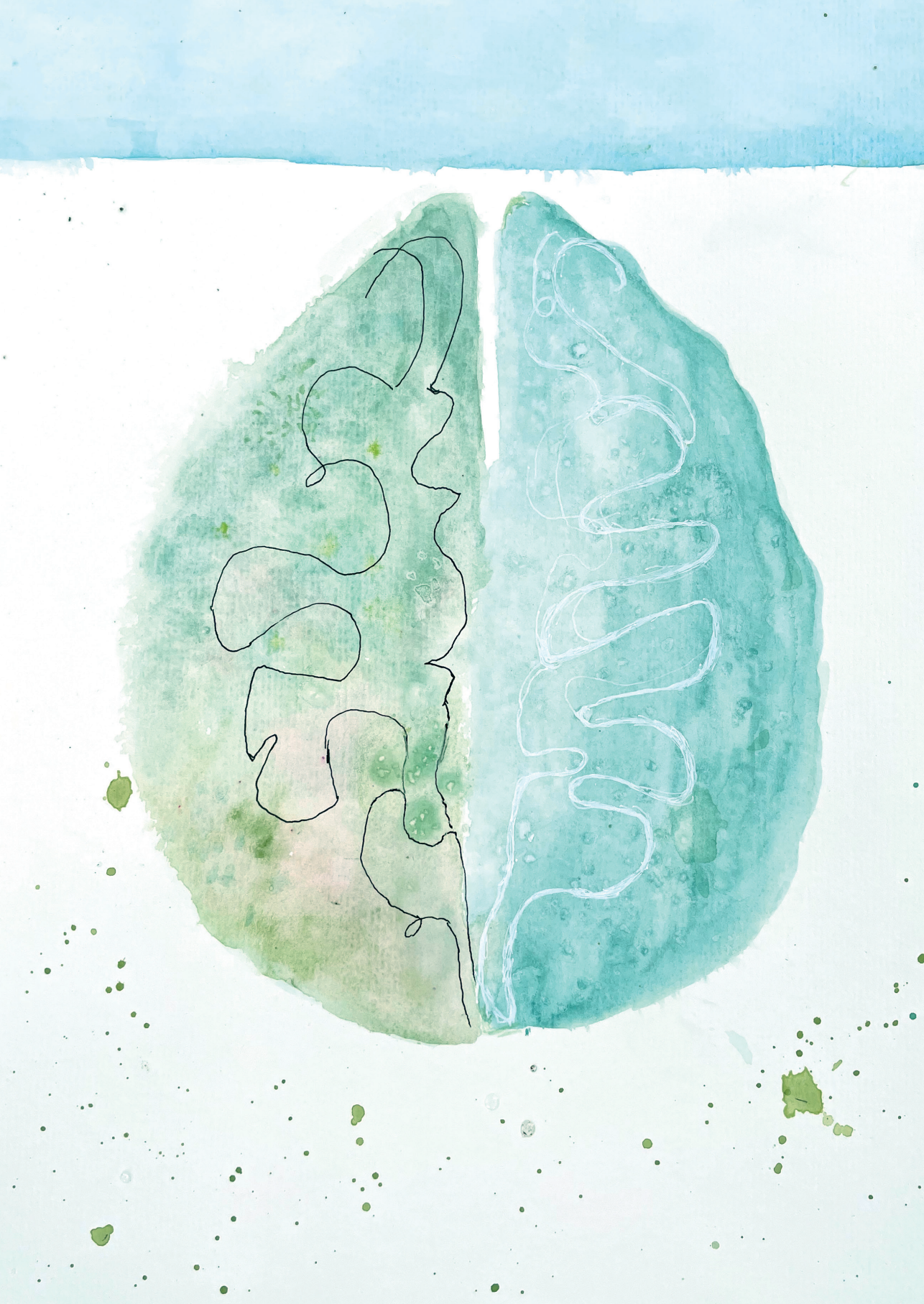
Van der Elst, W., Van Boxtel, M., Van Breukelen, G. & Jolles, J. (2005). Rey's verbal learning test: Normative data for 1855 healthy participants aged 24–81 years and the influence of age, sex, education, and mode of presentation. *Journal of the International Neuropsychological Society, 11*(3), 290-302, https://doi:10.1017/S1355617705050344

Van der Elst, W., Van Boxtel, M. P., Van Breukelen, G. J., & Jolles, J. (2006a). Normative data for the Animal, Profession and Letter M Naming verbal fluency tests for Dutch speaking participants and the effects of age, education, and sex. *Journal of the International Neuropsychological Society, 12*(1), 80–89, https://doi-org.mu.idm.oclc.org/10.1017/S1355617706060115

Van der Elst, W., Van Boxtel, M. P., Van Breukelen, G. J., & Jolles, J. (2006b). The Stroop Color-Word Test: influence of age, sex, and education; and normative data for a large sample across the adult age range. *Assessment*, *13*(1), 62–79, https://doi.org/10.1177/1073191105283427

Wilson, B., Watson, P., Baddeley, A., Emslie, H., & Evans, J. (2000). Improvement or simply practice? The effects of twenty repeated assessments on people with and without brain injury. *Journal of the International Neuropsychological Society, 6*(4), 469-479, https://doi:10.1017/S1355617700644053

Wechsler D. (2008). *Wechsler Adult Intelligence Scale–Fourth edition: Technical and interpretive manual*. San Antonio, TX: Pearson Assessment.



# CHAPTER 6

## Performance validity and outcome of cognitive behavior therapy in patients with chronic fatigue syndrome

Roor, J. J., Dandachi-FitzGerald, B., Peters, M. J., Knoop, H., & Ponds, R. W. (2022).
Performance validity and outcome of cognitive behavior therapy in
patients with chronic fatigue syndrome.
*Journal of the International Neuropsychological Society, 28*(5), 473–482

## ABSTRACT

### Objective

There is limited research examining the impact of the validity of cognitive test performance on treatment outcome. All known studies to date have operationalized performance validity dichotomously, leading to the loss of predictive information. Using the range of scores on a performance validity test (PVT), we hypothesized that lower performance at baseline was related to a worse treatment outcome following cognitive behavioral therapy (CBT) in patients with chronic fatigue syndrome (CFS) and to lower adherence to treatment.

### Method

Archival data of 1081 outpatients treated with CBT for CFS were used in this study. At baseline, all patients were assessed with a PVT, the Amsterdam Short-Term Memory test (ASTM). Questionnaires assessing fatigue, physical disabilities, psychological distress, and level of functional impairment were administered before and after CBT.

### Results

Our main hypothesis was not confirmed: the total ASTM score was not significantly associated with outcomes at follow-up. However, patients with a missing follow-up assessment had a lower ASTM performance at baseline, reported higher levels of physical limitations, and completed fewer therapy sessions.

### Conclusions

CFS patients who scored low on the ASTM during baseline assessment are more likely to complete fewer therapy sessions and not to complete follow-up assessment, indicative of limited adherence to treatment. However, if these patients were retained in the intervention, their response to CBT for CFS was comparable with subjects who score high on the ASTM. This finding calls for more research to better understand the impact of performance validity on engagement with treatment and outcomes.

## INTRODUCTION

The frequency of performance validity test (PVT) failure is substantial in non-forensic clinical settings (Dandachi-FitzGerald et al., 2016; Martin & Schroeder, 2020), and its impact on neuropsychological test performance is known to be as large or even greater than various medical and psychiatric conditions (Iverson, 2006; Sollman & Berry, 2011). PVT failure invalidates cognitive test results and hinders the clinician in making adequate diagnoses about cognitive (dis)functioning and, consequently, recommendations for treatment. One might, therefore, anticipate that the impact of performance invalidity is not limited to the diagnostic assessment, but extends to treatment efficacy.

The notion that PVT failure is relevant beyond the diagnostic domain and may also be related to everyday functioning has received some consideration. For example, Lippa et al. (2014) found that PVT failure is related to self-reported community participation in veterans with mild traumatic brain injury. Although research on this topic is limited, performance validity may serve as a behavioral proxy of how a patient copes with everyday life, and as such may convey potentially relevant information for treatment planning, adherence, and outcome.

To the best of our knowledge, only five studies have examined the relationship between performance validity and treatment. Moore et al. (2013) found that, in comparison with patients who passed a PVT, patients with schizophrenia and schizoaffective disorder ($n$ = 128) who failed a PVT had significantly lower therapy attendance -which is known to negatively affect treatment outcome (Long et al., 2012). Psychiatric symptoms or cognitive impairment did not predict group therapy attendance in this study. Horner et al. (2014) studied the relationship between invalid performance and healthcare utilization in a heterogeneous outpatient sample of a Veterans Affairs Medical Center ($n$ = 355). They found that PVT failure in these patients was associated with increased and longer inpatient hospitalizations and more emergency department visits. A recent study by Jurick and colleagues (2020) of veterans with posttraumatic stress disorder (PTSD) and mild-to-moderate traumatic brain injury ($n$ = 100) found that both subjects who passed and those who failed a PVT benefitted from treatment, although valid performers showed the greatest reduction in PTSD symptoms. No significant differences were found in treatment completion between patients who passed and those who failed a PVT (i.e., 57.9% and 46.5%, respectively). Williams and colleagues (2020) found that veterans ($n$ = 61) benefitted equally from PTSD treatment, regardless of PVT failure. Goedendorp et al. (2013) examined whether invalid performance at baseline assessment was related to treatment outcome in patients with chronic fatigue syndrome (CFS; $n$ = 169). These authors found a higher loss to follow-up (i.e., missing follow-up assessment) in CFS patients who failed a PVT (i.e., 23%), in comparison with patients who passed the PVT (i.e., 8%). For the patients who failed a PVT, no group differences were found in comparison with the patients who passed the PVT with regard to change in fatigue severity, functional impairments, or physical limitations following cognitive behavioral therapy (CBT) for CFS. Although research on this topic is limited, the studies described suggest that invalid performance is negatively associated with the response or adherence to treatment.

It is important to emphasize that the aforementioned studies all used performance validity dichotomously; subjects who failed a PVT were classified as "invalid performers" and subjects who passed

a PVT as "valid performers." Consequently, subjects who scored just on opposite sides of the cutoff were interpreted as being very different, when in fact their PVT scores were close to each other. Subjects who scored below the cutoff were also considered similar, even when their PVT scores varied greatly. However, the positive predictive value (PPP) of invalid performance - i.e., the probability that PVT failure represents true invalid performance - is related to the severity of PVT failure. Consequently, PVT scores well below cutoff are more likely to represent true invalid performance in comparison with PVT scores at cutoff. Consequently, adhering to a dichotomous approach will evidently lead to loss of information and a decrease in the statistical power to detect a relationship between performance validity and treatment outcome (Altman & Royston, 2006).

The current study addresses this limitation. The aim was to replicate the aforementioned study of Goedendorp et al. (2013), but using the total range of scores of a PVT and a considerably larger sample size to examine the impact that performance validity has on the outcome for CFS after CBT. We hypothesized that lower PVT scores (i.e., indicating lower levels of effort to perform to the best of one's abilities) at baseline would be related to worse treatment outcome (i.e., higher levels of self-reported symptoms of CFS after CBT) and lower treatment adherence (i.e., more loss to follow-up and fewer completed treatment sessions) in comparison with patients who produced higher PVT baseline scores.

## METHOD

### Participants

Archival pre- and post-treatment data on CBT for CFS was used for this study. Patients were consecutively referred to a tertiary treatment facility for chronic fatigue in a university hospital. First, the patients' medical status was assessed by consultants of the Department of Internal Medicine, to rule out other medical explanations for their fatigue, and second to scan for the potential need for additional medical examination. This procedure followed national CFS guidelines (Centraal Begeleidings Orgaan, 2013), which are in accord with the US Centers for Disease Control and Prevention (CDC) guidelines formulated in 2003 (Reeves et al., 2003). If patients met the CDC criteria for CFS, they were referred to the treatment center. All the patients in this sample were seeking treatment for CFS and were seen in the context of routine clinical care. Since it is known that being involved in a legal procedure with respect to disability claims is related to poor treatment outcome of CBT for CFS (Prins, et al., 2001), patients who were engaged in a disability claim were excluded from starting treatment.

Patients were included in this study if they were severely fatigued (i.e., scored 35 or higher on the fatigue severity subscale of the Checklist Individual Strength (CIS) questionnaire (Worm-Smeitink et al., 2017), and had significant functional impairments in daily life (i.e., had a weighted total score ≥ 700 on the Sickness Impact Profile 8 (SIP8) (Jacobs et al., 1990). Additional inclusion criteria were (1) Dutch language proficiency and (2) being 18 years or older. The data were collected between April 2007 and April 2015 in the context of treatment (i.e., CBT for CFS). The questionnaires and tests used in this study were part of the routine clinical assessment. It was standard practice that all patients completed the Amsterdam Short-Term Memory test (ASTM; Schmand & Lindeboom, 2005) at baseline.

The medical ethics committee of Radboud University Medical Centre approved this study. This research was conducted in accordance with the Helsinki declaration.

None of the participants from the Goedendorp et al. (2013) study are represented in the current sample. In the current study, the same clinical procedure, diagnostic criteria for CFS, and inclusion criteria (i.e., score of 35 or higher on the fatigue severity subscale of the Checklist Individual Strength questionnaire and a weighted total score ≥ 700 on the Sickness Impact Profile 8) were used, as in the study of Goedendorp et al. (2013).

**Procedures**

Before CBT treatment, a neuropsychological assessment was conducted, consisting of a clinical interview by a psychologist, followed by the administration of tests and questionnaires by a test assistant (*see* Instruments). After CBT was completed, a follow-up assessment was conducted, in which only the questionnaires were re-administered. Patients were invited for a follow-up assessment with a test assistant, separately from the last treatment session. If they did not respond to the initial invitation, they were contacted multiple times by telephone.

**Interventions**

Individual and group face-to-face CBT for CFS was provided according to a published treatment protocol (Knoop & Bleijenberg, 2010). The protocol is based on a model of cognitive behavioral fatigue-perpetuating factors (Knoop et al., 2010). The aim of CBT is to reduce fatigue and disabilities by changing fatigue-related cognitions and behaviors. CBT for CFS consists of about 12 to 14 sessions during a 6-month period.

**Measures**

Education was assessed by self-report, classifying formal schooling on an 8-point scale often used in the Netherlands (de Bie, 1987). Based upon Van der Elst et al. (2005), three groups of education level were formed: low (those with primary education at most), medium (those with junior vocational training at most) and high (those with senior vocational or academic training).

Performance validity was measured with the Amsterdam Short-Term Memory test (ASTM). The ASTM is a 30-trial forced-choice word recognition procedure. The total calculated score is used as a cutoff for invalid performance. In the original validation studies, a cutoff score lower than 84 was associated with a specificity of 93% and a sensitivity of 84% in discriminating experimental malingerers and a heterogeneous neurological patient group. The internal consistency was found to be excellent (Cronbach's $\alpha = 0.91$) (Schmand & Lindeboom, 2005).

The total score of the revised Dutch-language version of the Symptom Checklist (SCL-90) was used to measure psychological distress, and the Depression subscale (16 items) was used to measure symptoms of depression (Arrindell & Ettema, 2005; Derogatis, 1994). All items are rated on a 5-point Likert scale, ranging from "not at all" (0) to "extremely" (4). Reliability and validity of the revised Dutch-language version of the SCL-90 are qualified as good (Arrindell, et al., 2003).

Fatigue during the past two weeks was assessed with the fatigue severity subscale of the Checklist Individual Strength (CIS) questionnaire (Worm-Smeitink et al., 2017). The CIS fatigue severity subscale contains eight items, with a score range of 8-56. Higher scores indicate higher levels of fatigue. The CIS questionnaire is extensively validated for the assessment of fatigue (Worm-Smeitink et al., 2017).

Physical disabilities were measured with the physical functioning subscale of the Medical Outcomes Survey Short-Form-36 (Ware Jr & Sherbourne, 1992). Scores on this scale range from 0 to 100, with higher scores indicating fewer physical limitations. The SF-36 is a reliable and valid instrument (Scheeres et al., 2008; Ware Jr & Sherbourne, 1992).

Functional impairments in daily functioning were assessed using the Sickness Impact Profile 8 (SIP8) (Bergner, Bobbitt, Carter, & Gilson, 1981). The SIP8 total score consists of eight subscales; alertness behavior, sleep/rest, leisure activities, homemaking, work limitations, mobility, social interactions and ambulation. The eight subscales of the SIP are added to a weighted total score, with higher scores indicating more functional impairments [range 0-5799]. The SIP is a reliable instrument (Bergner et al., 1981) and has been validated for the Dutch population (Jacobs et al., 1990).

As with Goedendorp et al. (2013), loss to follow-up was determined by missing follow-up assessment after CBT for CFS. Since treatment dropout was not registered in this study, we examined the number of completed therapy sessions as a proxy of treatment adherence.

**Data Analyses**

When the amount of missing treatment outcome data is significant, it is likely that complete case analysis (CC) introduces bias and results in estimates with less precision, leading to loss of statistical power. Applying statistical methods that handle missing data appropriately is therefore advocated in reporting observational studies (von Elm et al., 2007). To this end, we used multiple imputation (MI). MI is a commonly used method for handling missing data. It has the potential to counteract the impact that CC has on the results, so that bias is reduced and precision is improved. Briefly, in MI, a model is fitted for the missing values of dependent variables. Predictor variables and auxiliary variables (i.e., variables that are not included in the final analyses but are related to variables of interest) are used for this purpose. An estimated (i.e., imputed) value is then calculated for every missing value, ensuring that these scores are near the collected scores of comparable subjects.

Following the suggested guidelines for MI reporting (Sterne et al., 2009), an imputation model with full conditional specifications was created on the assumption that data were missing at random (MAR). First, binary logistic regression analyses were conducted using the baseline measures (i.e., ASTM, CIS fatigue score, SCL-90 total score, SCL-90 depression score, SIP total score, and SF-36 physical functioning) and demographic information (sex, level of education, and age) to examine which variables predicted missing data at follow-up. We used a less strict significance level ($p < .10$) to include all potential confounding variables. The baseline measures of ASTM and SF-36 physical functioning were negatively associated with missing follow-up data. This suggests an inverse relationship; an increase on the ASTM (i.e., more effort to perform to the best of abilities) or SF36 physical functioning (i.e., reporting fewer physical limitations) was associated with a decrease in missing follow-up data. Therefore, these two variables were used to generate the imputations. Additionally, to preserve the association between

outcome measures and predictors, all follow-up variables (i.e., CIS fatigue subscale, SIP total score, SCL-90 total score, and SF36 physical functioning) were retained in the imputation model (Spratt et al., 2010). We used 25 imputations to reduce the impact that random sampling has on pooled data (Spratt et al., 2010). For all subjects ($n$ = 1081), missing follow-up values were calculated based upon the multiple imputation procedure outlined. Consequently, all analyses on treatment outcome were performed using the pooled imputed follow-up variables. Complete case (CC) analyses using non-imputed data were used for examining treatment adherence (i.e., missing follow-up assessment and number of completed therapy sessions).

Assumptions concerning linearity were assessed through visual inspection of residuals. To examine the appropriateness of the imputation model, results based on the original (non-imputed) data were compared with those based on the multiple imputations.

Hierarchical linear regression analyses were used to examine the relationship between the continuous ASTM (predictor) score and treatment outcome (criterion), controlling for potential confounding factors. Since we were specifically interested in the impact of performance validity on outcome after CBT for CFS, treatment outcome was defined by follow-up scores on a set of preferred outcome variables used in CFS-research (Janse et al., 2016): the CIS fatigue subscale, SIP total score, SCL-90 total score, and physical functioning subscale of the SF36, for which Bonferroni correction was applied (alpha = .0125). Older age in combination with depressive symptoms is associated with an increase of false-positive scores on the ASTM (Schmand & Lindeboom, 2005). In addition, low intelligence is known to negatively influence PVT performance (Lippa, 2018). Therefore, predictor variables were entered in two steps. The first step contained level of education (i.e., dummy variables Low and Medium levels of education) as a proxy of intelligence, depression (i.e., SCL-90 depression subscale), and age as predictors for treatment outcome. Step 2 included all of the above predictors and added the total range ASTM score as a predictor. Since SPSS is not able to provide pooled $R^2$ (change) data for imputed datasets, the mean $R^2$ (change) values for models 1 and 2 of the 25 imputed datasets were calculated manually. This is the preferred method for combining $R^2$ (change) across multiple imputed datasets (Van Ginkel, 2019).

Since a significant proportion of patients were lost to follow-up (i.e., did not complete follow-up measurement), we performed a secondary analysis to examine which patient characteristics were related to loss to follow-up, and whether loss to follow-up was related to the number of completed therapy sessions (as a proxy of therapy adherence). Mann-Whitney $U$ tests were used to examine differences in test and questionnaire scores administered at baseline. Differences in age, level of education, sex, and number of completed therapy sessions between subjects who completed the follow-up assessment and those who did not were examined using an independent $t$-test or Fisher's exact test as deemed appropriate.

All analyzes were performed using the Statistical Package for the Social Sciences software (SPSS), version 23.0, with $p < .05$ (two-tailed) used as the significance level for baseline analyses and $p < .01$ (Bonferroni correction) for follow-up analyses.

## RESULTS

### Sample Characteristics

A total of 1382 patients fulfilled the inclusion criteria. Only patients who had started with CBT ($n$ = 1081) were included. The variables of age, sex, and level of education were complete. All baseline measures were near to complete (missing < 1%). Data at follow-up were missing for the CIS fatigue subscale ($n$ = 222; 20.53%), SF-36 physical functioning ($n$ = 222; 20.53%), the SCL-90-R depression subscale and SCL-90-R total score ($n$ = 273; 25.25%), and for the SIP total score ($n$ = 221; 20.46%).

Table 1 provides an overview of demographics and treatment data at baseline and follow-up using the original (i.e., non-imputed) data of CFS patients provided with CBT. This sample consisted predominantly of women (75.39% female) in their thirties (mean age 36.98 years) with medium to high levels of education.

In addition, the continuous ASTM score was negatively related to all self-reported baseline measures (i.e., CIS fatigue, physical limitations, functional impairment, psychological distress, and depressive symptoms; all $p$'s < .0125).

**Table 1**

*Demographics and Treatment Data at Baseline and Follow-Up Using the Original (i.e., Non-Imputed) Data*

| | Baseline ($n$ = 1081) | | | Follow-up ($n$ = 859) | | |
|---|---|---|---|---|---|---|
| | **M** | **SD** | **%** | **M** | **SD** | **%** |
| **Patient characteristics** | | | | | | |
| Age (years) | 36.98 | (11.74) | | - | | |
| Education | | | | | | |
| Low | | | 9.53 | - | | |
| Medium | | | 59.63 | - | | |
| High | | | 30.92 | - | | |
| Female | | | 75.39 | - | | |
| **Treatment data** | | | | | | |
| Loss to follow-up | | | - | | | 20.53 |
| ASTM | 86.89 | (3.78) | | - | | |
| [a]ASTM fail (score < 84) | | | 11.38 | | | - |
| CIS fatigue | 50.51 | (5.03) | | 29.83 | (14.10) | |
| SF36 physical functioning | 57.25 | (20.33) | | 80.34 | (20.81) | |
| SIP total | 1572.31 | (551.32) | | 658.07 | (659.25) | |
| SCL-90 total | 164.56 | (38.43) | | 130.71 | (36.93) | |

*Note.* ASTM = Amsterdam short-term memory test; CIS fatigue = checklist for individual strength, fatigue subscale; SF36 physical functioning = medical outcomes survey short-form-36, physical functioning subscale; SIP total = sickness impact profile, total score; SCL-90 total = symptom checklist-90 total score.

[a] = cutoff used by Goedendorp et al. (2013)

**Hierarchical Linear Regression Analyses on Scores of Fatigue Severity, Physical Limitations, Functional Impairment, and Psychological Distress at Follow-Up**

The association between the ASTM and all outcome measures (i.e., CIS fatigue, SIP total score, SCL-90 total score, and SF36 physical subscale) were linear, based upon visual inspection of their respective residual plots.

The first model accounting for the combined explained variance in age, level of education, and depressive symptoms (i.e., SCL-90 depression subscale) on treatment outcome was significant ($p < .0125$) for all 25 imputation datasets of all criterion variables (i.e., CIS fatigue, SF-36 physical functioning subscale, SIP total score, and SCL-90 total score during follow-up; data not shown). The second model, with the added continuous ASTM score as predictor of treatment outcome, was also significant for all 25 imputed datasets of all criterion variables (data not shown). Importantly, the ASTM score added in Model 2 did not yield a significant improvement in the prediction of treatment outcome for any of 25 imputed datasets of all criterion variables on top of the predictors in Model 1 (i.e., the $R^2$ change was not significant; data not shown). Importantly, the continuous ASTM score was found not to be significantly associated with any of the follow-up scores (*see* Table 2). Furthermore, older age was found to be significantly associated with worse outcome on all follow-up scores. Higher levels of depressive symptoms (i.e., a higher SCL-90 depression score) at baseline were significantly associated with worse outcome on the CIS fatigue subscale score, SIP total score, and the SCL-90 total score after treatment. Low and medium levels of education were found to be significantly associated with a worse outcome on SF36 physical functioning. These findings were confirmed using the original (non-imputed) data, where the addition of the continuous ASTM in the regression model did not yield a significant improvement in the prediction of treatment outcome (*see* Appendix A). In addition, the association of the individual predictors with treatment outcome were comparable based on the original data, with an added significant association between higher levels of self-reported depressive symptoms and worse outcome on SF36 physical functioning (data not shown).

We chose to include depressive symptom reporting in combination with age in the regression models, since older subjects with higher levels of reported depression have a higher chance of producing false-positive ASTM scores - as mentioned in the ASTM manual. Leaving depressive symptom reporting (i.e., the SCL-90 depression subscale) out of the regression models, however, did not show different results: ASTM performance was still not significantly associated with any of the outcome measures at follow-up (i.e., CIS fatigue, the SIP total score, the SCL-90 total score, or SF-36 physical limitations).

Since in practice the ASTM is intended to be used categorically (sufficient versus insufficient performance validity), the mentioned hierarchical linear regression analyses were re-examined using the ASTM cutoff (i.e., score < 84) as a predictor instead of using its continuous score. ASTM failure was not found to be related to any of the outcome variables (data not shown).

When, in line with Goedendorp and colleagues (2013), change scores (baseline minus follow-up scores for CIS fatigue, SF36 physical functioning, the SIP total score, and SCL-90 total score) were used as criterion variables to define treatment outcome, the findings of the hierarchical linear regression analyses were replicated; the continuous ASTM score was found not to be significantly related with any of the change scores, using both the imputed datasets as well as the original data (data not shown).

**Table 2**

*Pooled Data from a Hierarchical Linear Regression Analysis Assessing the Relationship Between Level of Education, Depressive Symptoms, Age, and the Total Score Range of the Amsterdam Short-Term Memory Test at Baseline as Predictors, and Fatigue Severity, Physical Limitations, Functional Impairment, and Psychological Distress at Follow-Up as Dependent Variables*

| Step and predictor variables | $\bar{R}^2$ | B | 95% CI | | *t* | *p*-value | Effect size Cohen's $f^2$ |
|---|---|---|---|---|---|---|---|
| | | | Lower | Upper | | | |
| Subscale fatigue severity of the CIS at follow-up | | | | | | | |
| **Step 1** | .03 | | | | | | .03 |
| Constant | | 18.20 | 13.43 | 22.97 | 7.49 | < .01 | |
| Low education | | 1.34 | -2.23 | 4.90 | .74 | .46 | |
| Medium education | | .005 | -2.03 | 2.04 | .005 | .99 | |
| SCL-90, depression | | .19 | .08 | .291 | 3.53 | < .01* | |
| Age | | .16 | .07 | .241 | 3.64 | < .01* | |
| **Step 2** | .03 | | | | | | < .01 |
| Constant | | 21.22 | -2.51 | 44.96 | 1.76 | .08 | |
| Low education | | 1.25 | -2.39 | 4.90 | .68 | .50 | |
| Medium education | | -.02 | -2.07 | 2.03 | -.02 | .98 | |
| SCL-90, depression | | .18 | .08 | .29 | 3.49 | < .01* | |
| Age | | .15 | .07 | .24 | 3.61 | < .01* | |
| ASTM | | -.03 | -0.29 | .22 | -.26 | .80 | |
| Physical limitations subscale of the SF-36 at follow-up | | | | | | | |
| **Step 1** | .06 | | | | | | .07 |
| Constant | | 102.38 | 95.69 | 109.07 | 30.02 | < .01 | |
| Low education | | -8.36 | -13.62 | -3.09 | -3.12 | < .01* | |
| Medium education | | -5.11 | -8.02 | -2.19 | -3.44 | < .01* | |
| SCL-90, depression | | -.18 | -.33 | -.03 | -.25 | .01 | |
| Age | | -.35 | -.47 | -.23 | -5.89 | < .01* | |
| **Step 2** | .06 | | | | | | < .01 |
| Constant | | 78.19 | 44.55 | 111.85 | 4.56 | < .01 | |
| Low education | | -7.65 | -13.04 | -2.31 | -2.81 | < .01* | |
| Medium education | | -4.89 | -7.83 | -1.96 | -3.27 | < .01* | |
| SCL-90, depression | | -.17 | -.32 | -.02 | -2.29 | .02 | |
| Age | | -.35 | -.46 | -.23 | -5.79 | < .01* | |
| ASTM | | .27 | -.09 | .63 | 1.45 | .15 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | Functional impairment measured with the SIP at follow-up | | | | |
| **Step 1** | .06 | | | | | .06 |
| Constant | | -114.64 | -326.75 | 97.48 | -1.06 | .29 |
| Low education | | -1.27 | -166.79 | 164.25 | -.01 | .99 |
| Medium education | | 28.17 | -68.76 | 125.10 | .57 | .57 |
| SCL-90, depression | | 12.07 | 7.40 | 16.73 | 5.07 | < .01* |
| Age | | 10.52 | 6.88 | 14.15 | 5.68 | < .01* |
| **Step 2** | .06 | | | | | < .01 |
| Constant | | 703.53 | -354.50 | 1761.56 | 1.30 | .19 |
| Low education | | -24.47 | -193.41 | 144.19 | -.28 | .78 |
| Medium education | | 21.06 | -76.64 | 118.76 | .42 | .67 |
| SCL-90, depression | | 11.63 | 6.94 | 16.33 | 4.86 | < .01* |
| Age | | 10.33 | 6.69 | 13.98 | 5.57 | < .01* |
| ASTM | | -9.10 | -20.57 | 2.36 | -1.56 | .12 |
| | | Psychological distress measured with the SCL-90 at follow-up | | | | |
| **Step 1** | .01 | | | | | .11 |
| Constant | | 76.14 | 63.86 | 88.43 | 12.18 | < .01* |
| Low education | | 9.64 | .64 | 18.63 | 2.10 | .04 |
| Medium education | | 4.01 | -1.35 | 9.37 | 1.47 | .14 |
| SCL-90, depression | | 1.29 | 1.03 | 1.55 | 9.69 | < .01* |
| Age | | .34 | .13 | .55 | 3.24 | < .01* |
| **Step 2** | .11 | | | | | .13 |
| Constant | | 102.26 | 41.94 | 162.57 | 3.33 | < .01* |
| Low education | | 8.89 | -.28 | 18.08 | 1.90 | .06 |
| Medium education | | 3.78 | -1.59 | 9.15 | 1.38 | .17 |
| SCL-90, depression | | 1.27 | 1.01 | 1.54 | 9.57 | < .01* |
| Age | | .34 | .13 | .55 | 3.17 | < .01* |
| ASTM | | -.29 | -.95 | .37 | -.87 | .39 |

*Note.* ASTM = Amsterdam short-term memory test; B = unstandardized B; CI = confidence interval; CIS = checklist individual strength; Effect size: Cohen's $f^2 = \bar{R}^2$change / (1- $\bar{R}^2$change); SCL-90 = symptom checklist-90 total score; SCL-90 depression = symptom checklist-90 depression subscale; SF36 = medical outcomes survey short-form-36; SIP = sickness impact profile total score; $n$ = 1081; *$p$-value < .0125.

**Baseline and Treatment Characteristics of Subjects Without Follow-Up Assessment**

There were no differences in age ($t$[1079] = -1.80, $p$ = .07), level of education ($X^2$[2] = 3.85, $p$ = .15), or sex (Fisher's exact test, $p$ = .14) between patients who did or did not complete follow-up. Subjects lost to follow-up performed significantly lower on the ASTM (Mann-Whitney $U$ test, $p$ = .02; $\eta^2$ = .005) and reported higher levels of physical limitations (Mann-Whitney $U$ test, $p$ < .01; $\eta^2$ = .014) at baseline. When using the ASTM dichotomously (using the cutoff of < 84), subjects who did not complete

follow-up failed the ASTM significantly more often in comparison with the subjects who completed follow-up (resp. 15.8% and 10.1%; Fisher's Exact Test, $p = .023$, $\Phi = 0.17$). No group differences were found at baseline for fatigue severity (i.e., CIS fatigue), functional impairment level (i.e., the SIP total score) or psychological distress (i.e., the SCL-90 total score) between subjects with or without follow-up assessment. Moreover, subjects who did not complete the follow-up assessment finished fewer therapy sessions in comparison with subjects who completed follow-up ($t[1079] = 16.40$, $p < .01$; mean scores of 8.67 and 14.42 respectively; Hedge's $g = 1.28$). This suggests that loss to follow-up is closely related to therapy dropout.

## DISCUSSION

While considerable attention has been focused on examining the performance validity of diagnostic assessments in various clinical samples including CFS, few studies have examined the impact of performance validity on response or adherence to treatment. Previous studies all took a dichotomous approach to performance validity, leading to loss of information and consequently reducing the statistical power to detect a relationship between performance validity and criterion variables. We chose to use the total PVT score instead, taking into account the limitation of a dichotomous approach to performance validity. To our knowledge, the current study is the first to examine the association between the total score range of a PVT and treatment outcome.

Our main hypothesis that lower ASTM scores are associated with worse treatment outcome (i.e., higher levels of self-reported symptoms or disability following CBT for CFS) was disconfirmed. A continuous ASTM-score yielded no significant associations with any outcome measures (i.e., CIS fatigue, SF-36 physical functioning, SIP total score, SCL-90 total score) in subjects who completed follow-up. However, as hypothesized, loss to follow-up was found to be associated with lower ASTM scores, as well as with higher levels of self-reported physical limitations at baseline and fewer completed therapy sessions. The latter suggests that subjects with missing follow-up assessment were not as engaged in their treatment because they attended significantly fewer therapy sessions in comparison with subjects who completed follow-up. This conclusion is reasonable, since subjects with missing follow-up assessment completed fewer therapy sessions than the 12–14 therapy sessions described in the treatment protocol. To summarize, these results indicated that CFS patients who scored low on the ASTM during baseline assessment were more likely to complete fewer therapy sessions and have missing follow-up data. However, if low-scoring CFS patients are retained in the intervention, their response to CBT for CFS is comparable with that of subjects who scored high on the ASTM (i.e., indicating effort to perform to the best of their abilities).

Our study findings are consistent with those of Goedendorp et al. (2013), who found that ASTM failure was: 1. not associated with outcome after CBT for CFS, and 2. related to loss to follow-up. The replicated findings suggest that low ASTM scores (i.e., indicating lower levels of effort to perform to the best of one's abilities) impact treatment adherence, but are not related to responsiveness to treatment in CFS patients who completed follow-up. Moore and colleagues (2013) directly studied the

relation between PVT failure and treatment adherence in a sample of patients with schizophrenia and schizoaffective disorder, who were provided with a skills-training treatment. They found that PVT failure was associated with lower group therapy attendance. This suggests that PVT results are associated with subsequent treatment adherence across existing diagnostic groups.

A multitude of patient characteristics and situational factors may be associated with the relationship between low PVT performance and (study) dropout. For example, financial incentives (e.g., a pending disability claim) are linked to low PVT performance (Bianchini et al., 2006; Sherman et al., 2020). Obviously, these incentives may also negatively impact treatment outcome, since improvement in functioning may result in lower disability compensation. However, since participants were excluded when they were engaged in a disability claim, financial incentives are not likely to be present in the current study sample. Besides, external incentives also come in the form of avoiding more basic duties such as work, school, home responsibilities, or any undesirable outcome (Sherman et al., 2020). In most cases, the clinician is unaware of the presence of these incentives, which "may be detrimental to therapeutic success" (Van Egmond et al., 2005, p. 416). In general, a broader perspective on performance invalidity beyond malingering (i.e., intentionally feigning symptoms for external motives) is desirable. Besides factitious disorder (i.e., intentionally feigning symptoms for internal motives), various psychological constructs are suggested that might result in invalid performance (Silver, 2015). Empirical studies on this topic, however, are limited and focused, for example, on "diagnosis threat" (for a critical review, *see* Niesten et al., 2020), perceived injustice (Iverson et al., 2018), and the health locus of control and self-efficacy (Armistead-Jehle et al., 2020). Preferably, these constructs are measured independently (of self-reporting) and at least using a check on the validity of self-reported measures. For example, Armistead-Jehle et al. (2020) omitted subjects with noncredible symptom reports, and found no relationship between PVT failure and the self-reported health locus of control and self-efficacy. However, when reanalyzing their data including subjects who failed symptom validity measures, a trend was observed between PVT failure and reporting a lower internal locus of control and higher inefficacy. Taking these caveats into account, it is important to conduct empirical research into possible underlying mechanisms of invalid performance. If one thing is now clear, it is that performance invalidity is not restricted to the realm of malingering and that its relevance extends beyond "noise" during diagnostic decision-making.

This was an observational study using archival treatment data, which prevents causal inferences on the relationship between performance validity and treatment outcome. Furthermore, the current findings using the ASTM cannot readily be generalized to other PVTs, which may have shown different results. Additionally, in the current study, outcome measures relied fully on self-reporting instead of more objective measures. The validity of self-reporting is itself known to be influenced by, for example, intentional symptom exaggeration (Sherman et al., 2020), inattentive responding (Merckelbach, et al., 2019), and the unreliability of memory in general (Loftus et al., 1992). This is not only a limitation of the current study. In general, there is a lack of well researched methods for evaluating the validity of reported somatic symptoms (e.g., fatigue or pain). Without questioning the clinical value of more general measures of symptom validity (e.g., based upon the validity scales of the MMPI), there is a movement toward assessing the validity of specific symptoms/conditions (Sherman et al., 2020). Promising in this regard is the Self-Report Symptom Inventory (SRSI) - issued after the inclusion period of the current

study - containing subscales on pseudo items for fatigue and pain (Merten et al., 2016). However, in the current study, ASTM performance and self-reported symptoms were negatively related at baseline, in accordance with the findings of Goedendorp et al. (2013). Despite this association, baseline ASTM performance did not impact treatment outcome (based upon self-reporting), but did impact loss to follow-up and number of completed therapy sessions. Therefore, since invalid performance and symptom validity can be viewed as "separate but related aspects of the broader construct of symptom exaggeration" (Haggerty et al., 2007, p. 926), future studies may want to employ both symptom validity and performance validity when examining treatment outcome.

Finally, some may argue that the PVT utilized measured genuine cognitive (dis)functioning in CFS patients instead of performance validity. However, it is important to emphasize that the ASTM - and PVTs in general - are constructed to be relatively insensitive to cognitive dysfunction. By definition, these tests require little cognitive effort. For example, ASTM performance was examined in nonlitigating bona-fide neurology patients diagnosed with Parkinson's disease, multiple sclerosis, and cerebrovascular accidents without "obvious clinical cognitive symptoms" (e.g., repeating the same "story," not being able to refer to an earlier subject of conversation). It is highly unlikely that these cognitive symptoms were present in the current sample of relatively young, medium, and highly educated CFS patients. The mean ASTM score in the mentioned sample of neurology patients was 87.3 (SD 2.9), with 92% of these subjects passing the ASTM (Merten et al., 2007). On a related note, using known-groups design, the sensitivity (i.e., detection of insufficient effort to perform to the best of abilities) of the ASTM was found to be excellent in its original validation study (Schagen et al., 1997), and comparable with the TOMM Trial 2 and TOMM Retention Trial (Bolan et al., 2002). Therefore, low scores on the ASTM in the current sample of CFS patients were more likely to be reflective of poor performance validity than of genuine cognitive impairment.

Taken together, our findings have clinical implications. First, that low ASTM performance in CFS patients is not a reason to be excluded from CBT, since these subjects' response to treatment is comparable with subjects who performed to the best of their abilities (i.e., had higher PVT scores) during the baseline assessment. However, low performance on the ASTM was associated with loss to follow-up and fewer completed therapy sessions. Therefore, instead of being an indicator restricted to the assessment of symptom credibility, performance validity may also serve as a behavioral proxy of how patients engage in a behavioral treatment intervention (e.g., some might have reservations about the communicated diagnoses and/or treatment plans). Additional research is necessary to help understand the association between performance validity, adherence to treatment, and outcomes. Ultimately, a better determination of factors that are known to impact treatment adherence and treatment outcome may sharpen indications for treatment and, consequently, prevent costly specialized tertiary medical care.

## REFERENCES


Armistead-Jehle, P., Lippa, S. M., & Grills, C. E. (2020). The Impact of self-efficacy and health locus of control on performance validity testing. *Archives of Clinical Neuropsychology.* Advance online publication. https://doi.org/10.1093/arclin/acaa027

Altman, D. G. & Royston, P. (2006). The cost of dichotomizing continuous variables. *The British Medical Journal, 332,* 1080.

Arrindell, W. A., & Ettema, J. H. (2005). Symptom Checklist. Handleiding bij een multidimensionale psychopathologie indicator [Symptom Checklist. Manual of a multidimensional psychopathology indicator]. Amsterdam: Harcourt Test Publishers.

Arrindell, W., Ettema, H., Groenman, N., Brook, F., Janssen, I., Slaets, J., Hekster, G., Derksen, J., van der Ende, J., Land, H., Hofman, K., & Dost, S. (2003). De groeiende inbedding van de Nederlandse SCL-90-R*: Psychodiagnostisch gereedschap [Further Dutch experiences with the Symptom Checklist-90 Revised]. *Psycholoog, 38*(11), 576–582

Bergner, M., Bobbitt, R. A., Carter, W. B., & Gilson, B. S. (1981). The Sickness Impact Profile: development and final revision of a health status measure. *Medical Care, 19*(8), 787–805. https://doi.org/10.1097/00005650-198108000-00001

Bianchini, K. J., Curtis, K. L., & Greve, K. W. (2006). Compensation and malingering in traumatic brain injury: A dose-response relationship? *The Clinical Neuropsychologist, 20*(4), 831-847.

Bolan, B., Foster, J. K., Schmand, B., & Bolan, S. (2002). A comparison of three tests to detect feigned amnesia: the effects of feedback and the measurement of response latency. *Journal of Clinical and Experimental Neuropsychology*, *24*(2), 154–167. https://doi.org/10.1076/jcen.24.2.154.1000

Centraal Begeleidings Orgaan. (2013). *Richtlijn Diagnose, behandeling, begeleiding en beoordeling van patiënten met het chronisch vermoeidheidssyndroom (CVS) [Guideline: Diagnosis, treatment, coaching and evaluation of patients suffering chronic fatigue syndrome (CFS)].* Retrieved from *https:// www.diliguide.nl/document/3435/file/pdf/* 2013.

Dandachi-FitzGerald, B., van Twillert, B., van de Sande, P., van Os, Y., & Ponds, R. W. (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.

De Bie, S. (1987). Standaardvragen 1987: Voorstellen voor uniformering van vraagstellingen naar achtergrondkenmerken en interviews [Standard questions 1987: Proposal for uniformization of questions regarding background variables and interviews]. Leiden, the Netherlands: Leiden University Press.

Derogatis, L. R. (1994). SCL-90-R: Administration, scoring and procedures manual (3rd ed.). Minneapolis, MN: Nation Computer Systems.

Goedendorp, M. M., van der Werf, S. P., Bleijenberg, G., Tummers, M., & Knoop, H. (2013). Does neuropsychological test performance predict outcome of cognitive behavior therapy for chronic fatigue syndrome and what is the role of underperformance? *Journal of Psychosomatic Research, 75*, 242-248.

Haggerty, K. A., Frazier, Th. W., Busch, R. M., & Naugle, R. I. (2007). Relationships among Victoria Symptom Validity Test indices and Personality Assessment Inventory validity scales in a large clinical sample. *The Clinical Neuropsychologist, 21*, 917–928.

Horner, M. D., VanKirk, K. K. Dismuke, C. E., Turner, T. H., & Muzzy, W. (2014). Inadequate effort on neuropsychological evaluation is associated with increased healthcare utilization. *The Clinical Neuropsychologist, 28*(5), 703-713.

Iverson, G. L. (2006). Ethical issues associated with the assessment of exaggeration, poor effort, and malingering. *Applied Neuropsychology, 13*(2), 77-90.

Iverson, G. L., Terry, D. G., Karr, J.E., Panenka, W. J., & Silverberg, N. D. (2018). Perceived injustice and its correlates after mild traumatic brain injury. *Journal of Neurotrauma, 35*, 1156-1166.

Jacobs, H. M., Luttik, A., Touw-Otten, F. W., & de Melker, R. A. (1990). The Sickness Impact Profile; results of an evaluation study of the Dutch version (De sickness impact profile; resultaten van een valideringsonderzoek van de Nederlandse versie). *Nederlands Tijdschrift voor Geneeskunde, 134,* 1950–1954.

Janse, A., Wiborg, J. F., Bleijenberg, G., Tummers, M., & Knoop, H. (2016). The efficacy of guided self-instruction for patients with idiopathic chronic fatigue: A randomized controlled trial. *Journal of Consulting and Clinical Psychology, 84*(5), 377-388.

Jurick, S. M., Crocker, L. D., Merritt, V. C., Hoffman, S. N., Keller, A. V., Eglit, G. M. L., Thomas, K. R., Norman, S. B., Schiehser, D. M., Rodgers, C. S., Twamley, E. W., & Jak, A. J. (2020). Psychological symptoms and rates of performance validity improve following trauma-focused treatment in veterans with PTSD and history of mild-to-moderate TBI. *Journal of the International Neuropsychological Society*, *26*(1), 108–118. https://doi.org/10.1017/S1355617719000997

Knoop, H., & Bleijenberg, G. (2010). Het chronisch vermoeidheidssyndroom. Behandelprotocol cognitieve gedragstherapie voor CVS [Chronic fatigue syndrome. Cognitive behavior therapy for CFS treatment protocol]. Houten, The Netherlands: Bohn Stafleu Van Loghum.

Knoop, H., Prins, J. B., Moss-Morris, R., & Bleijenberg, G. (2010). The central role of cognitive processes in the perpetuation of chronic fatigue syndrome. *Journal of Psychosomatic Research, 68*(5), 489-494.

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

Lippa, S. M., Pastorek, N. J., Romesser, J., Linck, J., Sim, A.H., Wisdom, N. M., & Miller, B.I. (2014). Ecological validity of performance validity testing. *Archives of Clinical Neuropsychology, 29*(3), 236-244.

Loftus, E. F., Levidow, B., & Duensing, S. (1992). Who remembers best? Individual differences in memory for events that occurred in a science museum. *Applied Cognitive Psychology, 6,* 93–107.

Long, C., Dolley, O. & Hollin, C. (2012). Engagement in psychosocial treatment: Its relationship to outcome and care pathway progress for women in medium-secure settings. *Criminal Behaviour and Mental Health, 22*, 336–349.

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.

Merckelbach, H., Dandachi-FitzGerald, B., van Helvoort, D., Jelicic, M., & Otgaar, H. (2019). When patients overreport symptoms: More than just malingering. *Current Directions in Psychological Science*, *28*(3), 321–326.

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

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.

Moore, R., Davine, T., Harmell, A., Cardenas, V., Palmer, B., & Mausbach, B. (2013). Using the repeatable battery for the assessment of neuropsychological status (RBANS) effort index to predict treatment group attendance in patients with schizophrenia. *Journal of the International Neuropsychological Society, 19*(2), 198-205.

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

Prins, J. B., Bazelmans, E., Van der Werf, S. P., Van de Meer, J., & Bleijenberg, G. (2001). Cognitive-behaviour therapy for chronic fatigue syndrome: predictors of treatment outcome. Paper presented at the psycho-neuro-endocrino-immunology (PNEI): A common language for the whole human body: Proceedings of the 16th World Congress of Psychosomatic Medicine, Göteborg, Sweden.

Reeves, W. C., Lloyd, A., Vernon, S. D., Klimas, N., Jason, L. A., Bleijenberg, G., Evengard, B., White, P. D., Nisenbaum, R., Unger, E. R., & International Chronic Fatigue Syndrome Study Group (2003). Identification of ambiguities in the 1994 chronic fatigue syndrome research case definition and recommendations for resolution. *BMC Health Services Research*, *3*(1), 25. https://doi.org/10.1186/1472-6963-3-25

Scheeres, K., Wensing, M., Knoop, H., & Bleijenberg, G. (2008). Implementing cognitive behavioral therapy for chronic fatigue syndrome in a mental health center: a benchmarking evaluation. *Journal of Consulting and Clinical Psychology, 76*(1), 163-171.

Schagen, S., Schmand, B., de Sterke, S., & Lindeboom, J. (1997). Amsterdam Short Term Memory Test: A new procedure for the detection of feigned memory deficits. *Journal of Clinical and Experimental Neuropsychology, 19*, 43-51.

Schmand, B., & Lindeboom, J. (2005). Amsterdam Short-Term Memory Test: Manual. Leiden, The Netherlands: Psychologische Instrumenten, Tests en Services.

Sherman, E. M., 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

Silver, J. M. (2015). Invalid symptom reporting and performance: What are we missing? *NeuroRehabilitation*, *36*, 463–469.

Sollman, M. J. & Berry, D. T. (2011). Detection of inadequate effort on neuropsychological testing: A meta-analytic update and extension. *Archives of Clinical Neuropsychology, 26*(8), 774-789.

Spratt, M., Carpenter, J., Sterne, J. A., Carlin, J. B., Heron, J., Henderson, J., & Tilling, K. (2010). Strategies for multiple imputations in longitudinal studies. *American Journal of Epidemiology, 172*, 478-487.

Sterne, J., White, I. R., Carlin, J. B., Spratt, M. P., Royston, P., Kenward, M. G., Wood, A.M., & Carpenter, J.R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *The British Medical Journal, 338*, 157 - 160.

Van der Elst, W., van Boxtel, M.P., van Breukelen, G. J., & Jolles, J. (2005). Rey's verbal learning test: Normative data for 1855 healthy participants aged 24-81 years and the influence of age, sex, education, and mode of presentation. *Journal of the International Neuropsychological Society, 11*, 290–302.

Van Egmond, J., Kummeling, I., & Balkom, T. (2005). Secondary gain as hidden motive for getting psychiatric treatment. *European Psychiatry, 20*(5-6), 416-421.

Von Elm, E., Altman, D.G, Egger, M., Pocock, S.J., Gøtzsche, P.C., Vandenbroucke, J.P, & STROBE initiative (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Lancet, 370*, 1453-1457.

Van Ginkel, J. R. (2019). Significance tests and estimates for $R^2$ for multiple regression in multiply imputed datasets: A cautionary note on earlier findings, and alternative solutions. *Multivariate Behavioral Research, 54(*4), 514-529.

Ware Jr., J. E., & Sherbourne, C.D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection, *Medical Care, 30,* 473–483.

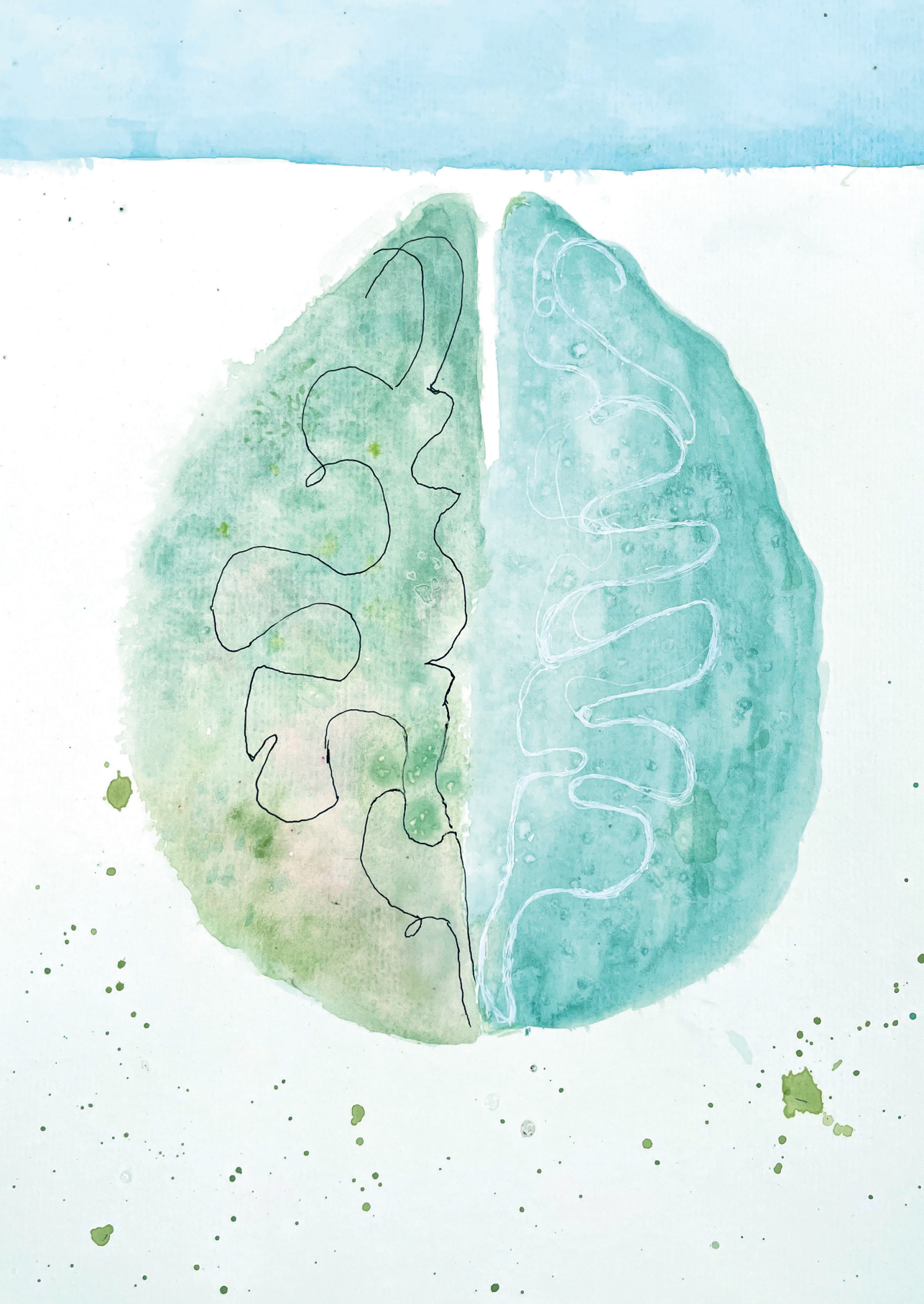
Williams, M. W., Graham, D., Sciarrino, N. A., Estey, M., McCurry, K. L., Chiu, P., & King-Casas, B. (2020). Does validity measure response affect CPT group outcomes in veterans with PTSD? *Military Medicine*, *185*(3-4), 370–376.

Worm-Smeitink, M., Gielissen, M., Bloot, L., van Laarhoven, H. W., van Engelen, B. G., van Riel, P., Bleijenberg, G., Nikolaus, S., & Knoop, H. (2017). The assessment of fatigue: Psychometric qualities and norms for the Checklist individual strength. *Journal of Psychosomatic Research*, *98*, 40–46. https://doi.org/10.1016/j.jpsychores.2017.05.007

## APPENDIX A

**Complete Case Analyses (CC) of Hierarchical Linear Regression Analyses on Scores of Fatigue Severity, Physical Limitations, Functional Impairment, and Psychological Distress at Follow-Up**

The first model accounting for the combined explained variance of age, level of education, and depressive symptoms (i.e., SCL-90 depression subscale) was significant for all criterion variables (CIS fatigue: $R^2 = .04$, $F[4, 850] = 10.25$, $p < .01$; SF-36 physical functioning: $R^2 = .08$, $F[4, 850] = 19.14$, $p < .001$; SIP total score: $R^2 = .09$, $F[4, 851] = 20.57$, $p < .01$; SCL-90 total core: $R^2 = .18$, $F[4, 799] = 42.68$, $p < .01$). This finding was repeated in the second model with the added continuous ASTM score as predictor of treatment outcome (CIS fatigue: $R^2 = .04$, $F[5, 849] = 8.20$, $p < .01$; SF-36 physical functioning: $R^2 = .08$, $F[5, 849] = 15.180$, $p < .01$; SIP total score: $R^2 = .09$, $F[5, 850] = 16.96$, $p < .01$; SCL-90 total core: $R^2 = .18$, $F[5, 798] = 34.18$, $p < .01$). Importantly, the added ASTM score in Model 2 did not yield a significant improvement in the prediction of treatment outcome for all criterion variables (CIS fatigue: $R^2$ change <.01, $F[1, 849] = .07$, $p = .79$; SF-36 physical functioning: $R^2$ change = < .01, $F[1, 849] = 2.35$, $p = .13$; SIP total score: $R^2$ change = < .01, $F[1, 850] = 2.36$, $p = .12$; SCL-90; $R^2$ change = < .01, $F[1, 798] = .30$, $p = .58$).



# CHAPTER 7

## General Discussion

Neuropsychological assessment is used to determine "whether patients have objectively verified (i.e., credible) cognitive dysfunction" (Boone, 2007, p. 37). In order to make accurate inferences about their cognitive abilities, patients are expected to put forth their best efforts when cognitive tests are administered. It is important to emphasize that the implications when patients do not perform to the best of their capabilities during neuropsychological assessment, can be significant. For example, when gone undetected, underperformance on cognitive tests can potentially lead to several consequences: (1) misdiagnoses, (2) reinforcement of noncredible symptom reporting, (3) unnecessary diagnostic workup (e.g., magnetic resonance imaging [MRI] of the brain), (4) unnecessary and potentially harmful treatment, limiting the exploration of more effective therapeutic options, and (5) self-imposed restrictions on activities, such as quitting work, hobbies or driving (Schroeder & Martin, 2022). Hence, the relevance of performance validity assessment is nowadays considered crucial in *every* neuropsychological assessment (i.e., clinical and forensic), ensuring confidence in the credibility of test performance (Moore et al., 2021; Sweet, et al. 2021).

As performance validity assessment is becoming more and more standard practice in clinical settings (Hirst et al., 2017), still some important questions remain unanswered. The most important being the frequency of clinical patients who exhibit indications of performance below best of capabilities. Additionally, there are uncertainties surrounding how to approach these patients who show signs of invalid performance, and as to whether this behavior is clinically relevant beyond the diagnostic realm. Therefore, in this dissertation, we aimed to investigate (1) the base rate of performance validity test (PVT) failure, (2) the effects of feedback interventions upon indications of invalid performance to improve test performance, and (3) the impact of performance validity on treatment outcome in adult clinical patients. This was performed by using several study designs including a case report, systematic review, meta-analysis, randomized controlled trial, and cross-sectional studies.

This chapter discusses the main findings, methodological strengths and constraints concerning our studies, implications for clinical practice, and recommendations for future directions.

## MAIN FINDINGS

### Base Rates of Performance Validity Test (PVT) Failure in Routine Clinical Care

We conducted a systematic review and meta-analysis for calculating the pooled base rate of PVT failure in adult patients seen for routine clinical neuropsychological assessment **(chapter 3)**. To limit false positive PVT findings (Lippa, 2018), strict in- and exclusion criteria were used by only including studies that applied freestanding PVTs with proven psychometric qualities (Sollman & Berry, 2011) in adult patients without obvious cognitive impairment due to dementia or intellectual disability. We found that failing one well-researched freestanding PVT in these patients occurred in 16% (95% CI [14, 19]) of all included cases. Our empirical findings show a remarkable resemblance with the reported median base rate (i.e., 15%) of invalid test results in clinical non-forensic settings from a recent survey amongst 178 practicing neuropsychologists (Martin & Schroeder, 2020). Furthermore, our findings confirm that PVT failure occurs in a sizeable minority of patients seen for clinical neuropsychological assessment. It should

be noted that for some additional subgroup-analyses, the pooled PVT failure rates are based upon relatively few subjects or small numbers of studies, impacting the stability of these reported estimates. Therefore, a first step for future studies on the base rate of invalid performance would be to use larger group sizes across relevant clinical characteristics (i.e., type of clinical setting, presence of external gains) and preferably homogenous diagnosis groups.

Subgroup analyses were performed to examine whether the base rate of PVT failure was related to specific clinical contexts, distinct diagnostic patient groups, PVTs used, and the presence of potential external gains. Strikingly, pooled PVT failure rates varied quite dramatically across these relevant characteristics. This finding is reminiscent of the notion that invalid performance is not a static trait of examinees undergoing neuropsychological assessment, but related to the personal circumstances (such as diagnosis group membership), specific objectives (such as external gain incentives being present), and to the particular setting (such as private practice versus rehabilitation clinic) (Rogers, 2008). By providing pooled PVT failure rates across these relevant patient and context characteristics, clinically applied statistics such as positive- and negative predictive value (PPV/NPV) can be calculated more accurately by choosing the characteristics that fit best with the patient sample in question (*see* Dandachi-FitzGerald & Martin, 2022 for a detailed description).

The neuropsychological test results of the patient from the case report **(chapter 2)** can be used to illustrate the clinical application of our review-findings by contrasting two hypotheses. Consider for the first hypothesis, that the patient indeed has a credible diagnosis of mild cognitive impairment (MCI). The corresponding pooled base rate of PVT failure in this diagnosis group is 9%. The alternative hypothesis would be that PVT failure was related to the external incentives being present. The corresponding pooled base rate of PVT failure in patient samples in which subjects with potential external gains were *not* excluded, is 16%. Therewith, instead of a pass/fail approach to a PVT score to determine performance validity, the reported base rate of PVT failure closest to the examined diagnostic hypotheses (i.e., MCI versus External gains) can be used in a next step to further increase the accuracy of performance validity determination. This is illustrated by calculating the likelihood that PVT failure was indeed a true positive (i.e., positive predictive value, PPV) and its counterpart (i.e., negative predictive value, NPV). For both PVTs (ASTM and WMT[2]) and both associated base rates (for MCI and External gains), the likelihood that passing a PVT was a true-negative finding was high, as NPV values varied from 93.6% (WMT; External gains) to 98.3% (ASTM; MCI). Thus, passing either PVT would help to rule out performance invalidity with a high level of confidence (at least 93.6%). Importantly, the likelihood that PVT failure was due to invalid performance (PPV) varied quite dramatically. For the WMT, PPV varied from 19.5% (MCI) to 31.9% (External gains), and for the ASTM from 45.4% (MCI) to 61.5% (External gains). This clearly shows that the likelihood that a single PVT failure is truly indicative of invalid performance greatly depends on the context of the evaluation, due to its impact on base rates. Although an increase in PPV on the WMT (19.5% to 31.9%) and ASTM (45.4% to 61.5%) is substantial and in favor of the alternative hypotheses of External gains being related to PVT failure, these values are still not sufficient to determine performance invalidity. Additionally, the more that a PVT score falls beyond a pre-established cutoff for invalid performance,

2 An important sidenote is the substandard low specificity for the WMT of 69.4% (per Sollman & Berry, 2011).

the higher the PPV. In other words, scoring farther below this cutoff reduces the likelihood that PVT failure is a false-positive (with significant below chance performance on a PVT as an extreme example). Another method for increasing the classification accuracy, in which the examiner can directly influence PPV, is by administering multiple independent PVTs (Larrabee, 2008). Nonetheless, while considerable weight should be given to the psychometric evaluation of performance validity, the clinician should also include other test and extra-test information (e.g., degree of PVT failure, (in) consistency of the clinical presentation) to draw conclusions about the validity of an individual patient's neuropsychological assessment (Dandachi-FitzGerald & Martin, 2022; Sherman et al., 2020). The patient from the case report reported considerable impairments in cognitive functioning, showed low performance across formal cognitive tests, and, importantly, failed multiple well-validated freestanding PVTs. These diagnostic outcomes were highly inconsistent with her fully intact level of daily functioning (e.g., pursuing a successful career as an artist), compromising the MCI diagnosis.

**Impact of Feedback**

In **chapter 4**, we examined the effects of a brief feedback intervention in adult patients diagnosed with chronic fatigue syndrome (CFS) who failed a PVT at baseline. During the feedback intervention, the psychologist (a) addressed that the CFS symptoms of the patient were difficult to evaluate because of the lower-than-expected performance on a previously administered test, (b) emphasized exerting the patient's best effort and that improvement was expected, and (c) explained that therefore tests needed to be repeated. The feedback group showed no improvements on a repeated PVT compared to a matched no-feedback group of patients with CFS. In fact, both groups showed comparable improvements in performance on this PVT during re-administration. On one of out of two administered measures of speed of information processing (i.e., complex reaction time task, CRT), the feedback group showed significant improvement during follow-up compared to the no-feedback group. However, it is important to consider that this specific result could also have been influenced by the difference in time intervals for the repeated assessments between the two groups. Next to this potentially limited feedback-effect, pass/fail fluctuations in PVT performance between assessments were apparent. The latter implies that performance validity should be checked in every repeated neuropsychological assessment.

**Chapter 5** describes a multisite single-blind randomized-controlled trial (RCT) in a general hospital setting on the effects of a brief immediate feedback intervention upon PVT failure. By default, most patients were encouraged to do their best by the clinician (neuropsychologist) and/or technician before starting the assessment. We found no group (feedback vs. no-feedback) differences on a repeated PVT and repeated standard cognitive tests. Such group differences were also absent on a subsequently single-administered PVT and standard cognitive tests. We found the vast majority of participants continued to fail at least one PVT, independent of feedback. These study results challenge the findings from Suchy and colleagues (2012), which is the only other study to examine the impact of immediate feedback upon PVT-failure on subsequent test performance in clinical patients. These authors found patients diagnosed with multiple sclerosis (MS) significantly improved their performance on a repeated PVT and memory test, but this study was lacking a no-feedback control condition.

Together, our results from **chapters 4 and 5** indicate that performance below best of capabilities was persistent, independent of type of feedback examined in both studies. Consequently, these findings suggest that there might be limitations to using feedback upon indications of invalid performance for increasing patients' efforts to perform at the best of their capabilities.

### Treatment Outcomes

Treatment recommendations based upon invalid data are likely not targeting the expected (medical) conditions, and these treatments may therefore be less efficient and/or not in line with the actual needs of the patient. In a worst-case-scenario, potentially iatrogenic treatments are started for non-existent (medical) conditions. Therefore, we were interested in examining the relationship between invalid performance and treatment outcome. In **chapter 6**, we used the total score range of a PVT to detect a relationship with treatment outcome. Archival data of 1081 outpatients treated with protocolled cognitive behavioral therapy (CBT) for chronic fatigue syndrome (CFS) were used in this study. Lower scores on the PVT were associated with completing fewer therapy sessions and missing follow-up assessment. However, for those patients who completed the intervention, their response to CBT was comparable to those who scored high on the ASTM, despite their initial lower performance on the PVT. These findings are in line with those from Goedendorp et al. (2013), who found that PVT failure at baseline impacted treatment adherence, but not the responsiveness to treatment in patients who completed therapy. This suggests that performance validity may serve as a behavioral proxy of how patients engage in a behavioral treatment intervention, but additional research on this topic is needed.



## METHODOLOGICAL CONSIDERATIONS

### Strengths

Our studies have a number of strengths. First, all studies in this dissertation were performed in adult patients seeking treatment in a context of routine clinical care, which strengthens external validity as compared to, for example, results from experimental studies using simulation designs in healthy subjects (Rogers & Bender, 2018). Furthermore, necessary precautions were made to ensure that performance validity measures would provide accurate information on the credibility of cognitive abilities. This was done in several ways. First, by stringently excluding patients with clinically obvious cognitive symptoms and intellectual disability, as standard PVT cutoff application would lead to an increase in false-positives in these patients (Lippa, 2018). Second, by using freestanding PVTs with proven psychometric qualities (Sollman & Berry, 2011). Third, by controlling for the presence of potential external incentives, as this factor is recognized as an important driver behind PVT failure (Schroeder et al., 2022). Another strength is that we utilized various study designs to illustrate and examine performance validity assessment in clinical settings. A final strength of this dissertation is that in the two feedback intervention studies, a control group (Chapter 4) and experimental study design (Chapter 5) were employed. With that, we improved the accuracy on current knowledge about the effects of feedback interventions compared to a prior non-experimental observational study where a control condition was lacking (Suchy et al., 2012).

### Limitations

After more than a decade, the highly influential consensus statement on validity assessment was updated (Sweet et al., 2021). In these updated guidelines, the use of multiple PVTs ("two-failure rule") instead of one to determine psychometric evidence of performance below best of capabilities is underscored. The primary goal for using multiple PVTs is preventing false-positive conclusions of invalid performance (i.e., PVT failure due to other factors than invalid performance, such as severe cognitive disorder due to dementia). However, studies in this dissertation used, for different reasons, a one-failure approach to determine invalid performance. One could counter that in all of our studies, various major factors related to false-positive PVT classification were addressed (i.e., intellectual disability, low educational level, fluency in Dutch language, severe cognitive impairment, external gain incentives; Lippa, 2018; Sweet et al., 2021). By adhering to these strict in- and exclusion criteria, precautionary measures were taken to minimize the impact of these factors on PVT failure. Additionally, only well-validated freestanding PVTs with generally high specificity levels (≥ .90) were used in an attempt to further limit false-positive PVT classifications. Consequently, because of their inverse relationship, sensitivity levels are lowered. This leaves a proportion of patients who performed below best of capabilities undetected (i.e., false-negatives). By applying these measures, we tried to reduce the risk of false-positive classifications. Nonetheless, for to enhance the classification accuracy it would have been better if more PVTs were employed - embedded and freestanding - to determine the validity status.

Finally, we used only "high-stake" definitions of external incentives in our studies, such as involvement in applying for a disability claim (Chapters 4, 5, and 6). Moreover, the way potential external incentives were examined and defined varied significantly if we also look at the studies that were included in our systematic review (Chapter 3). With the publication of the updated Malingered Neurocognitive Disorder (MND)-criteria (Sherman et al., 2020), the concept of external gain was broadened by also including "avoiding having to fulfill more basic duties and responsibilities such as avoiding work, school examinations, or home responsibilities" (p. 739). Therefore, the definition of external incentives in our studies may have been too restricted in favor of a more financial incentive-type description (i.e., social security benefits and litigation; Dandachi-FitzGerald et al., 2020; van Egmond & Kummeling, 2005).

## IMPLICATIONS FOR CLINICAL PRACTICE

### Performance Validity Assessment in Routine Clinical Care

Several clinically relevant considerations can be drawn from the studies presented in this dissertation. One of our most important findings is that in adult clinical patient groups presenting for routine clinical care, a substantial minority fails a well-validated freestanding PVT. These findings confirm the need that *all* clinical neuropsychological evaluations should address the issue of performance validity to ensure reliable and clinically meaningful results. Since in one third of cases, surveyed clinical neuropsychologists reported not to include validity measures in all neuropsychological assessments, this clearly leaves room for improvement (Hirst et al., 2017). Moreover, the pooled base rates of PVT failure (Chapter 3, Table 2) can be used for determining the accuracy of a PVT finding for an individual patient by taking relevant

clinical and patient characteristics into account. This was illustrated by using the neuropsychological test results from the case report (Chapter 2).

**The Impact of Feedback on Subsequent Test Performance**

An ongoing discussion among clinicians is if, when, and how we should provide feedback to increase patients' test-taking efforts in case of PVT failure. We examined two different feedback interventions and found PVT failure to persist. After excluding patients involved in a personal injury claim during the clinical assessment, PVT failure persisted in the majority of remaining patients, indicating that factors other than litigation-specific external gain incentives may contribute to the ongoing noncredible responding. These findings suggest that a neutral feedback-approach for improving test performance is of limited clinical value to improve test performance in routine clinical care.

Although more research is warranted to further improve our understanding of the possibilities and limitations of using feedback to increase test performance, other directions for improving neuropsychological assessment outcomes in clinical patients may be considered. Promising in this regard is approaching psychological assessment as a therapeutic intervention that already starts with the initial interview, coined Therapeutic Assessment (TA; Poston & Hanson, 2010) or Collaborative Therapeutic Neuropsychological Assessment (CTNA; Gorske, 2008). Key aspects of these approaches are (1) developing and maintaining an empathic connection, (2) working collaboratively to define individualized assessment goals, and (3) inviting patients to actively collaborate in discussing (and making sense of) the psychometric test results. In their recent meta-analyses, Durosini and Aschieri (2021) found TA to have statistically significant effects on treatment process, self-reported symptoms, and improvements during the assessment ("self-enhancement"). Longley and colleagues (2023) were the first to examine the therapeutic effects of neuropsychological assessment in adult patients with multiple sclerosis (MS) using a high-quality experimental study design. These authors found significant improvements in perceived everyday cognitive functioning, MS self-efficacy, and stress and depression at 1-month follow-up. Waldronn-Perrine and colleagues (2021) discuss the potential benefits of TA-based feedback in patient who show non-credible responding (i.e., fail an SVT of PVT), but empirical data on its effects are lacking. More in general, a therapeutic assessment-approach to performance validity may be promising. By elaborating on the patients' assessment goals, expected benefits, and various assessment outcome-scenario's during the initial interview, subsequent test performance or other assessment outcomes may be improved. One potential outcome may in fact be to postpone the neuropsychological assessment as a whole, as therapeutic goals are already clear irrespective of further assessment outcomes. Illustratively is the study by Jurick et al. (2020). These authors found that patients who failed a PVT exhibited a clinically meaningful reduction in symptoms of PTSD, depression, and post-concussive symptoms after treatment. Strikingly, their PVT performance also increased *after* treatment was completed, leading to significantly less subjects who failed a PVT. Therefore, these authors suggest that patients "may benefit from neuropsychological assessment after, rather than before, treatment" (p. 108).

**Performance validity and its Relationship with Treatment Outcome**

Research on performance invalidity and its implications for treatment in clinical patients is limited. Our findings indicate that patients who show indications of invalid performance should not be excluded from therapy, as those who follow-through with treatment do benefit from it, comparable to patients with higher scores on a baseline PVT (i.e., indicating that their test performance is likely valid). However, we did find that lower PVT performance is related to limited treatment adherence. Therefore, additionally to diagnostic conclusions, invalid performance may also indicate that factors relevant for subsequent (indications for) treatment are in need of clinical attention. The underlying motives of patients who invalidate testing and show limited treatment adherence are likely heterogeneous in nature and for a clinician difficult to unravel. Nevertheless, when patients show indications of invalid performance, it may be clinically useful to first discuss outcome expectations and reaching agreement about the indicated treatment, treatment goals and related tasks and activities (i.e., such as graded activity) –instead of neglecting the issue of noncredible performance and continue with care-as-usual. For example, patients' beliefs about the mental health consequence of participating in psychotherapy (i.e., outcome expectations; OE) are found to be significantly associated with treatment success (Constantino et al., 2018). Therefore, these authors outlined practice suggestions to help clinicians cultivate and respond to their patients' OE (p. 480). In short, invalid performance may be viewed as a behavioral proxy of how patients engage in subsequent treatment (e.g., having reservations about the communicated therapy proposal). Therefore, patients displaying this behavior may be in need of specific clinical attention - instead of losing empathy and abandoning attempts to provide clinical aid. Ultimately, as clinicians, it is our mandate to provide adequate and useful clinical services.

## FUTURE DIRECTIONS

The results of the studies from this dissertation give rise to several recommendations for future research and clinical directions to be considered.

**Base Rates of PVT Failure in Routine Clinical Care**

The findings from our systematic review and meta-analyses clearly indicate that PVT failure in clinical contexts is substantial and shows large variability, depending on patient-, assessment-, and context factors. Therefore, to further improve the accuracy of performance validity determination, additional research in examining these factors related to invalid performance is warranted.

During the identification phase of our systematic review, it was apparent that PVTs were not by default administered in line with their respective manuals. For example, a substantial number of studies were excluded because of the application of non-standard PVT cutoffs. Additionally, studies examined performance validity by only using embedded PVTs (without the recommended combination with at least one freestanding PVT), or using freestanding PVTs which only have been validated in small and/or restricted samples. Evidently, such a proliferation of PVT application in research is undesirable for establishing accurate base rates of PVT failure. Therefore, future research on PVT failure in clinical patients

may want to address failure rates using the standard per manual PVT cutoffs -possibly in combination with a different (for research purposes) PVT application or cutoff. Relatedly, future research may also want to examine the base rate of failing multiple PVTs (i.e., two-failure rule) in clinical patients, as this is in line with recommended clinical practices (Sweet et al., 2021). Moreover, failing two or more out of multiple PVTs as an external criterion for invalid performance, has been shown to lead to improvement in classification accuracy compared to using a single PVT failure (Schroeder et al., 2019). A major advantaged is that the outcomes of multiple PVTs provide the opportunity to calculate additional diagnostic statistics. More specifically, by "chaining" the likelihood ratios (LRs) of multiple PVTs, the positive predictive value (PPV) and the confidence that can be placed on positive findings is markedly increased (Larrabee, 2008).

Research clearly shows that incentives cause PVT failure in a majority of cases, even in routine clinical care (Schroeder at al., 2022). Gaining more insights in the nature of "incentives" may help in understanding why some forms of incentives are related to non-credible test performance, and other forms are not. Therefore, a first crucial step in all performance validity research should be to carefully screen for potential incentives. In the updated MND criteria (Sherman et al., 2020), the criteria for presence of external gains for malingering are expanded, making them more relevant for routine clinical care contexts. These added "lower-stakes" incentives (such as avoiding work, school examinations, or home responsibilities) can be used for future research additionally to "high stakes" incentives (e.g., avoiding criminal prosecution). The relevance of lower-stakes incentives in routine clinical care is illustrated by the study of van Egmond and colleagues (2005), who found that 41% of clinical patients anonymously reported to expect gaining support to obtain non-treatment related advantages from being in therapy (e.g., related to work, social security, compensation for healthcare costs, or their insurance). In contrast, less than 10% communicated these expectations with their consulting psychiatrist.



**Feedback Interventions to Improve Subsequent Test Performance**

Although it can be argued that our feedback interventions were too "soft" (i.e., non-accusatory/neutral), or not clear enough to elicit improvements on subsequent test performance, these approaches were chosen to mimic common clinical practices, as clinicians and technicians typically encourage patients to perform to their best of capabilities in a non-confrontational, empathetic manner. However, since we found no feedback-effects on subsequent test performance, other types of feedback may be examined. For example, future research may contrast various and more extreme types of direct feedback (sympathetic, neutral, and confrontational) to examine whether other feedback approaches may elicit improvements or normalization (to performance at best of capabilities) on subsequent test performance. Relatedly, although the publication of feedback-frameworks for providing clinical feedback when patient invalidate testing is clinically helpful (Carone et al., 2010; Martin & Schroeder, 2022), future research should test such feedback approaches empirically in order to establish their impact on subsequent test performance.

**Performance Validity and its Potential Relationship with Treatment Outcome**

To date, very little research examined the impact of invalid performance on treatment outcome, as most attention is devoted to its relevance for diagnostic practices. This represents a missed opportunity, as

invalid performance may convey potentially relevant information for treatment planning, adherence, and outcomes (Lippa et al., 2014). Our study findings indicate that performance validity is relevant for treatment-related factors, as performing low on a PVT was found to be related to limited treatment adherence (i.e., completing fewer therapy sessions and study drop-out). Relatedly, as stated by Lippa and colleagues (2014), "failure of PVTs may lend some insight into the patients' ability to cope with the stressors of everyday life and may even contribute to decisions regarding treatment planning" (p. 240). Therefore, future research may want to examine the relationship between invalid performance and the subsequent course of treatment (e.g., medical consumption, treatment adherence/drop-out), potentially contributing to developing interventions for specific subgroups of invalidly performing patients.

**"Think Dirty!": Relevance for Clinicians, Education, and (Postdoctoral) Training in Psychology**

Although invalid performance is found to be substantial in routine clinical care, textbooks and (postdoctoral) training devote little attention to this area of expertise. This may contribute to clinicians oftentimes struggling with the correct application of performance validity assessment, as well as interpreting and incorporating these results into their practices. A welcome development is the issuing of criteria for competency-based assessment in clinical neuropsychology for practicum training (Nelson et al., 2015) and the postdoctoral level (Heffelfinger et al., 2022). These guidelines explicitly mention performance validity assessment as a required competency. Formally addressing this topic in teaching and training in psychology contributes to the understanding that performance validity assessment is warranted in all neuropsychological evaluations, and that guidelines for its proper application are available. Relatedly, Beach and colleagues (2017) signaled a lacuna and stated that "despite the prevalence with which trainees encounter patients who manipulate, deceive, or withhold information, trainees receive little formal guidance in "thinking dirty" -incorporating elements of hidden patient motives into their interview, formulation, and plans" (p. 474). They therefore propose a multi-modal approach for teaching trainees to recognize hidden motives and deception. Their goal is to improve patient care, and help trainees to normalize their experiences of being deceived or having information withheld from them.

To date, the Guideline for the Use of Tests by the Dutch Association of Psychologists (NIP, 2017), detailing recommended practices for psychodiagnostic assessment and competency levels, is silent about the issue of validity assessment. The section Neuropsychology of the Dutch Association of Psychologists published a guideline for independent medico-legal neuropsychological assessment, in which the importance of performance validity assessment is explicated (NIP, 2016). This is in contrast with the United States and Great Britain, where professional organizations issued practical guidelines on the assessment of performance validity in *all* neuropsychological assessments (Moore et al., 2021; Sweet et al., 2021). Therefore, an important step towards the appropriate recognition of the importance of validity assessment in all clinical neuropsychological assessments would be the issuing of specific practical guidelines by the Dutch professional psychological associations.

## CONCLUSIONS

Nowadays, most clinically active neuropsychologists recognize the importance for examining the validity of test results, in order to place confidence in the accuracy of the obtained neuropsychological assessment data. However, there is evidence that still a significant proportion of clinicians does not assess performance validity status by default. A reasonable explanation may be that – although its importance is recognized – they find themselves struggling with the correct application of performance validity assessment. In academia and (postgraduate) training in psychology, the topic of validity assessment only plays a marginal role. Therefore, we hope that our studies provide clinicians with more knowledge, insights, and practical guidance about performance validity assessment in routine clinical care. Providing practical tools for improving performance validity status determination and indications how to approach patients who display noncredible performance, may increase clinicians' comfort in actually integrating proposed validity assessment guidelines in their daily practices. Ultimately, instead of losing empathy and abandoning attempts to provide clinical aid, non-credible responding may be perceived as the common clinically meaningful behavior that needs specific clinical attention.

## REFERENCES


Beach, S. R., Taylor, J. B., & Kontos, N. (2017). Teaching psychiatric trainees to "think dirty": Uncovering hidden motivations and deception. *Psychosomatics*, *58*(5), 474–482. https://doi.org/10.1016/j.psym.2017.04.005

Boone, K. B. (Ed.) (2007). *Assessment of feigned cognitive impairment: A neuropsychological perspective.* The Guilford Press.

Boone, K. B., Lu, P., & Herzberg, D. (2002). Rey dot counting test. Western Psychological Services.

Carone, D. A., Iverson, G. L., & Bush, S. S. (2010). A model to approaching and providing feedback to patients regarding invalid test performance in clinical neuropsychological evaluations. *The Clinical Neuropsychologist, 24*(5), 759–778, https://doi:10.1080/13854041003712951

Constantino, M. J., Vîslă, A., Coyne, A. E., & Boswell, J. F. (2018). A meta-analysis of the association between patients' early treatment outcome expectation and their posttreatment outcomes. *Psychotherapy, 55*(4), 473–485. https://doi-org.mu.idm.oclc.org/10.1037/pst0000169

Dandachi-FitzGerald, B., & Martin, P. K. (2022). Clinical judgement and clinically applied statistics: Description, benefits, and potential dangers when relying on either one individually in clinical practice. In R. W. Schroeder & P. K. Martin (Eds.), *Validity assessment in clinical neuropsychological practice; evaluating and managing noncredible performance* (pp. 107–125). The Guilford Press

Dandachi-FitzGerald, B., Merckelbach, H., Bošković, I., & Jelicic, M. (2020). Do you know people who feign? Proxy respondents about feigned symptoms. *Psychological Injury and Law, 13*(3), 225–234. https://doi.org/10.1007/s12207-020-09387-6

Dandachi-FitzGerald, B., Merckelbach, H., & Merten, T. (2022). Cry for help as a root cause of poor symptom validity: A critical note. *Applied neuropsychology. Adult*, 1–6. Advance online publication. https://doi.org/10.1080/23279095.2022.2040025

Durosini, I., & Aschieri, F. (2021). Therapeutic assessment efficacy: A meta-analysis. *Psychological Assessment*, *33*(10), 962–972. https://doi-org.mu.idm.oclc.org/10.1037/pas0001038

Goedendorp, M. M., van der Werf, S. P., Bleijenberg, G., Tummers, M., & Knoop, H. (2013). Does neuropsychological test performance predict outcome of cognitive behavior therapy for chronic fatigue syndrome and what is the role of underperformance? *Journal of Psychosomatic Research, 75*, 242–248.

Gorske, T. T., & Smith, S. R. (2008). *Collaborative therapeutic neuropsychological assessment*. Springer Science & Business Media.

Heffelfinger, A. K., Janecek, J. K., Johnson, A., Miller, L. E., Nelson, A., & Pulsipher, D. T. (2022). Competency-based assessment in clinical neuropsychology at the post-doctoral level: Stages, milestones, and benchmarks as proposed by an APPCN work group. *The Clinical Neuropsychologist*, *36*(6), 1209–1225. https://doi-org.mu.idm.oclc.org/10.1080/13854046.2020.1829070

Hirst, R. B., Han, C. S., Teague, A. M., Rosen, A. S., Gretler, J., & Quittner, Z. (2017). Adherence to validity testing recommendations in neuropsychological assessment: A survey of INS and NAN members. *Archives of Clinical Neuropsychology*, *32*(4), 456–471. https://doi.org.mu.idm.oclc.org/10.1093/arclin/acx009

Jurick, S. M., Crocker, L. D., Merritt, V. C., Hoffman, S. N., Keller, A. V., Eglit, G. M. L., Thomas, K. R., Norman, S. B., Schiehser, D. M., Rodgers, C. S., Twamley, E. W., & Jak, A. J. (2020). Psychological symptoms and rates of performance validity improve following trauma-focused Treatment in veterans with PTSD and history of mild-to-moderate TBI. *Journal of the International Neuropsychological Society, 26*(1), 108–118. https://doi.org/10.1017/S1355617719000997

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

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

Lippa, S. M., Pastorek, N. J., Romesser, J., Linck, J., Sim, A. H., Wisdom, N. M., & Miller, B. I. (2014). Ecological validity of performance validity testing. *Archives of Clinical Neuropsychology*, *29*(3), 236–244. https://doi-org.mu.idm.oclc.org/10.1093/arclin/acu002

Longley, W. A., Tate, R. L., & Brown, R. F. (2023). The psychological benefits of neuropsychological assessment feedback as a psycho-educational therapeutic intervention: A randomized-controlled trial with cross-over in multiple sclerosis. *Neuropsychological Rehabilitation*, *33*(5), 764–793. https://doi-org.mu.idm.oclc.org/10.1080/09602011.2022.2047734

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

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. mu. idm. oclc. org/ 10. 1093/ arclin/acaa0 17

Martin, P. K., & Schroeder, R. W. (2022). A framework for providing clinical feedback when patients invalidate testing. In R. W. Schroeder, & P. K. Martin (Eds.), *Validity assessment in clinical neuropsychological practice: Evaluating and managing noncredible performance* (pp. 47-69). The Guilford Press.

Martin, P. K., Schroeder, R. W., & Odland, A. P. (2015). Neuropsychologists' validity testing beliefs and practices: A survey of north American professionals. *The Clinical Neuropsychologist, 29*(6), 741-776. https://doiorg.mu.idm.oclc.org/10.1080/13854046.2015.1087597

Moore, P., Bunnage, M., Kemp, S., Dorris, L., & Baker, G. (2021). *Guidance on the assessment of performance validity in neuropsychological assessment*. The British Psychological Society.

Nederlands Instituut van Psychologen, sectie Neuropsychologie (2016). Guidelines for neuropsychological expertise.

Nederlands Instituut van Psychologen (2017). Guidelines for the use of tests.

Nelson, A. P., Roper, B. L., Slomine, B. S., Morrison, C., Greher, M. R., Janusz, J., Larson, J. C., Meadows, M. E., Ready, R. E., Rivera Mindt, M., Whiteside, D. M., Willment, K., & Wodushek, T. R. (2015). Official position of the American Academy of Clinical Neuropsychology (AACN): Guidelines for practicum training in clinical neuropsychology. *The Clinical Neuropsychologist*, *29*(7), 879–904. https://doi-org.mu.idm.oclc.org/10.1080/13854046.2015.1117658

Poston, J. M., & Hanson, W. E. (2010). Meta-analysis of psychological assessment as a therapeutic intervention. *Psychological Assessment*, *22*(*2*), 203–212. https://doi.org/10.1037/a0018679.

Rogers, R. (2008). *Clinical assessment of malingering and deception* (3rd ed.). Guilford Press

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

Schroeder, R. W., Clark, H. A., & Martin, P. K. (2022). Base rates of invalidity when patients undergoing routine clinical evaluations have social security disability as an external incentive. *The Clinical Neuropsychologist*, *36*(7), 1902–1914, https://doi-org.mu.idm.oclc.org/10.1080/13854046.2021.1895322

Schroeder, R. W. & Martin, P. K. (2022). Validity assessment in clinical settings; How it differs from forensic settings and why it is important. In R. W. Schroeder & P. K. Martin (Eds.) *Validity assessment in clinical neuropsychological practice; Evaluating and managing noncredible performance* (pp. 3-10). Guilford Press

Sherman, E. M., 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

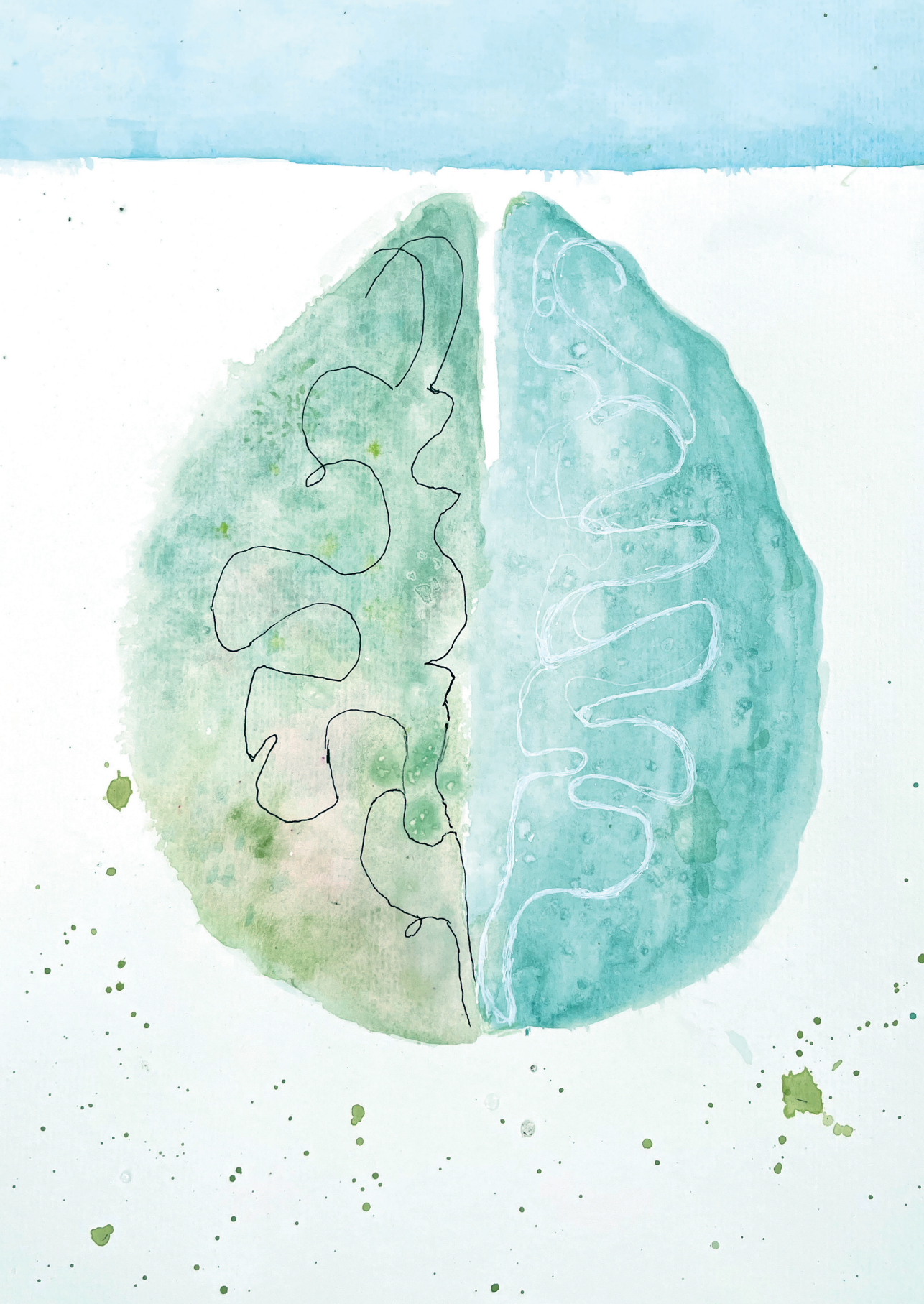
Sollman, M. J., & Berry, D. T. (2011). Detection of inadequate effort on neuropsychological testing: a meta-analytic update and extension. *Archives of Clinical Neuropsychology*, *26*(8), 774–789. https://doi-org.mu.idm.oclc.org/10.1093/arclin/acr066

Suchy, Y., Chelune, G., Franchow, E. I., & Thorgusen, S. R. (2012). Confronting patients about insufficient effort: the impact on subsequent symptom validity and memory performance. *The Clinical Neuropsychologist*, *26*(8), 1296–1311. https://doi.org/10.1080/13854046.2012.722230

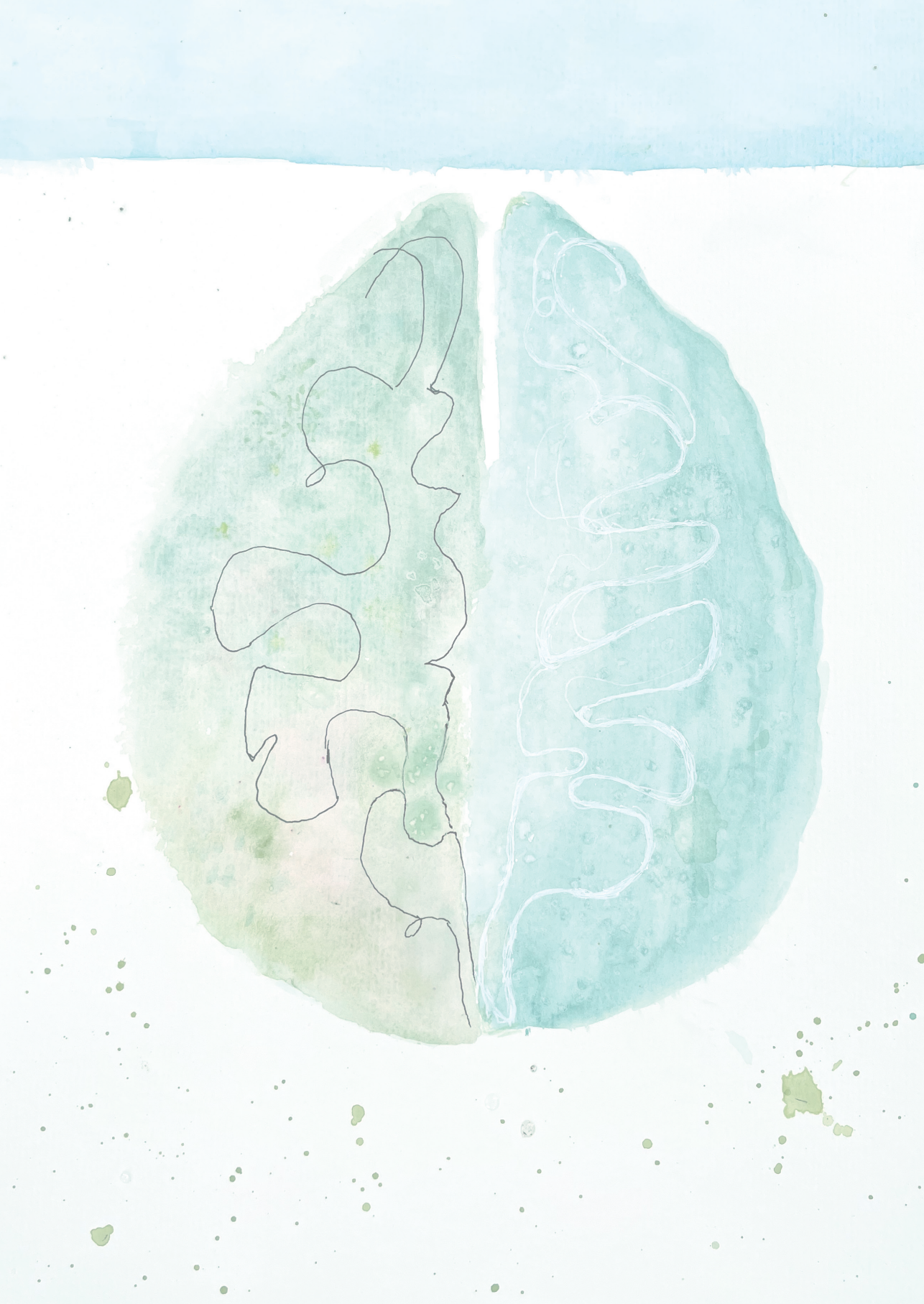
Sweet, J. J., Heilbronner, R. L., Morgan, J. E., Larrabee, G. J., Rohling, M. L., Boone, K. B., Kirkwood, M. W., Schroeder, R. W., Suhr, J. A., & Conference Participants (2021). American Academy of Clinical Neuropsychology (AACN) 2021 consensus statement on validity assessment: Update of the 2009 AACN 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

Van Egmond, J., Kummeling, I., & Balkom, T. A. (2005). Secondary gain as hidden motive for getting psychiatric treatment. *European Psychiatry*, *20*(5-6), 416–421. https://doi-org.mu.idm.oclc.org/10.1016/j.eurpsy.2004.11.012

Waldron-Perrine, B., Rai, J. K., & Chao, D. (2021). Therapeutic assessment and the art of feedback: A model for integrating evidence-based assessment and therapy techniques in neurological rehabilitation. *NeuroRehabilitation*, *49*(2), 293–306. https://doi.org/10.3233/NRE-218027



# ADDENDUM



## SUMMARY

The work in this dissertation focusses on the validity of test performance of adult patients who present for routine care in a clinical setting. The main objective of the studies described in this dissertation was to gain more insight into the prevalence rate of invalid performance, the impact of feedback interventions upon indications of invalid performance, and the relevance of performance validity to treatment outcome. These aims are addressed in the five studies of this dissertation.

**Chapter 1** introduces the evolving concept of performance validity from forensic to routine clinical care. Methods for measuring performance validity are addressed, as are the continuing questions related to performance validity assessment in routine clinical care. Finally, the aims and outlines of this dissertation are presented at the end of this chapter.

**Chapter 2** presents a case report of a patient who was referred by a Neurologist in a general hospital setting for neuropsychological assessment because of persisting cognitive complaints and fatigue. Approximately ten years earlier, this patient was diagnosed with mild cognitive impairment (MCI) by her treating medical specialist, where low performance on cognitive tests were crucial for these diagnostic conclusions. During a new neuropsychological assessment, the patient failed multiple PVTs and showed a marked discrepancy between her low-test performance and actual level of functioning. We discuss how performance validity assessment sheds a different light on het former MCI diagnosis. This case report illustrates the clinical relevance of PVT usage, but also its challenges and complications in routine clinical care.

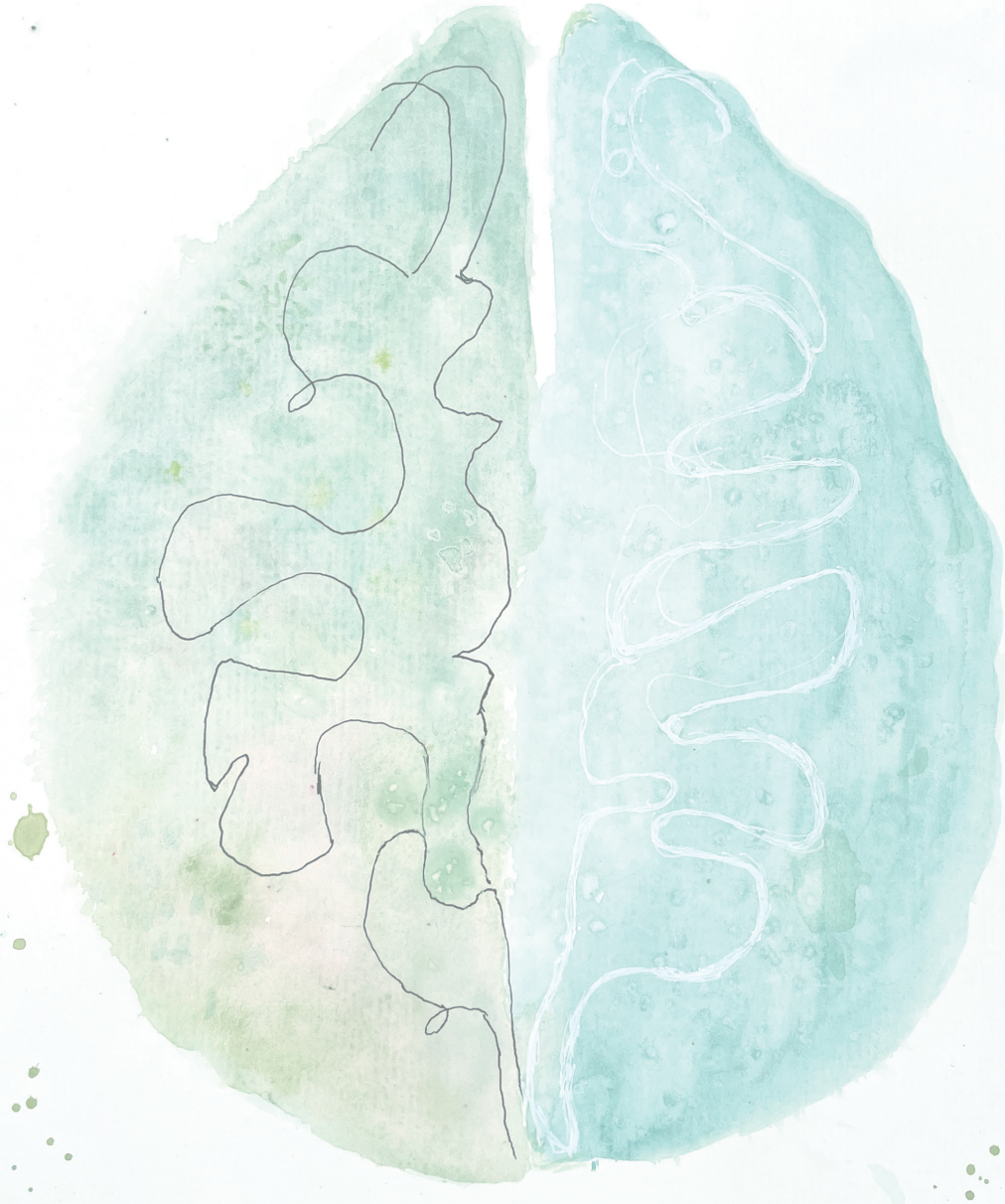
The systematic review study in **chapter 3** aimed to evaluate how often adult clinical patients fail a well-validated freestanding PVT in the context of routine clinical care. Meta-analyses were carried out to calculate pooled base rates of PVT failure, and an overall pooled PVT failure rate of 16%, 95% CI [14, 19] was found. Type of clinical context, diagnosis group, presence of external gain incentives, and psychometric properties of the utilized PVT were found to impact the rate of PVT failure. To our knowledge, this study is the first to provide high-quality information about PVT failure, which can be utilized for calculating clinically relevant statistics such as the positive/negative predictive values and likelihood ratios. Thereby, the diagnostic accuracy of performance validity can be increased for both research and clinical purposes.

In the second part of this dissertation **(chapters 4 and 5)**, we examined the impact of interventions to counter performance below best of capabilities (i.e., PVT failure). In other words, our focus was on examining how feedback interventions impact the performance on subsequent tests when a patient fails a PVT. Ultimately, such interventions might contribute to increasing the overall quality of neuropsychological assessment outcomes and therewith improve appropriate diagnostic conclusions and recommendations for treatment. For this purpose, we performed an observational cross-sectional study using retrospective data and a multicenter single-blind randomized controlled trial (RCT). In the observational study **(chapter 4),** we found that performance on a PVT equally improved during re-assessment in both the group that was provided with feedback versus the group in which invalid performance was left unaddressed. In the feedback group, a significant improvement on a repeated reaction time test was apparent compared to the no-feedback group. However, it is important

to consider that this specific result could also have been influenced by the difference in time intervals for the repeated assessments between the two groups. In the multisite RCT **(chapter 5)**, we found that a brief neutral direct feedback intervention upon PVT failure had no effects on subsequent repeated and single-administered PVT performance and standard cognitive tests. Combining the findings of chapters 4 and 5, these results suggest that there are limitations to using feedback upon indications of invalid performance if the goal is to increase patients' efforts to perform at the best of their capabilities.

In **chapter 6**, we examined the impact of performance validity on treatment outcome. Instead of employing a dichotomous pass/fail approach to PVT results, we utilized the complete range of scores from a freestanding PVT. This was done to enhance statistical power when examining its relationship with response and adherence to cognitive behavioral therapy (CBT) in patients with chronic fatigue syndrome (CFS). We found that, CFS patients with low PVT performance (i.e., higher likelihood of performance below best of capabilities) are more likely to attend fewer therapy sessions and not complete the follow-up assessment, indicative of limited adherence to treatment. However, for those patients who completed the intervention, their response to CBT was comparable to those who scored high on the ASTM, despite their initial lower performance on the PVT. Therefore, instead of being an indicator restricted to the assessment of the credibility of performance on cognitive tests, performance validity may also serve as a behavioral proxy about the level of engagement a patient has regarding a behavioral treatment intervention.

**Chapter 7** presents the general discussion of this dissertation, integrating all study results, reflecting on both the methodological strengths and weaknesses, and detailing the implications for clinical practice, education, and future research. The studies conducted in this dissertation demonstrated that invalid performance is prevalent in a substantial minority of adult patients seen for routine clinical care. The relevance of invalid performance extends beyond diagnosis and encompasses (a) management strategies for patients who show indications of noncredible performance, and (b) adherence to subsequent treatments. Our studies provide clinicians with more knowledge, insights, and practical guidance about performance validity assessment in routine clinical care. By offering practical tools for improving the determination of performance validity status and by clarifying its importance, clinicians may feel more comfortable integrating the proposed validity assessment guidelines into their daily practices. Ultimately, patients may benefit from these developments as this may lead to a specific clinical focus on noncredible performance, rather than overlooking the possibility of non-credible performance, dismissing PVT failure, or losing empathy and abandoning attempts to provide clinical aid.



## SAMENVATTING (DUTCH SUMMARY)

Neuropsychologisch onderzoek wordt verricht om het cognitieve (dis)functioneren van een patiënt in kaart te brengen en daarmee beter te begrijpen. Deze informatie wordt vervolgens gebruikt om goed onderbouwde diagnostische beslissingen te nemen en passende behandelindicaties te stellen. Van patiënten wordt verwacht dat zij goed meewerken en zich optimaal inspannen om zo goed mogelijk op cognitieve testen (zoals van geheugen, aandacht, of executieve functies) te presteren. Echter, er is een deel van de patiënten waarbij dit niet gebeurd. Dit wordt onderpresteren genoemd; het fenomeen waarbij door een ondermaatse inzet de verkregen testscores geen adequate weergave zijn van de daadwerkelijke cognitieve capaciteiten van een patiënt. Onderpresteren heeft allerhande oorzaken, zoals niet inspannen tijdens een geheugentest vanwege desinteresse of het belang van de testafname niet inzien. Andersom kan een patiënt ook de idee hebben dat het tonen van *goede* prestaties op cognitieve tests niet in zijn of haar belang is. Denk bijvoorbeeld aan een binnenkort geplande UWV-beoordeling, waar volgens de patiënt goede testprestaties haaks staan op het aantonen van cognitieve problematiek. Niet standaard op onderpresteren testen, kan resulteren in de valkuil dat de psycholoog afwijkende prestaties op cognitieve tests toeschrijft aan cognitieve stoornissen. Dit kan vervolgens resulteren in onjuiste diagnostische beslissingen en daarmee ook onjuiste behandelindicaties. Derhalve heeft onderpresteren mogelijk niet alleen impact op diagnostische beslissingen, maar ook op behandeluitkomst.

De centrale doelstelling van het onderzoek in dit proefschrift is het kaart brengen van hoe vaak onderpresteren voorkomt binnen de reguliere patiëntzorg, of en hoe clinici door toepassing van feedback-interventies onderpresteren kunnen beïnvloeden, en of onderpresteren implicaties heeft voor behandeluitkomsten. Deze doelen zijn geadresseerd in de vijf studies van dit proefschrift.

**Hoofdstuk 1** geeft een algemene inleiding over de ontwikkeling van het concept prestatievaliditeit. Waar men tot ongeveer 25 jaar geleden vooral in de forensische zorgcontext beducht was op onderpresteren, - en dit in de eerste plaats toeschreef aan het simuleren van cognitieve problemen vanwege de overduidelijke externe belangen (onder een straf uitkomen vanwege geheugenverlies, bijvoorbeeld) - heeft het belang van het standaard beoordelen van de prestatievaliditeit binnen de reguliere klinische zorg de afgelopen decennia postgevat. Ten slotte worden de doelstellingen en hoofdlijnen van dit proefschrift aan het eind van dit hoofdstuk gepresenteerd.

In **hoofdstuk 2** wordt een casus beschreven van een patiënt die, binnen een algemeen ziekenhuis, door de neuroloog werd verwezen voor neuropsychologisch onderzoek vanwege aanhoudende cognitieve klachten en vermoeidheid. Ongeveer tien jaar eerder werd bij deze patiënt door een medisch specialist een milde cognitieve stoornis (MCI –een mogelijk voorstadium van dementie) vastgesteld, waarbij lage prestaties op cognitieve tests cruciaal waren voor deze diagnostische beslissing. Tijdens een nieuwe neuropsychologische beoordeling scoorde deze patiënt afwijkend op meerdere PVT's en vertoonde ze een duidelijke discrepantie tussen haar opmerkelijk lage testprestaties en haar volledig intacte niveau van functioneren in het dagelijks leven. We bespreken hoe prestatievaliditeitsbeoordeling een ander licht werpt op de vroegere MCI-diagnose. Deze casus illustreert de klinische relevantie van PVT-toepassing, maar ook de uitdagingen en complicaties ervan in de routinematige klinische zorg.

**Hoofdstuk 3** beschrijft een systematische review en meta-analyse naar de prevalentie van PVT-falen bij volwassen patiënten die gezien worden binnen routinematige klinische zorg. Psychologen worden opgeleid in het bepalen van onderpresteren, en nemen daarvoor prestatievaliditeitstesten (PVT's) af. Echter, wanneer er geen rekening wordt gehouden met het theorema van Bayes – in dit geval de prevalentie van onderpresteren in een specifieke patiëntpopulatie -, loopt men de kans foute conclusies te trekken over de aan- of afwezigheid van onderpresteren. Bij een afwijkende PVT-score kan de clinicus concluderen dat er sprake is van onderpresteren, en dat de testuitslagen daarom invalide en onbruikbaar zijn. En dat terwijl bij een lage prevalentie in de doelpopulatie, de kans op daadwerkelijk onderpresteren laag is. Andersom is bij dezelfde afwijkende PVT-score bij een patiënt afkomstig uit een populatie waar vaak sprake is van onderpresteren, de kans op daadwerkelijk onderpresteren hoog. Als de clinicus in dit geval beslist dat er geen sprake is van onderpresteren (de patiënt leek immers zijn of haar best te doen?), kunnen afwijkende scores op cognitieve taken onterecht als bewijs voor de aanwezigheid van cognitieve stoornissen worden beschouwd. Informatie over de prevalentie van onderpresteren is daarom cruciaal voor het beoordelen van de kwaliteit van neuropsychologische testresultaten. Er werden meta-analyses uitgevoerd om de gepoolde percentages van PVT-falen te berekenen. Van de aanvankelijke 12524 geïncludeerde studies bleven er na strenge kwaliteitsselectie 47 over met daarin de gegevens 6484 individuele patiënten. De resultaten toonden dat 16% (95% CI [14, 19]) van deze volwassen patiënten die werd gezien binnen een routinematige zorgcontext, afwijkend scoorde op een PVT. Tevens werd vastgesteld dat PVT-falen afhankelijk is van diagnosegroep (MCI, PNES, ADHD, traumatisch hersenletsel, epilepsie, MS, en Parkinson), klinische context (ziekenhuis, GGZ, revalidatie, tertiaire epilepsiekliniek, en eerstelijnspraktijk), aanwezigheid van potentiële externe belangen (zoals een lopende letselschadeprocedure), en PVT (TOMM, WMT, MSVT, en VSVT). Door voor deze moderatoren de gepoolde percentages PVT falen te presenteren (Tabel 2), kan een op maat gesneden beoordeling van prestatievaliditeit worden toegepast door rekening te houden met deze relevante eigenschappen van een individuele patiënt (diagnose, klinische context, extern belang, en specifieke PVT). De uitkomsten van deze studie kunnen tevens gebruikt worden voor het berekenen van klinisch toepasbare statistiek, zoals positief voorspellende waarde (PPV), negatief voorspellende waarde (NPV), en Likelihood ratio's (LRs), om de nauwkeurigheid van prestatievaliditeitbeoordeling verder te verbeteren.

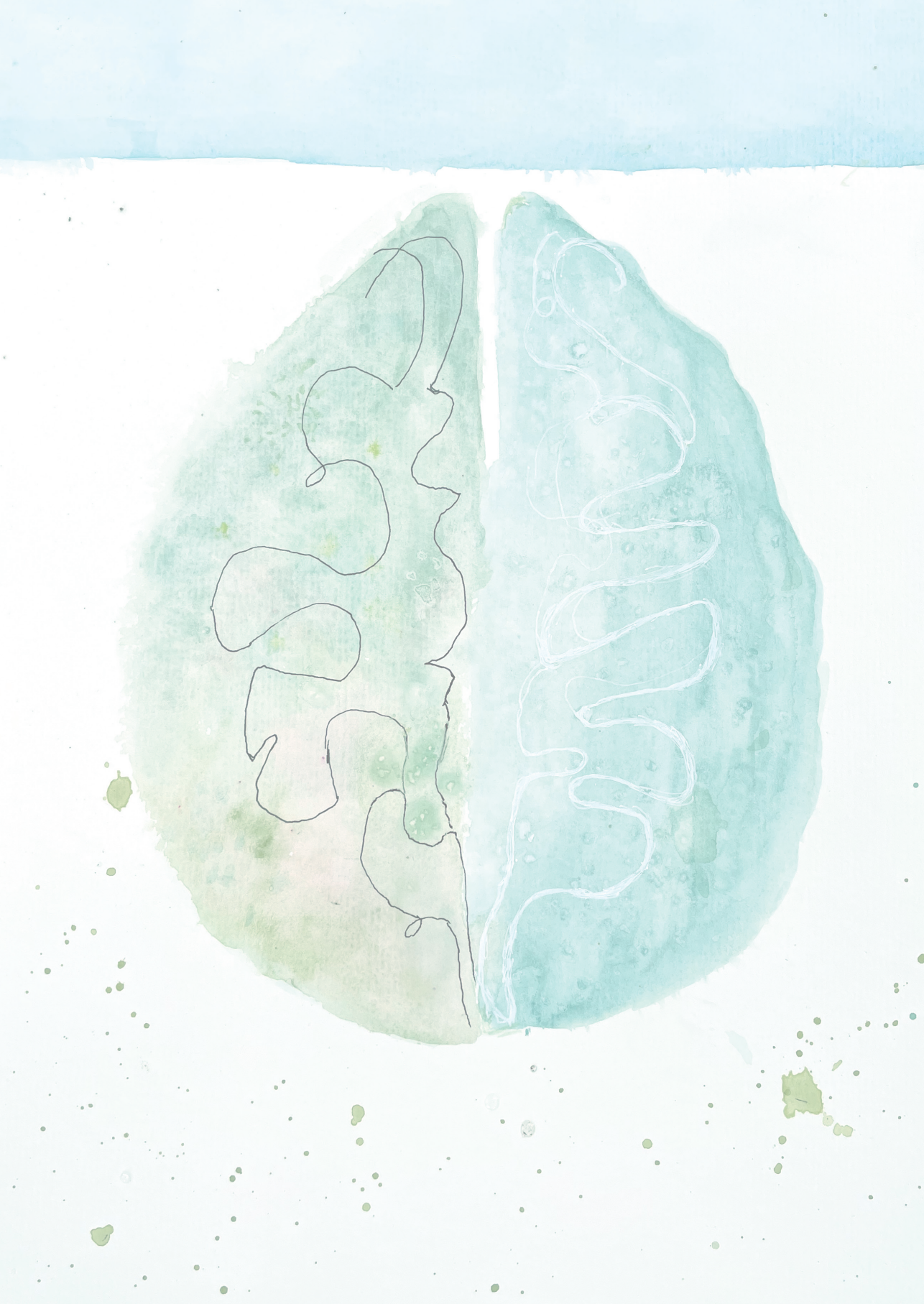
In het tweede deel van dit proefschrift worden twee studies beschreven **(hoofdstukken 4 en 5)** waarbij werd onderzocht wat de impact is van feedbackinterventies om ondermaats presteren op cognitieve tests te verbeteren. Oftewel, wanneer een patiënt onder de afkapwaarde voor onderpresteren scoorde op een PVT, werd een interventie toegepast met als doel de inzet en daarmee de testprestaties op de cognitieve tests die vervolgens werden afgenomen te verbeteren. Uiteindelijk zouden dergelijke interventies kunnen bijdragen aan het verhogen van de algehele kwaliteit van neuropsychologische testresultaten en de daarop gebaseerde diagnostische conclusies en behandeladviezen. Voor dit doel hebben we een observationeel cross-sectioneel onderzoek uitgevoerd met behulp van retrospectieve gegevens en een multicenter, enkelblind, gerandomiseerd gecontroleerd onderzoek. **Hoofdstuk 4** beschrijft de bevindingen uit het cross-sectionele onderzoek bij patiënten met de diagnose chronisch vermoeidheidssyndroom (CVS). Als gevolg van naturalistische veranderingen in de klinische procedures

kreeg een subgroep van deze patiënten feedback na PVT-falen. Tijdens de feedbackinterventie stelde de psycholoog (a) dat de CVS-symptomen van de patiënt moeilijk te beoordelen waren vanwege de lager dan verwachte prestaties op een eerder afgenomen test, (b) benadrukte dat de patiënt zijn/haar uiterste best moest doen en dat verbetering werd verwacht, en (c) legde uit dat de tests daarom herhaald moesten worden. Deze groep werd vergeleken met een vergelijkbare groep CVS-patiënten die geen feedback kreeg na PVT-falen. De resultaten van deze studie toonden dat de prestaties op een herhaalde PVT in gelijke mate verbeterden in zowel de groep die feedback kreeg als de groep waarin ongeldige testprestaties niet werden geadresseerd. In de feedbackgroep was wel een significante verbetering op één van de twee herhaalde reactietijdtests zichtbaar vergeleken met de groep zonder feedback. Een belangrijke beperking in dit onderzoek was echter het aanzienlijke verschil in de gemiddelde tijdsintervallen tussen de eerste en tweede meting voor de feedbackgroep (één maand) en de groep zonder feedback (zes maanden). In het gerandomiseerd gecontroleerde onderzoek **(hoofdstuk 5)** werden volwassen klinische patiënten bij PVT falen op willekeurige wijze toegewezen aan een korte, neutrale feedback-interventie. Tijdens de feedback-interventie, die direct volgde op het PVT-falen, werd aangegeven dat de patiënt tijdens het eerste gedeelte van het onderzoek lager presteerde dan verwacht, en dat daarom enkele testen zouden worden herhaald. De resultaten toonden dat patiënten in zowel de feedback als de niet-feedback groep gelijke prestaties lieten zien op zowel de enkelvoudig afgenomen en de herhaalde PVT's. Ook lieten de resultaten op reguliere cognitieve tests (zoals geheugen, tempo van informatieverwerking, en woordvloeiendheid) geen groepsverschillen zien. Opvallend genoeg scoorde de overgrote meerderheid (80%) van de deelnemers afwijkend op ten minste één van de twee later afgenomen PVT's, onafhankelijk van feedback. De bevindingen uit **hoofdstukken 4 en 5** suggereren dat er beperkingen zijn aan het gebruik van feedback om - bij indicatie op onderpresteren - de inspanningen van patiënten te vergroten om zo goed mogelijk te presteren op cognitieve tests.

**Hoofdstuk 6** beschrijft de resultaten van de studie naar de impact van prestatie(in)validiteit op behandelsucces. Hoewel onderzoek op dit vlak beperkt is, is het is goed voorstelbaar dat onderpresteren niet alleen relevant is voor diagnostische vraagstukken, maar ook informatief kan zijn over hoe patiënten zich verhouden ten opzichte van een behandeling. Immers, wanneer een patiënt onder zijn/haar niveau presteert op cognitieve taken zou een dergelijke suboptimale inzet ook voorzien kunnen worden tijdens een gedragsmatige behandelaanpak. Eerdere onderzoeken op dit vlak gebruikten een dichotome (onderpresteren vs. niet-onderpresteren) aanpak en vonden wisselende resultaten. Om de statistische toetsingskracht te vergroten, werd in deze studie de totale score-range op een PVT gebruikt om de relatie met therapietrouw en respons op cognitieve gedragstherapie (CGT) bij patiënten met CVS te onderzoeken. De resultaten van dit onderzoek toonden dat CVS-patiënten met lage PVT-prestaties (d.w.z. een grotere kans op presteren ónder hun feitelijke cognitieve mogelijkheden) minder therapiesessies bijwoonden en de follow-upbeoordeling niet voltooiden, wat een beperkte therapietrouw suggereert. Voor de patiënten die de behandeling afmaakten, was het behandelsucces vergelijkbaar met die van degenen die hoog scoorden op een PVT, ondanks hun aanvankelijk lagere prestaties op deze PVT. Tezamen genomen suggereren deze bevindingen dat prestatievaliditeit ook kan dienen als een gedragsmatige proxy van *hoe* patiënten deelnemen aan een gedragsmatige behandelinterventie

(sommigen kunnen bijvoorbeeld bedenkingen hebben bij de gecommuniceerde diagnoses en/of behandelplannen).

**Hoofdstuk 7** beschrijft de belangrijkste bevindingen van dit proefschrift en geeft een algemene discussie waarin alle onderzoeksresultaten zijn geïntegreerd. Ook wordt gereflecteerd op de methodologie, implicaties voor de klinische praktijk, onderwijs, en toekomstig onderzoek. De studies in dit proefschrift toonden aan dat onderpresteren voorkomt bij een aanzienlijke minderheid van de volwassen patiënten binnen een routinematig klinisch zorgkader. De relevantie van onderpresteren voor de klinische praktijk reikt verder dan het beoordelen van cognitieve testprestaties op hun geldigheid, en omvat tevens (a) manieren om middels feedbackinterventie onderpresteren te beïnvloeden, en (b) therapietrouw bij psychologische behandelingen. Deze bevindingen bieden clinici - en potentieel ook onderzoekers - meer kennis over en praktische handvatten voor de beoordeling van prestatievaliditeit in de routinematige neuropsychologische klinische zorg. Door gestratificeerde prevalentiegegevens van PVT falen aan te bieden, en de toepassing ervan toe te lichten door middel van concrete voorbeelden, kunnen clinici zich gesterkt voelen bij het integreren van vigerende richtlijnen voor prestatievaliditeitsbeoordeling in hun dagelijkse klinische werk. Uiteindelijk kunnen patiënten profiteren van deze ontwikkelingen, omdat dit kan leiden tot een specifieke klinische focus op onderpresteren, in plaats van invalide testprestaties binnen neuropsychologisch onderzoek over het hoofd te zien, deze niet (of onjuist) te gebruiken, of compassie voor de patiënt te verliezen en geen passende hulp aan te bieden.



## IMPACT PARAGRAPH

In neuropsychological assessment, performance tests (e.g., memory, attention, planning, or language) are used to assess cognitive functioning. For example, this performance-based approach is used in clinical practice for examining the consequences of a brain injury on patients' memory functioning and related learning potential. However, if patients do not perform to the best of their capabilities on these cognitive tests, this leads to invalid data and potentially inaccurate diagnostic conclusions and recommendations for treatment. This is illustrated in the case study from chapter 2, where a patient was incorrectly diagnosed with mild cognitive impairment (MCI) based upon invalid test performance. While clinical judgment alone is insufficient for determining the validity of a patients' test performance, the use of designated freestanding performance validity tests (PVTs) is essential. This dissertation focusses on (1) how often adult patients fail a PVT, (2) the impact of feedback interventions upon indications of invalid performance, and (3) the impact of performance invalidity on treatment outcome in routine clinical care.

**Main Findings**

First, a systematic review using meta-analyses was carried out to calculate pooled base rates of performance validity test (PVT) failure in adult patients seen for routine clinical care. We found an overall PVT failure rate of 16% (95% CI [14, 19]). Type of clinical context (e.g., medical hospital or mental healthcare institute), diagnosis group (e.g., ADHD or traumatic brain injury), presence of external gains (e.g., financial incentives), and psychometric properties of the utilized PVT (i.e., sensitivity and specificity) were found to impact the rate of PVT failure.

In the second part of this dissertation, we examined the impact of feedback interventions on subsequent test performance when patients failed a PVT. Such interventions might contribute to enhancing the overall quality of neuropsychological assessment outcomes and therewith improve appropriate diagnostic conclusions and treatment recommendations. We performed two studies: an observational cross-sectional study using retrospective archival data and a multicenter single-blind randomized controlled trial (RCT). In the observational study, we found that performance on a PVT equally improved during re-assessment in both the group that was provided with feedback versus the group in which invalid performance was left unaddressed. In the feedback group, a significant improvement on a repeated reaction time test was apparent compared to the no-feedback group. However, in the multisite RCT, we found that a brief neutral direct feedback intervention upon PVT failure had no effects on subsequent repeated and single-administered PVT performance or standard cognitive test performance. Combined, these results suggest that there might be limitations to using feedback upon indications of invalid performance for increasing patients' efforts to perform at the best of their capabilities.

In the final part of this dissertation, we examined the impact of performance validity on treatment outcome. Instead of employing a dichotomous pass/fail approach to PVT results, we utilized the complete range of scores from a freestanding PVT. This was done to enhance statistical power when examining its relationship with response and adherence to cognitive behavioral therapy (CBT) in patients with chronic fatigue syndrome (CFS). We found that CFS patients with low PVT performance (i.e., higher

likelihood of performance below best of capabilities) are more likely to attend fewer therapy sessions and not complete the follow-up assessment, indicative of limited adherence to treatment. However, for the for those patients who completed the intervention, their response to CBT was comparable to those who scored high on the ASTM, despite their initial lower performance on the PVT. Therefore, instead of being an indicator restricted to the assessment of the credibility of performance on cognitive tests, performance validity may also serve as a behavioral proxy about the level of engagement a patient has regarding a behavioral treatment intervention

**Scientific Impact**

Four of our five studies have been published in various international peer-reviewed journals. One study is submitted and under review. As such, our findings contribute to scientific research and clinical practice by providing freely accessible information regarding the base rate of PVT failure across relevant contextual, personal, and assessment characteristics (Table 2 from Chapter 3). These data provide clinicians and research alike with the opportunity for increasing the accuracy of performance validity determinations in neuropsychological examinations. Our studies on the effects of feedback following PVT failure and the impact of performance validity on treatment outcome, represent crucial initial steps towards advancing validity assessment in these areas. These findings provide valuable insights that may inspire future research on communicating and handling performance invalidity in clinical assessments. By shedding light on these aspects, our research contributes to the ongoing development of validity assessment practices.

**Societal Impact**

As all our studies concerned adult patients seen for routine clinical care, our study findings may have direct implications for current clinical (neuropsychological) practices. Our meta-analyzed results on how often adult clinical patients fail PVTs, can be directly implemented in both routine clinical care. To our knowledge, this study is the first to provide high-quality information about PVT failure that can be used for calculating clinically applied statistics (i.e., positive-/negative predictive values, likelihood ratios). Thereby, the diagnostic accuracy of performance validity determination can be increased for both research and clinical purposes. Illustratively, our review-study findings are currently displayed at a Dutch publishing house of commonly used PVTs (Hogrefe), highlighting its clinical implications (https://www.hogrefe.com/nl/nieuw/zijn-de-door-jou-gemeten-klachten-wel-valide). The research insights may potentially enhance the quality of diagnostic conclusions and the treatment recommendations derived from the neuropsychological assessment. Or to put in other words, misdiagnosis and inaccurate treatments may be prevented, ultimately leading to improved patient care.

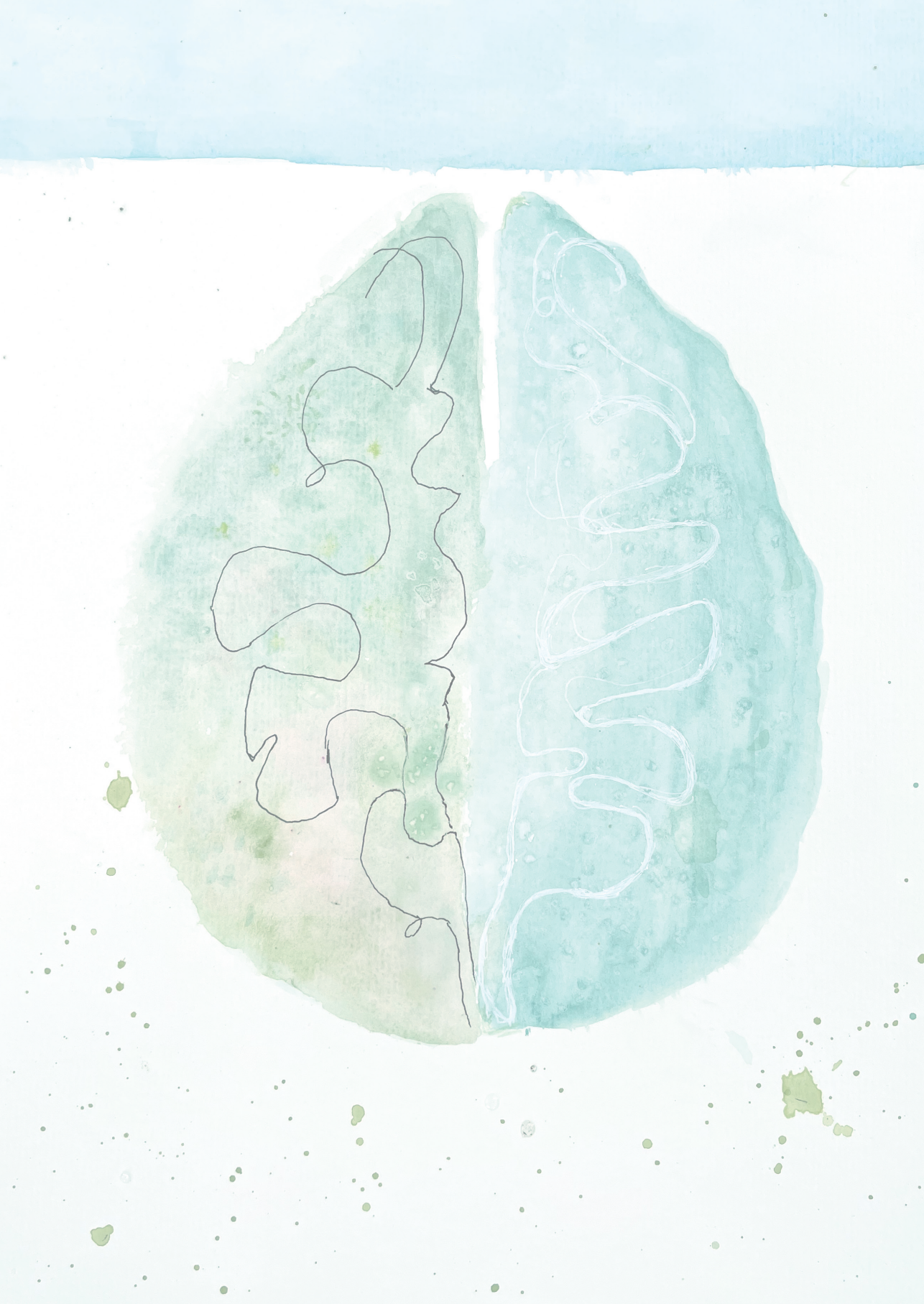
In addition, current practices on feedback strategies for improving patients' test-taking behavior were empirically tested and found to have little to no impact. This urges for additional research and alternative approaches to dealing with invalid performance in clinical patients (e.g., patients may benefit from neuropsychological assessment after, rather than before, treatment). In the meantime, the apparent lack of influence that clinicians seem to have on test-taking behavior trough feedback

interventions, underlines the importance of assessing performance validity continuously during the neuropsychological assessment and in every test session.

Lastly, as response to treatment for patients who show indications of invalid performance is comparable to subjects who performed to the best of their abilities, low PVT performance should not be a reason to exclude patients from treatment. This is an important implication, as clinicians may view this behavior as a sign of non-compliance and consequently may question whether they would benefit from costly medical treatment. We, however, did find proof for the first notion that performing low on a PVT is in fact related to limited treatment adherence (i.e., completing fewer therapy sessions and study drop-out). As such, the clinician might instead view invalid performance as a behavioral proxy of how patients engage in treatment (e.g., having reservations about the communicated therapy proposal) that may need clinical attention, instead of losing empathy and abandoning attempts to provide clinical aid.

**Dissemination Activities**

The findings from the studies in this dissertation have been communicated in various ways. The results have been presented at **national and international conferences**. For fellow **researchers and clinicians**, an introduction into the topics of this dissertation and the study findings were communicated during a **webinar** of the Limburg Brain Injury Centre (2021). **Clinicians** (neuropsychologists, technicians, and interns) involved in the multicenter randomized controlled trial from this dissertation (chapter 5) conducted in seven hospitals in the Netherlands, were trained onsite on the study procedure but also on the (developing) concepts of performance validity and related feedback interventions. The proceedings of the studies in this dissertation were shared with **clinicians** through contributions to a local **science magazine** (VieCuri Medical Center), **(invited) oral presentations** at RINO Groep Utrecht, VieCuri Medical Center, Radboud University Medical Center; departments of Psychiatry, and Medical Psychology, Psychotrauma Expertise Center (Psytrec), Dutch Institute for Forensic Psychiatry and Psychology (NIFP), and clinicians working in occupational health services. The review-study from chapter 3 was awarded with the **research prize 2023** by RINO Zuid for being the most clinically relevant of all submissions. Three studies from this dissertation were **published open access** and are therefore accessible to the general public. In addition, these open access articles were also shared via online platforms such as LinkedIn and ResearchGate. As a **trainer and supervisor** for psychologists in training to become a registered health care psychologist and clinical neuropsychologist, the topics of validity assessment, approaches on how to manage clinical patients who show non-credible responding, and its potential influence on both assessment- and treatment outcomes were specifically addressed and incorporated the study findings as mentioned in this dissertation. Finally, the study findings related to validity assessment, diagnostic decision making, and managing invalid presentations were also integrated into the curriculum of the **postdoctoral training** to become a registered health care psychologist (2-year program) and registered clinical psychologist (4-year program) at the **Radboud Centre for Social Sciences (RCSW)**, Nijmegen.



## CURRICULUM VITAE

Jeroen Roor was born in Enschede on January 28th 1980. After dropping out of senior secondary education (HAVO), he started secondary vocational education (MBO) to be a medical laboratory technician. Upon completion, and after working as a caretaker for people with physical disability, he passed the colloquium doctum at Katholieke Universiteit Nijmegen (now Radboud University) and in 2001 moved to Nijmegen to study psychology. In 2005, Jeroen completed the Interdisciplinary Honours Programme and in 2006 obtained his doctorate degree in Neuro- and Rehabilitation Psychology from Radboud University. After a brief period studying International Health at Humboldt Universität and Charité Universitätsmedizin in Berlin, he returned to the Netherlands and started working as a psychologist in mental healthcare. In 2010, he completed his postdoctoral training as a health care psychologist at Medisch Centrum Alkmaar. In 2019, he completed his postdoctoral training as a clinical neuropsychologist at VieCuri Medisch Centrum. In 2015, Jeroen started his PhD project at Maastricht University on performance validity in clinical patients under the supervision of prof. Rudolf Ponds, prof. Brechje Dandachi-FitzGerald, and dr. Maarten Peters. Jeroen currently works as a clinical neuropsychologist, supervisor, and researcher at VieCuri Medical Center. He teaches at Radboud Centrum voor Sociale Wetenschappen (RCSW), and holds a private practice for forensic and medico-legal neuropsychological assessment. Jeroen lives in Nijmegen together with Evelien and their two sons Kees and Willem. He is a passionate trail runner and an outdoor enthusiast.

**Publications and Presentations**

***Publications***

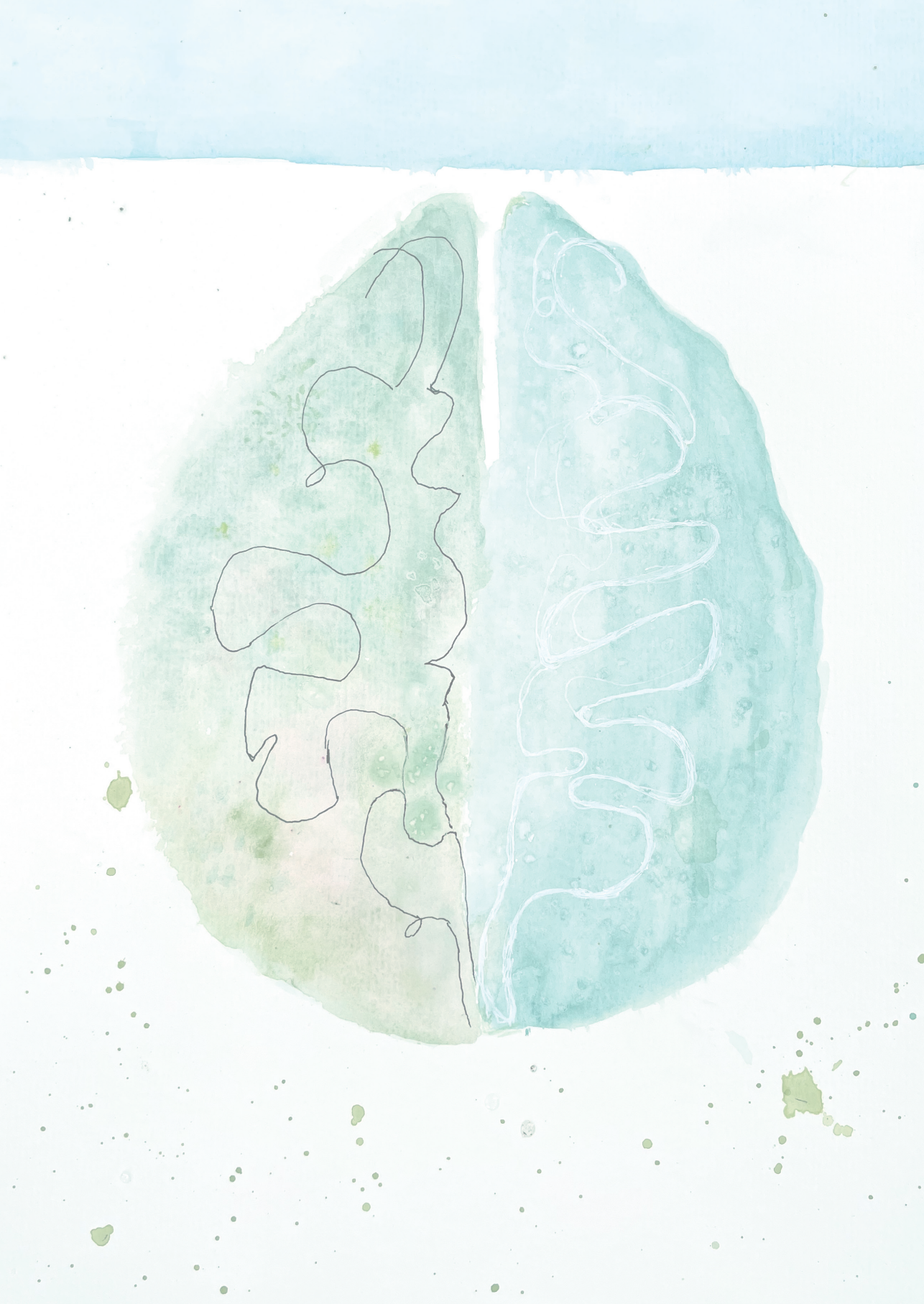
- Roor, J. J., Dandachi-FitzGerald, B., & Ponds, R. W. H. M. (2016). A case of misdiagnosis of mild cognitive impairment: The utility of symptom validity testing in an outpatient memory clinic. *Applied Neuropsychology: Adult*, *23*(3), 172–178. https://doi.org/10.1080/23279095.2015.1030018
- Roor, J. J., Knoop, H., Dandachi-FitzGerald, B., Peters, M. J. V., Bleijenberg, G., & Ponds, R. W. H. M. (2020). Feedback on underperformance in patients with chronic fatigue syndrome: The impact on subsequent neuropsychological test performance. *Applied Neuropsychology: Adult*, *27*(2), 188–196. https://doi.org/10.1080/23279095.2018.1519509
- Roor, J. J., Dandachi-FitzGerald, B., Peters, M. J. V., Knoop, H., & Ponds, R. W. H. M. (2022). Performance validity and outcome of cognitive behavior therapy in patients with chronic fatigue syndrome. *Journal of the International Neuropsychological Society*, *28*(5), 473–482. https://doi.org/10.1017/S1355617721000643
- Roor, J. J., Peters, M. J. V., Dandachi-FitzGerald, B., & Ponds, R. W. H. M. (2023). Performance Validity Test Failure in the Clinical Population: A systematic review and meta-analysis of prevalence rates. *Neuropsychology Review*, 10.1007/s11065-023-09582-7. Advance online publication. https://doi.org/10.1007/s11065-023-09582-7
- Roor, J. J., Dandachi-FitzGerald, B., Peters, M. J. V., & Ponds, R. W. H. M. No impact of a clinical feedback intervention when patients invalidate testing; A multi-site, single-blind randomized controlled trial. (Under review).

***Presentations at International Conferences***

- Roor, J. J. & van Kempen, C. (2012). *Onverklaarde cognitieve klachten: Somatisatie in de neuropsychologie*? Workshop Davos Sessies. Landelijke Vereniging voor Medische Psychologie (LVMP). Davos, Switzerland.
- Ponds, R. H. W., de Jonghe, J. F. M., Roor, J. J., Niesten, I., Dandachi-FitzGerald, B., & Merckelbach, H. L. G. J. (2016, July 6-8). Symposium session: *Symptom validity: The blurred lines between "crooks" and genuine patients* [Conference presentation]. INS 2016 Mid-Year Meeting, London, United Kingdom. https://www.the-ins.org/meetings/ldn2016/
- Roor, J. J. (2017, July 22-23). *Feedback on performance validity test failure in patients with chronic fatigue syndrome: The impact on subsequent neuropsychological test performance and treatment outcome* [Conference presentation]. 5th European Conference on Symptom Validity Assessment. Basel, Switzerland.
- Roor, J. J., Boone, K., Suhr, J., & Ponds, R. W. H., & Kessels, R. P. C. (2019, February 20-23). Symposium session: *The Evolution of the concept of performance validity: From malingering to illness behaviors in the clinical context* [Conference presentation]. 47th Annual Meeting INS, New York City, USA. https://www.the-ins.org/files/meetings/ny2019/

***Presentations at National Conferences***

- Roor, J. J. (2010). *Cogniforme stoornis, een welkome aanvulling?* [Conference presentation]. .1e Symposium Neuropsychologie, Medisch Centrum Alkmaar.
- Roor, J. J., Dandachi-FitzGerald, B., & Ponds, R. W. H. M. (2015). *Misdiagnosis of mild cognitive impairment (MCI) in a memory clinic: a case report* [Poster session]. 4th European Conference on Symptom Validity Assessment, Maastricht.
- Roor, J. J. (2016). *Onderpresteren bespreken? Doen!* [Conference presentation]. 2e Symposium Neuropsychologie, Medisch Centrum Alkmaar.
- Roor, J. J., Dandachi-FitzGerald, B. Peters, M. J. V., & Ponds, R. W. H. M. (2019, July 4-5). *Feedback on invalid performance in the clinic; Preliminary RCT Results* [Conference presentation]. 6th European Conference on Symptom Validity Assessment, Amsterdam.
- Roor, J. J., Meyer, S., Schoemaker, T., van Leeuwen, M., & Ponds, R. W. H. M. (2019, June 19-22). Symposium session: *Symptom validity in the clinical context* [Conference presentation]. 7th Annual Scientific of the European Association of Psychosomatic Medicine, Rotterdam, The Netherlands. https://www.eapm.eu.com/event/eapm-conference-rotterdam-2019/



## DANKWOORD

Wat een tocht! Na acht jaar is het proefschrift eindelijk af. Als buitenpromovendus is het vaak zoeken geweest naar het creëren van ruimte en mogelijkheden om het onderzoek voort te zetten. Ik ben dan ook blij dat ik kan zeggen dat dat is gelukt.

Werken aan mijn proefschrift deed ik grotendeels tussen het kinderspeelgoed en de was op zolder, aan de keukentafel, of in de Universiteitsbibliotheek. Ondanks het solistische karakter van al die onderzoeksuren - wat ik eigenlijk wel prettig vond naast mijn werkzaamheden in het ziekenhuis -, stond ik er gelukkig niet alleen voor. Mijn dank gaat dan ook uit naar iedereen die op enigerlei wijze heeft bijgedragen aan het verwezenlijken van dit proefschrift en mij daarin heeft gesteund.

Allereerst bedank ik alle deelnemers die aan de onderzoeken hebben deelgenomen. Zonder jullie geen data, geen onderzoek, en geen proefschrift.

Ook wil ik hierbij graag mijn collega-psychologen en psychologisch medewerkers bedanken die – ondanks een krappe agenda en bijkomende administratieve rompslomp – belangeloos tijd investeerden om de verschillende studies in dit proefschrift mogelijk te maken.

Mijn promotoren Rudolf Ponds en Brechje Dandachi-FitzGerald en copromotor Maarten Peters wil ik bedanken voor hun begeleiding. Rudolf, vanaf de start was je positief-kritisch, en warm betrokken bij mij en mijn onderzoeksplannen. Na onze overleggen kon ik altijd met een gerustgesteld gevoel en nieuwe energie verder. Je humor, vertrouwen in en geduld met mij hebben daar zeker aan bijgedragen. Brechje, wat heb ik veel van je geleerd en wat ben ik onder de indruk van de manier waarop je alle professionele taken die je op je bord hebt uitvoert en combineert. Je enthousiasme en gedrevenheid hebben bijgedragen aan mijn ontwikkeling als onderzoeker, en aan de voldoeding en plezier die ik aan het onderzoek doen heb beleefd. Maarten, tijdens onze overleggen was je positief over de vorderingen van het promotietraject (ook als die nog niet zo zichtbaar waren), wat mij vertrouwen gaf om door te zetten. Tevens gaf je toelichting op de ongeschreven regels die bij het verrichten van wetenschappelijk onderzoek komen kijken. Daarmee had je aandacht en zorg voor mij als beginnend onderzoeker. Dank daarvoor.

Via deze weg wil ik ook alle leden van de beoordelingscommissie en alle leden van de promotiecommissie bedanken voor het lezen en beoordelen van dit proefschrift.

Coauteurs Hans Knoop en Gijs Bleijenberg wil ik hierbij bedanken voor de prettige samenwerking. Hans, speciale dank voor je vertrouwen in onze samenwerking en de moeite die je daarin hebt gestoken.

Ron Mengelers en Nico Rozendaal, bedankt voor jullie ondersteuning bij het uitvoeren van onze randomized controlled trial. Jullie vlotte reageren en heldere uitleg hebben geholpen in de het opzetten van een goed geoliede online dataverzameling.

Paranimfen Peter en Rogier, wat fijn dat jullie naast mij willen staan 19 januari. Peter, het is leuk en uniek om onze gedeelde interesse in symptoom- en prestatievaliditeit te bespreken tijdens gezamenlijke lange duurlopen of rondjes over de N70. Rogier, als collega-scepticus zitten we meestal snel op één lijn als we ontwikkelingen binnen ons vakgebied bespreken. In combinatie met je absurdistische en soms onnavolgbare humor is dat een waar feestje. Dank allebei voor jullie vriendschap.

Collega's van de afdeling Medische Psychologie van het VieCuri Medisch Centrum, jullie waren de afgelopen jaren mijn steun en toeverlaat. Als buitenpromovendus, en daarmee op flinke afstand van de Universiteit, kon ik bij jullie terecht voor het bespreken van mijn plannen, vorderingen, en tegenslagen op onderzoeksgebied. Eric van Balen, je was niet alleen mijn opleider maar ook daarna betrokken en richtinggevend in mijn professionele ontwikkeling. Speciale dank voor je steun, je bemoedigende woorden en tips die mij verder hebben geholpen om het onderzoek doorgang te laten vinden en vorm te geven. Esther Castermans, dank voor je onvoorwaardelijke ondersteuning en gezelligheid. Simone Traa en Sylvia Raaijmakers, dank voor het meedenken in het creëren van de randvoorwaarden om wetenschappelijk onderzoek binnen onze afdeling en in het ziekenhuis mogelijk te maken.

Bedankt ook alle andere Viecurianen voor jullie interesse in mijn promotieonderzoek, in het bijzonder de collega's van het Wetenschapsbureau waar ik met vragen over statistiek of datamanagement altijd terecht kon.

Ook wil ik Caroline van Heugten hierbij bedanken voor haar vertrouwen in mij en in onze samenwerking binnen het Expertisecentrum Hersenletsel Limburg (EHL). Samen met Jolein van Leeuwen-Manders, Robin van Pinxteren, Gisela Claessens, Dennis Barten, en Ellen Notting hebben we al mooie stappen gezet in het onderzoek naar licht traumatisch schedel-hersenletsel (LTSH). Ik hoop dat we nog lang zullen samenwerken.

Jos de Jonghe, van jou heb ik veel geleerd tijdens mijn eerste stappen als neuropsycholoog. Ook wakkerde jij mijn interesse in onderzoek aan. Dank daarvoor. Met veel plezier kijk ik terug op onze samenwerking.

Het besluit om een promotieonderzoek te starten kwam tijdens mijn opleiding tot klinisch neuropsycholoog. De KNP14 opleidingsgroep, bestaande uit Laura, Michel, Dymphie, Hanneke, Olga, Carla, Martin †, Kim, Gwenny, Willemijn, en Yvette wil ik bedanken voor hun betrokkenheid en het sparren over onderzoek doen en alles wat daarbij komt kijken.

Tjerk Schoemaker, met veel plezier denk ik terug aan onze gezamenlijke buitenlandse congresbezoeken. Zoals toen we uit interesse ons lieten doormeten door de Church of Scientology, dichtbij Times Square. Hun psychologische testen wezen uit dat we beiden een zeer defect karakter hebben. Derhalve werd ons geadviseerd om intensieve trainingen volgen, om er nog wat van te maken. Daarbij zagen we ook dat normscores ontbraken, IQ-scores van 0 en boven de 200 mogelijk waren, en dat nagenoeg elke testscore aanleiding gaf tot 'urgent clinical attention' ☺.

Marc, Arno, Jelle, Anne, Piotr, Menso, en Jorik, dank voor jullie jarenlange vriendschap. Ik hoop nog lang met jullie naar de kroeg te gaan, een concert te bezoeken, buitenlandse tripjes naar de bergen te maken, af te spreken voor een rondje rennen of voor koffie/een hapje eten in de stad.

Arnoud Roor †, ik ben dankbaar voor je steun in mijn toen onverwachte en ook wat ongeloofwaardige plannen om te gaan studeren aan de Universiteit. Ter voorbereiding op een toets om tot de Universiteit toegelaten te worden (colloquium doctum), heb ik dankbaar gebruik gemaakt van je uitzonderlijke kennis op het gebied van wiskunde en in het bijzonder van kansberekening. Met veel plezier kijk ik terug op de tijd dat ik bij jou in Amsterdam mocht bivakkeren. Je geweldig grappige anekdotes over je ervaringen bij het toepassen van die wiskundige inzichten in de praktijk zullen mij altijd bijblijven. Wat zou je trots zijn geweest en wat had je dit allemaal prachtig gevonden. Je zou de show hebben gestolen op de dansvloer tijdens het feest 's avonds ☺.

Lieve pap en mam, eindelijk is mijn 'scriptie' af ☺. Dank voor jullie steun en interesse in het onderzoek. Speciale dank mam voor het mooie schilderij dat is gebruikt als omslag voor dit proefschrift.

Lieve Janneke en Karlijn, wat fijn te ervaren dat jullie zo trots zijn op mij.

En natuurlijk allerliefste Evelien, Kees, en Willem! Evelien, je hebt me altijd gesteund in mijn werk en onderzoek, en hebt (soms meer dan ikzelf) het volste vertrouwen in mij. Het was bijzonder om de afgelopen periode op de vrijdagen samen aan de keukentafel te zitten werken aan ons beider onderzoek, en tijdens een lunchwandelingetje daarover te sparren. Kees en Willem, bij de start van mijn promotieonderzoek (Willem was net geboren, Kees was toen 2 jaar) nam ik jullie wel eens in een duo kinderwagen mee naar het Radboudumc als ik overleg had over het onderzoek bij CVS-patiënten. Inmiddels zijn jullie al grote kerels geworden, en hebben jullie de afgelopen jaren weinig meegekregen van alle uren dat ik bezig was met mijn promotieonderzoek. Behalve als jullie niet op een scherm mochten en ik wel ☺. Ik ben dankbaar dat jullie in mijn leven zijn!