Psychological Injury and Law (2026) 19:6
https://doi.org/10.1007/s12207-026-09557-y

Criteria-Based Validity Assessment in Legal Cases of Claimed Reduced Work Capacity

N. Klett[1,4] · R. Dohrenbusch[2] · A. Fischer[1] · T. Geiger[1] · F. Keller[3] · J. Kornhuber[4] · O. Littke[1] · A. Schütz[5] · E. M. Siegmann[4] · W. Käfferlein[1] · T. Grömer[1,4] · E. S. Capito[1]

Received: 15 October 2025 / Accepted: 5 January 2026 / Published online: 28 January 2026

Abstract
Medical Expert Witness Assessments (MEWA) are international standard to generate information involving medical questions in litigations, such as the capacity to work. While this procedure is widely utilized and guidelines for assessing the validity of symptoms using psychometric tools and non-psychometric criteria have been developed, the scientific foundation of this multimodal Criteria-Based Validity Assessment (CVA) is weak. This study aims to provide empirical validation of CVA using psychometric Symptom (SVT) and Performance Validity Tests (PVT) as areference point. 466 MEWA conducted in the law of the German Statutory Pension Insurance (GPI), all uniformly having addressed the question of the capacity to work, were analyzed. Information about scores regarding the Structured Inventory of Malingered Symptomatology (SIMS), Amsterdam Short-Term Memory Test (ASTM) aswell as the seven CVA criteria were extracted. A logistic regression using CVA data to group the MEWA into plausible and implausible (over- and/or under reporting of symptoms) cases showed a significant association between implausible cases and SIMS scores (*OR*= 1. 067; *95%- CI* [1.037, 1.098]; *p*<. 0 01) as well as ASTM scores (*OR*= 0.965; *95%- CI* [0.936, 0.994]; *p*<.05). In this highly comparable real-world dataset, we find evidence that CVA is indeed a valid tool in terms of convergent validity using a SVT and PVT as the point of reference and provide descriptive data on SVT and PVT that facilitates the interpretation of psychometric test results in such cases.

Keywords Criteria-Based validity assessment · Validation · Medical expert witness assessment · Pension insurance · Symptom validity test · Performance validity test

Introduction

Medical Expert Witness Assessments (MEWA) conducted by physicians are a standard procedure in German litigations. These independent medical evaluations provide administrative officials and judges with information on an individual's diseases, disabilities and functioning. The most frequently encountered issue in such assessments is the evaluation of an individual's capacity to work. When a reduced capacity to work is identified, financial support is provided through the national pension insurance system, the legal framework governing the German Statutory Pension Insurance (GPI). Since 2001, mental disorders have been the most common reason for granting a pension in Germany. While in 2000 only 24.2% of all pensions were granted due to mental impairments, this figure rose to 42.7% in 2018. Main causes are affective and anxiety disorders with a rising tendency (von Kardorff et al., 2020). In 2022, 163,907

T. Grömer and E. S. Capito contributed equally to this work.

✉ N. Klett
noah.klett@gmx.de; sekretariat@bamberg-neurologie.de

1 Practice Clinic for Neurology, Psychiatry, Psychosomatic Medicine and Psychotherapy, Bamberg 96047, Germany

2 University of Bonn, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn 53113, Germany

3 Thüringer Landessozialgericht, Erfurt 99092, Germany

4 Department of Psychiatry and Psychotherapy: Friedrich-Alexander-Universität Erlangen-Nürnberg, FAU Erlangen Nuremberg, Erlangen 91054, Germany

5 University of Bamberg, Bamberg 96047, Germany

pensions were awarded due to diminished work capacity caused by medical conditions. Concurrently, 338,014 applications for a pension due to reduced work capacity were filed (Grundsatz- und Querschnittsabteilung: Finanzen und Statistik, 2023, p. 67). Accordingly, MEWA exert considerable influence on the outcomes of social law cases. Aside from their application in social law, MEWA are also used to decide legal cases concerning the capacity to work in specific occupations under occupational disability insurance or matters of compensation for mental or physical injuries.

In 2019, a comprehensive practice guideline for MEWA in psychiatric and psychosomatic disorders was published by an alliance of scientific medical organizations (AWMF, 2019). According to the guideline, evaluating seven criteria (Table 1) is recommended to assess the validity of a case. We refer to this recommendation as Criteria-Based Validity Assessment (CVA). CVA aligns with the long-standing tradition of using experience-based criteria for case validity regarding response styles (Bianchini et al., 2005). Consistent with the Official Position of the American Academy of Clinical Neuropsychology (Chafetz et al., 2015), this assessment of feigning is inherently probabilistic and requires the integration of multiple, independent validity indicators rather than reliance on single test results, as has been the core principle of neuropsychological assessments for years (Chafetz, 2011). In this context, CVA can be conceptualized as a structured, case-level framework for aggregating psychometric and non-psychometric validity information, in line with current AACN recommendations. CVA is commonly employed in MEWA as well as in other areas, such as pain evaluation (AWMF, 2023).

The presence of external incentives (such as financial gain or relief from mandatory job-seeking) in GPI-related cases makes the occurrence of various types of malingering (Young et al., 2025a, b) expectedly common. Due to the uncertainty in the attribution of motivation in GPI-related cases and CVA acting solely as a measure of response validity, the term 'feigning' will be used in the following and 'invalid response set' for Symptom Validity Tests (SVT) and Performance Validity Tests (PVT) showing an invalid or false representation of the neuropsychological profile. Response styles in forensic assessments vary and can be further categorized into two groups: overstated pathology (symptom overreport) and simulated adjustment (symptom underreport) (Rogers & Bender, 2018). Additionally, mixed response styles in the form of overreporting of some symptoms and underreporting of others can also occur (Boskovic et al., 2024).

The general trend of validation research using experimental feigning designs shows that SVT and PVT are evaluated using participants trained to simulate severe illness. As for applying these findings to MEWA in litigation, a central problem is that the research conditions are not replicated in actual legal scenarios. In legal cases, even 'controls' — those who do not meaningfully feign their symptoms — nonetheless have a vested interest in receiving a determination of incapacity to work. Thus, the presentation and report of symptoms by these individuals cannot be assumed to be as neutral as in a more artificial study situation. In addition, some SVT like the Structured Inventory of Malingered Symptomatology (SIMS) contain items regarding genuine symptoms, particularly regarding affective disorders (van Impelen et al., 2014), potentially leading to elevated scores in the German MEWA population, which often experiences such disorders (von Kardorff et al., 2020). Because of this, in MEWA, cut-off values for SVT and PVT, even below the recommended values, do not justify the classification of one's response style as invalid and should always be accompanied by converging evidence from additional validity indicators (Bush et al., 2005; Young et al., 2025a). This shows a discrepancy between validation research and real-world applicability of SVT and PVT.

Among the seven validity criteria, six do not involve psychometric testing. As a result, data derived from legal cases in which CVA is applied, including SVT and PVT, provides a unique benefit: it yields evidence on feigning under real-world conditions and, at the same time, makes it possible to evaluate how effectively SVT and PVT identify feigning. In a legal setting, the nature of the response style often depends on the claimant's specific aims. Pension-related disputes

Table 1 Criteria for Criteria-Based validity assessment (CVA) in MEWA

Criterion	Discrepancies the medical expert witness is recommended to evaluate
1	Discrepancies between the subjectively reported intensity of the complaints and the vagueness with which they are described.
2	Discrepancies between severe subjective complaints (including self-assessments in questionnaires) and the observable physical and psychological impairments noted during the clinical examination.
3	Discrepancies between self-reported information and information from third-party reports (including the documented medical history)
4	Discrepancies between severe subjective impairment and a largely intact level of psychosocial functioning when coping with everyday life
5	Discrepancies between the extent of the complaints described and the intensity of previous use of therapeutic help
6	Discrepancies between the recognizable clinical picture and the results in self-assessment scales and/or psychometric tests (including SVT)
7	Discrepancies between the medications that were reported to have been taken at the time of the examination and a lack of evidence in the blood serum

typically feature attempts to appear broadly ill, with a recent study showing that 87% of work disability claimants reported to have multiple problems regarding mental (76%) and physical symptoms (76%) (Brongers et al., 2022). In theory, due to the underlying incentives, symptom overreport is generally anticipated to be the predominant strategy. In contrast, accident-related cases frequently involve one or few specific disorders like Post-Traumatic Stress Disorder (PTSD) or Complex Regional Pain Syndrome (CRPS). While most SVT generally excel at uncovering either symptom over- or underreporting, they may struggle with mixed response styles, which can include overreporting of some symptoms and simultaneously underreporting of other symptoms, therefore, presenting a genuine clinical picture (Boskovic et al., 2024). Consequently, these mixed response styles can evade detection by SVT. Although this has yet to be confirmed, combining various categories of legal cases—such as those involving accident outcomes and those regarding work capacity—risks complicating the interpretation of results. From a methodological standpoint, it is thus preferable to rely on a dataset that does not merely contain MEWA, but specifically MEWA focusing on the same core question.

In summary, even though MEWA are standard practice and play a key role in clarifying medical facts, data on non-psychometric validity criteria as well as on the use of SVT and PVT in MEWA for legal cases are lacking, and no study has yet validated this specific CVA. We hypothesized that the result of non-psychometric validity criteria and an empirically validated SVT would show a strong correlation, demonstrating convergent validity. Additionally, we posited that, due to SVT measuring symptom report validity and PVT measuring cognitive performance validity, therefore describing different aspects of invalid responding (Boe & Evald, 2022; Young et al., 2025a), the non-psychometric validity criteria would have a weaker association with PVT. To examine this, we analyzed CVA data including SVT and PVT of approximately 1% of the annual volume of MEWA concerning GPI-related questions handled by national social courts, about 2% of the annual volume in the area of mental disorders.

Materials and methods

Data

This study screened 933 MEWA performed from 23.03.2019 to 17.10.2023 at the Institute for Neurological-Psychiatric Assessments in Bamberg, Germany. Each document typically spanned 40–50 pages and contained a detailed case history including anamnesis, medical findings, psychometric test reports, an evaluation of plausibility and consistency of response behavior incorporating psychometric test results within the framework of CVA as well as diagnoses and responses to the specific legal questions. Of the 933 MEWA initially reviewed, 482 were deemed relevant as they addressed work capacity issues under GPI law (§ 43 SGB VI). Among these, 16 were excluded for reasons indicated in Fig. 1, leaving a final dataset of 466 cases. The study received approval from the ethics committee of the Otto-Friedrich-University of Bamberg on 21.03.2020 (No. 2020-02/07).

Psychometric Tests

Amsterdam Short-Term Memory Test (ASTM/AKGT) (Schagen et al., 1997). This PVT detects feigning and insufficient motivation to perform. It is employed in neuropsychological evaluations of individuals who report memory and/or concentration issues that are not clinically evident. Test-takers are asked to memorize five words, then complete a simple addition task before being presented with five words again—three of which were previously shown and must be recognized. This task is repeated 30 times. The total number of correctly recognized words is then summed, with a minimum score of 30 and maximum score of 90. A score lower than 86 points indicates a performance below actual performance levels and the lower the sum score the more conspicuous the result. To visualize the difference in case plausibility frequencies, especially below a score of 86, and to adhere to an unbiased analysis regarding CVA as well as to not rely on a priori assumptions about cut-offs, ASTM scores were grouped into 10-step increment categories (score categories: ≤50, 51–60, 61–70, 71–80, >80).

Structured Inventory of Malingered Symptomatology (SIMS/SFSS) (Cima et al., 2003). The SIMS is a self-report measure designed to detect overreporting of symptoms. Participants receive a list of 75 statements and respond with either 'yes' (true) or 'no' (false), and the number of responses indicative of feigning is summed up. A score of ≥17 is indicative of feigning. To visualize the difference in plausibility frequencies, SIMS scores were also grouped into categories (≤15, 16–20, 21–25, 26–30, >30).

CVA Criteria Rating

The validity criteria (AWMF, 2019) (Table 1) were extracted from each MEWA using a structured rating process. 50 cases were also rated by two additional raters to calculate the inter-rater reliability. After the calculation of the inter-rater reliability, the ratings of the first rater were used for

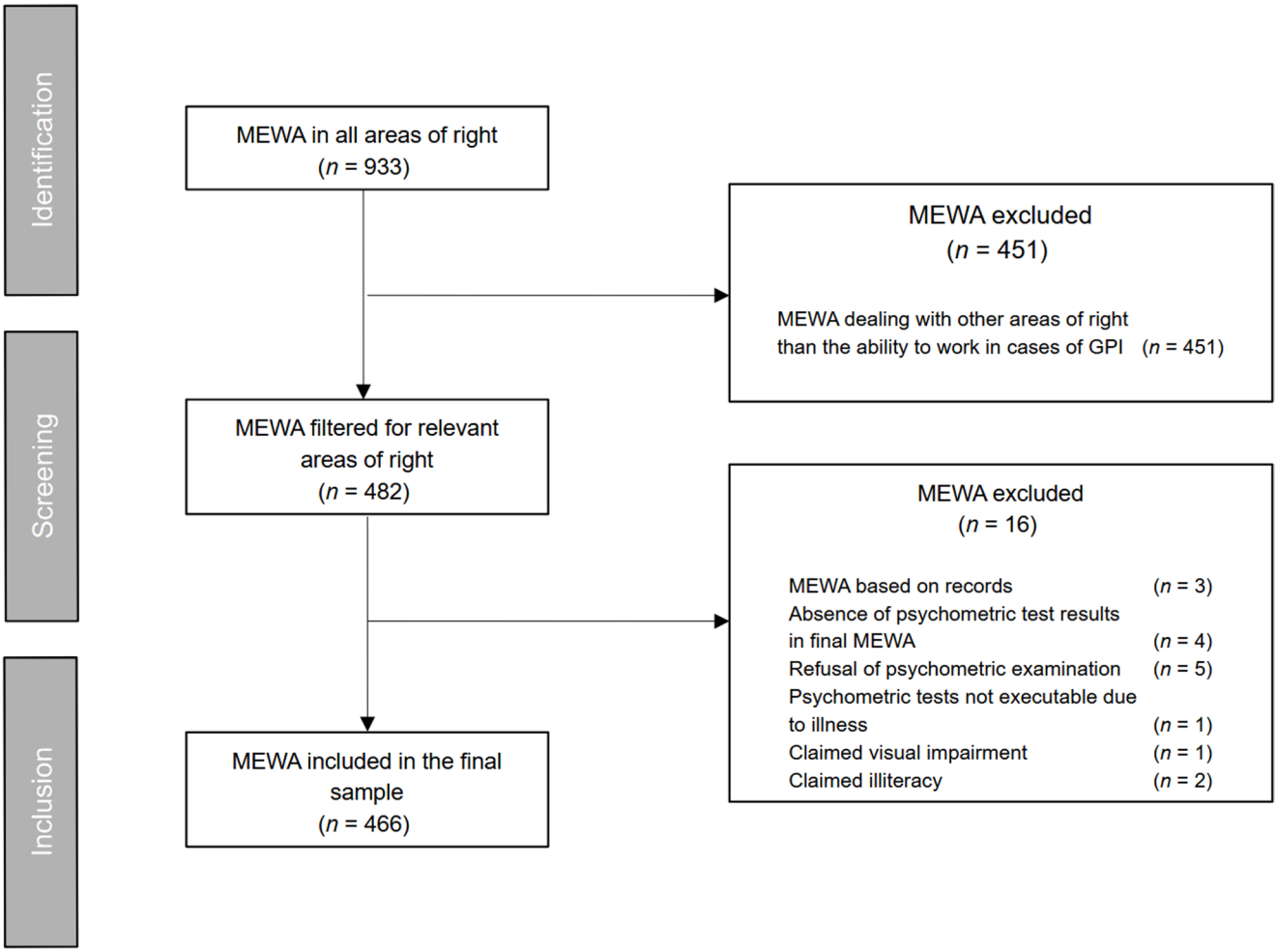

Fig. 1 Flowchart regarding all MEWA screened and included in the final sample

analysis. Criterion 6 (C6), which concerns psychometry itself, was excluded from the analysis to avoid any artificial correlations in subsequent regression analysis, due to the results of the SIMS and ASTM being used to rate criterion C6. Going forward, the remaining six CVA criteria will be referred to as non-psychometric validity criteria.

By searching for predefined key phrases or values within specified sections of the MEWA, the rater assigned values from 0 to 2 for each criterion (0 = not fulfilled/inconspicuous, i.e., no discrepancies; 1 = fulfilled/conspicuous, i.e., discrepancies exist; 2 = not evaluated). The rater instructions are available in Supplemental Material A. A conspicuous criterion indicated that a discrepancy was identified between the individual's symptom reports and the objective evaluation.

Table 2 displays the inter-rater reliability of the validity criteria, based on ratings from three independent assessors of 50 sample cases. Except for criteria C3 and C7, Krippendorff's alpha (Krippendorff, 2004) indicated acceptable levels of agreement across all criteria. Given the conservative bias of Krippendorff's alpha (Lombard et al., 2002), such outcomes are not unexpected. Reason for this could lie within the nature of these criteria itself. Criterion 3, involving the interpretation of third-party accounts about the condition's genesis and prior discrepancies, is inherently subjective and prone to variability. Similarly, C7's assessment of laboratory results invites interpretative differences. Recognizing that context influences the interpretation of reliability measures (Bajpai et al., 2015; Hallgren, 2012; O'Connor & Joffe, 2020), a value like 0.54 for C3 can be considered sufficiently acceptable, given the subjective nature of the task. Viewed collectively, the mean inter-rater reliability ($M = 0.68$, $SD = 0.17$) indicated a substantial level of agreement for non-psychometric CVA rating.

Case Plausibility

Case plausibility was operationalized as a binary variable based on the non-psychometric validity criteria C1-C5 and

Table 2 Inter-Rater reliability (Krippendorff's α) for CVA criteria

Criterion	Krippendorf's α
C1 Report	0.81
C2 Observation	0.88
C3 Third party	0.54
C4 Activities	0.79
C5 Therapy	0.62
C7 Medication	0.45

Note. This table shows inter-rater reliability using Krippendorf's alpha for 50 cases rated by three independent raters using the rating instructions found in the supplementary material

C7 to accommodate regression analysis. The two categories refer to 'plausible' and 'implausible' response behavior with a cut-off of ≥4 conspicuous CVA criteria. Implausible response behavior incorporates symptom over- and underreporting as CVA was designed to assess both response styles. A case rated as plausible is a case with valid response behavior in terms of CVA. The classification of cases into these two categories was based on the frequency distribution of conspicuous non-psychometric validity criteria (Fig. 2).

The distribution of the number of conspicuous non-psychometric validity criteria displayed a bimodal pattern, with distinct peaks at 0 and 4. Curve fitting was done with Matlab (The MathWorks Inc., 2024). A dual-distribution modeling approach was adopted to capture this pattern: a half-Gaussian distribution was used to model the peak at 0 and a standard Gaussian distribution was centered around 4. The half-Gaussian was selected because it is restricted to non-negative values, making it suitable for count data such as CVA criteria (Gaussian mean: 3,90, Gaussian σ: 1.23, Half-Gaussian σ: 1.20) (Fig. 2). Goodness-of-Fit for these distributions reached an approximate value of 0.95, suggesting that the model captures most of the data variability and, thereby, supports the model's validity.

It is important to emphasize that the case population cannot be strictly divided into 'plausible' and 'implausible' response behavior. The model is artificial and serves only as a conceptual framework to derive a binarization cut-off. It is expected that each individual case will contain both plausible and implausible elements of response behavior, and the two statistical distributions could be interpreted as inherent components within a single case. Their ratio determines the number of conspicuous non-psychometric validity criteria. This conceptual model may thus be suitable for explaining the bimodal distribution of non-psychometric validity criteria.

Using the proportions derived from the model to categorize the MEWA data into binary groups proved highly effective. The threshold for binary classification was guided by assessments of accuracy, precision, sensitivity, and especially the F1-score, which balances precision and recall. Given the bimodal distribution, a cut-off of ≥4 emerged as

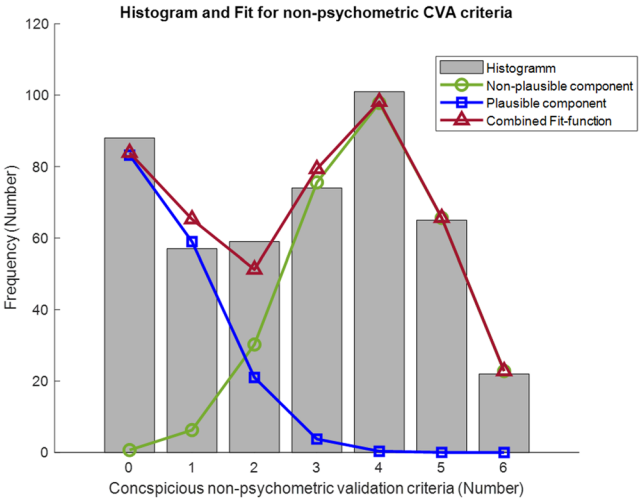

Fig. 2 Frequency histogram of conspicuous non-psychometric validity criteria with modeled components representing plausible and implausible contributions in the original dataset. Note. This figure shows a histogram of the absolute frequencies of conspicuous CVA criteria (without criterion C6) across all MEWA. The histogram exhibits a bimodal distribution. Using a conceptual model, the data was fitted with a Gaussian and a half-Gaussian distribution. These represent hypothesized components for plausible and implausible contributions, as the true underlying functions are unknown

the optimal choice, as the contribution of plausible response behavior elements was <5% in this model. This represents an approach focusing on specificity (Bianchini et al., 2005; Young et al., 2025a), i.e. maximizing the amount of true negative results, due to the immense personal disadvantages of one's response behavior being rated as implausible. Overall, this structured approach ensures that the classification results are robust from a statistical perspective and meet practical demands, making them highly relevant in real-world applications.

In contrast, for the original data — in which certain criteria remained undiscussed in a relevant number of cases — a cut-off was established at two-thirds (rounded to the nearest integer) of the criteria that were actually evaluated. By employing a CVA-based threshold, rather than depending on the medical expert witness's global validity statement, the study reduced reliance on an experience-driven, thus non-systematic and subjective, assessment.

Statistical Analysis

Data preprocessing and analysis were performed using RStudio Version 4.4.0 (RStudio Team, 2015). All Figures have been created using RStudio or Matlab. To address missing data of CVA criteria and psychometric tests, Multiple Imputation by Chained Equations (MICE) (Buuren & Groothuis-Oudshoorn, 2011) was employed and analyzed following Rubin's rules (White et al., 2011).

Following these recommendations, the minimum required number of imputed datasets corresponds to the percentage of incomplete cases. Given that 71.46% of the cases (333 out of 466) had missing values regarding CVA criteria and SIMS as well as ASTM scores, 72 datasets were generated in 20 iterations. To preserve the highest level of informational value, all variables intended for the final statistical model were first calculated using complete cases. The imputation procedure converged effectively and the imputed values proved consistent upon inspection. Box-Tidwell transformations of the continuous variables were carried out once the imputation was finalized.

A logistic regression analysis was employed to align with the data structure and to establish convergent validity of the non-psychometric CVA classification system, using the SIMS and ASTM as external, empirically validated benchmarks.

Results

Descriptives

Table 3 presents the demographic and psychometric test variables of the selected MEWA cohort of the original data, including descriptive data on claimants' sex. Table 4 provides descriptive statistics on the non-psychometric validity criteria and highlights the considerable number of missing values found in the real-world dataset. After imputation and inclusion of the non-psychometric validity criteria 251.51 out of 466 cases (53.97%) were rated as implausible or, in other terms, were rated as cases which show feigning. The decimal number is a product of the imputation and its subsequent uncertainty.

As shown in Table 5, the percentage of implausible response behavior increased with higher SIMS scores and the mean amount of conspicuous non-psychometric validity criteria rose from 2.75 to 4.60. By contrast, in the ASTM data, the proportion of implausible rated response behavior increased with lower ASTM scores, and the mean amount of conspicuous non-psychometric validity criteria increased from 3.32 to 5.59. For more details on the original data, see Supplementary Material B. Table 5 therefore indicated an increasing trend in implausible response behavior associated with higher SIMS scores. Furthermore, the table also indicated that implausible response behaviour was increasing with lower ASTM scores. The following logistic regression examines this potential correlation more closely.

Tables 7 and 8, which provide results regarding the SIMS/ASTM score categories and the individual non-psychometric validity criteria, are of marginal importance to this investigation and can be found in the Supplementary Material C.

Logistic Regression

The linearity assumption was tested using the Box and Tidwell method (Box & Tidwell, 1962). In both the original and the imputed data none of the interaction terms reached statistical significance, implying that a linear relationship can be assumed (Table 11 in Supplementary Material B). None of the Variance Inflation Factors (VIF) values exceeded a value of 2 (Table 12, Supplementary Material B) deeming multicollinearity not an issue (Kutner et al., 2005). Furthermore, studentized residuals, leverage values, and Cook's distance were calculated to identify potential outliers. None of the selected imputed datasets presented any values above the commonly used thresholds, indicating the

Table 3 Sex-Based Demographic, psychometric and CVA profiles

Gender	Age		SIMS		ASTM		Implausible Cases	Plausible Cases
	Min - Max	M (SD)	Min - Max	M (SD)	Min - Max	M (SD)	n (RF)	n (RF)
Female	24–64	54.18 (7.34)	2–52	21.90 (8.99)	40–90	78.47 (10.32)	142 (0.56)	111 (0.44)
Male	30–64	54.15 (7.20)	1–64	21.78 (10.34)	38–90	79.36 (9.96)	121 (0.57)	92 (0.43)
Total	24–64	54.16 (7.27)	1–64	21.85 (9.62)	38–90	78.87 (10.15)	263 (0.56)	203 (0.44)

M Mean, *SD* Standard Deviation, *RF* Relative Frequency. This table depicts a sex-based demographic as well as psychometric and CVA profiles of the original data without imputation. Due to the amount of missing data regarding CVA criteria, a case was rated as implausible if two-thirds of criteria assessed were rated as conspicuous

Table 4 Descriptive statistics regarding Non-Psychometric validity criteria

	C1 Report	C2 Observation	C3 Third party	C4 Activities	C5 Therapy	C7 Medication
recorded	354	454	409	313	416	325
conspicuous (n)	197	272	166	248	275	100
inconspicuous (n)	157	182	243	65	141	225
NA (n)	112	12	57	153	50	141

Table 5 Distribution of conspicuous Non-Psychometric validity criteria and case plausibility across ranges of ASTM and SIMS score categories

Score Category	Total amount (*n*)	Case implausibility (*p*)	Mean conspicuous validity criteria/mean number of evaluated validity criteria
SIMS (Imputed Data)			
≤15	128.95	0.30	2.75/6.00
16–20	98.64	0.52	3.51/6.00
21–25	92.71	0.69	4.17/6.00
26–30	69.14	0.63	4.11/6.00
>30	76.57	0.73	4.60/6.00
ASTM (Imputed Data)			
≤50	8.27	0.86	5.59/6.00
51–60	24.45	0.81	4.79/6.00
61–70	55.83	0.72	4.46/6.00
71–80	126.73	0.58	3.77/6.00
>80	250.72	0.45	3.32/6.00

Case implausibility refers to the relative frequency of cases with ≥4 conspicuous non-psychometric validity criteria. This table refers to the imputed data. A complete- and all-case-analysis can be found in Supplementary Material B. The total amount of cases in each category includes decimal numbers due to the changing total amount in each of the datasets as a result of the uncertainty of the imputation

absence of outliers and removing the need for further sensitivity analyses (Heiberger & Holland, 2015; Huber, 1981; Pardoe, 2012; Yan & Su, 2009). Detailed results concerning these assumptions can be found in the supplementary material.

A Generalized Linear Model (GLM) was employed to evaluate the relationship between the SIMS/ASTM and the binary CVA rating in terms of convergent validity. In the supplementary material, a logistic regression analysis based on the original data, with an adaptive two-thirds cut-off for case plausibility, is also provided.

The likelihood ratio test, using the D3 statistic to compare the imputed model with an intercept-only model, revealed a significant improvement of the imputed model ($D3(2, 853.739) = 17.976$, $p < .001$, $riv = 0.666$). The Hosmer-Lemeshow test showed a good fit ($F(8, 2140) = 1.126$, $p = .342$). This F-statistic originates from aggregating separate chi-square tests. The mean Nagelkerke R^2 across all analyses stood at 0.164, reflecting the amount of explained variance.

As presented in Table 6, the pooled results from the imputed datasets revealed a highly significant correlation between SIMS and binary case plausibility as well as a significant correlation between ASTM and binary case plausibility. Higher SIMS scores and lower ASTM scores are therefore significantly associated with cases rated as implausible according to CVA. Figure 3 provides a graphical overview of these relationships. These results confirm the hypothesis regarding a strong correlation between SVT and CVA, confirming convergent validity between the two measures. They do not, however, support the hypothesis that CVA and PVT are not correlated.

Sensitivity and Specificity

Utilizing binary CVA-classified outcomes and a dichotomization of SIMS results using the validated 17-point threshold (Cima et al., 2003), the non-psychometric CVA's sensitivity and specificity was computed. This calculation indicates a sensitivity of 0.64 and a specificity of 0.68.

Discussion

In this study, we present the commonly applied CVA in MEWA and its relationship with SVT and PVT in terms of convergent validity. Our findings indicate that, given the highly significant correlation between non-psychometric CVA and SIMS, CVA—like the SIMS and ASTM—functions as a measure of feigning and can be used independently of psychometric test scores. The hypothesis that non-psychometric CVA and PVT are not correlated was not supported by the mildly significant association between non-psychometric CVA and the ASTM. Moreover, it is not surprising that a modest correlation emerged, since both act as validity measures of response behavior, albeit different aspects of it, and are not entirely distinct. These results, however, can be a potential benefit of CVA in which it can reliably be applied to assess not only symptom but also performance validity. Given the large sample size and the modest correlation between CVA and the ASTM, this interpretation has to be handled with care until future studies and more data are available.

The observed divergence between non-psychometric CVA and the SIMS, visible by the relatively low sensitivity and specificity, could be the result of cases that escape psychometric testing. Such cases might involve the simultaneous occurrence of underreporting of some symptoms (e.g. of a mental disorder) and overreporting of other symptoms, leading to SVT outcomes comparable to valid response behavior (Boskovic et al., 2024) but abnormal

Table 6 Logistic regression results for imputed data: significant associations observed between case plausibility and SIMS and ASTM scores

Predictor	Estimate	SE	OR	95% CI for OR	*P*-Value
Constant	1.620	1.294	5.052	[0.392; 65.041]	0.212
SIMS	0.065	0.015	1.067	[1.037; 1.098]	0.000***
ASTM	−0.036	0.015	0.965	[0.936; 0.994]	0.019*

OR Odds Ratio, *SE* Standard Error, *CI* Confidence Interval. Model pooled across 72 imputed datasets. Results regarding the original data can be found in the supplementary material

*$p < .05$. **$p < .01$. ***$p < .001$

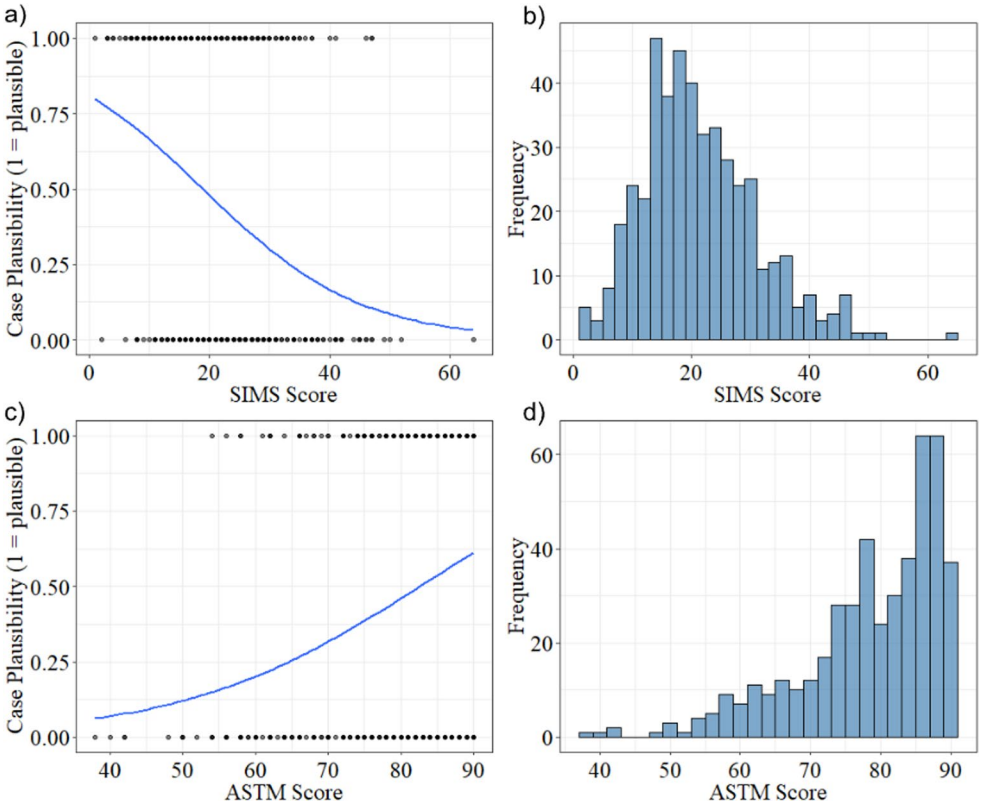

Fig. 3 Depiction of the distribution of the SIMS and ASTM scores as well as the predicted probabilities regarding case plausibility. Note. Part (b) and (d) of this figure visualize the distribution of SIMS and ASTM scores in the original dataset. Part (a) and (c) show the results of the logistic regression regarding the imputed data in the form of predicted probabilities of case plausibility taking the value of 1, which, in this case, is analogous to a plausible response behavior and validity according to non-psychometric CVA

non-psychometric CVA findings. CVA, thus, may represent an effective detection method within MEWA. In addition, it is possible for a case to be rated as implausible (conspicuous CVA) despite yielding no irregularities in SVT results, possibly suggesting familiarity with the test procedures in advance (Allen & Green, 2001; Youngjohn, 1995).

As a third-party rating, the non-psychometric CVA is intrinsically more robust against these forms of distortion. At the same time, high SVT scores can coexist with a plausible rating according to CVA, given that factors, such as certain personality profiles characterized by emotional and thought dysfunction, somatic complaints and suicidal ideation, can lead to somatic and cognitive complaint overreporting (Lang et al., 2024). Patients with genuine symptoms regarding personality disorders also achieved higher scores on most MMPI validity scales (Aparcero et al., 2023). Additionally, somatoform disorders may play a role by increasing SIMS scores due to the endorsement of symptoms with no identifiable physical cause (Wisdom et al., 2010). These potential distortions represent one reason why CVA, including both psychometric and non-psychometric criteria, is recommended in guidelines for the evaluation of psychiatric and psychosomatic disorders. Further, this professional/legal context (pension insurance claims under German social law) imposes incentives that intensify symptom overreporting in claimants. A German study investigating the response behavior in GPI claimants found that claimants who fear that their burdens and complaints will not be sufficiently recognized tend to answer the questions not in an objective manner (Kobelt-Pönicke et al., 2020). Symptom overreporting in this population can also be an expression or consequence of a work or social situation that is experienced as hopeless (Kobelt-Pönicke & Walter, 2020). The data presented here enables a more nuanced investigation of non-psychometric CVA, SVT and PVT, especially in cases involving GPI cases regarding the capacity to work.

In our study, the threshold of ≥4 conspicuous criteria for classifying a case as implausible using CVA was intentionally set quite high. This decision was informed by descriptive data, a theoretical model indicating two distinct populations and the objective of ensuring a high degree of certainty regarding implausible response behavior. Therefore, these findings mainly apply to highly conspicuous cases that carry a substantial probability of being implausible. While this study mainly focused on cases regarding the German statutory pension insurance, future studies need to empirically investigate these plausibility cut-offs in other legal contexts and each new population.

It is essential to emphasize that the data does not support a mere binary classification of response behavior based on SVT or PVT. As demonstrated by the sensitivity and specificity analysis, there are numerous cases with an inconspicuous SVT result that are deemed implausible according to non-psychometric CVA, while some cases

with elevated SVT scores are still considered plausible. In addition to the commonly used binary classifications of 'fail' and 'pass', other research has suggested that 'indeterminate' is a legitimate third outcome of SVT and PVT validity assessments (Erdodi, 2019). Accordingly, each individual case must be assessed thoroughly, with SVT and PVT serving as indicators among other components of a multimodal validity assessment (Sherman et al., 2020; Young et al., 2020). The likelihood of an implausible case rises with increasing conspicuousness in SVT and PVT as shown in the descriptive data. However, cut-off values in SVT and PVT do not alone suffice to classify a case as implausible with the required certainty in a high-stakes legal context and should be regarded as one criterion among both psychometric and non-psychometric approaches. From the perspective of current AACN guidelines, the present findings support the view that validity assessment should not be based on isolated SVT or PVT outcomes, but on the aggregation of multiple sources of validity evidence to reduce false-positive classifications. Accordingly, CVA does not replace SVT or PVT, but operationalizes their integration within a multimodal, probability-based framework congruent with recent work regarding neuropsychological assessments (Chafetz, 2011; Chafetz et al., 2015).

Despite the concerns outlined, our analyses show that SVT such as the SIMS and PVT such as the ASTM are indeed useful. When the SIMS score exceeds 30 points, there is a marked increase in the probability of feigning, even if psychometric testing is not part of the assessment (i.e. via CVA). In practical application, SIMS scores around 17 points require a thorough examination of potential feigning. As the score climbs—particularly into the 25–30 range—the risk of feigning intensifies. These results are in line with previous research (Cima et al., 2003), which documented a mean SIMS score of 30 in participants who had been trained to simulate.

In the ASTM, an analogous pattern emerged. A score below 60 was strongly correlating with feigning analogous to CVA implausibility, while a score above 70 did not necessarily imply feigning, despite the handbook suggesting the interpretation of scores below 86 points as an invalid response set (Schagen et al., 1997). It is reasonable to assume that an ASTM score hovering around the guessing probability (49–54 points) already signifies a substantial feigning of test results, as the examinee is tasked with providing correct answers. When the score indicates random responding, it constitutes a notable phenomenon. Although some contexts consider an invalid response set proven only at scores below or equal 48, exclusively interpreting such scores as evidence of feigning would be inaccurate. Instead, such scores fall even further below the guessing threshold. A score at the guessing probability itself points to non-cooperation or severe aggravation and a score well below that threshold suggests deliberate production of incorrect responses. Typically, non-cooperation is purposeful, and random answers are incongruous with the individual's actual performance during the assessment interview—consistent with the marked non-psychometric validity criteria observed in these situations.

Limitations

While centering on a specific legal issue enhances the homogeneity of the sample, it should be noted that the results' applicability to other evaluation contexts remains limited and has not yet been investigated. It is conceivable that in cases, such as accident-related inquiries, different values might emerge. Further research is needed to determine how CVA performs in detecting varied response styles (e.g. naive vs. informed) (Röhner et al., 2013). Consequently, the findings of this study cannot simply be generalized to all legal domains or every form of symptom reports. Nonetheless, they may still prove relevant to broader medical questions, including social security cases involving severe disabilities. More research focused on other legal contexts is necessary to improve CVA's transferability to additional areas. Further, the data consisted only of claimants of the German statutory pension insurance. The generalizability of the results to other cultures or populations is, therefore, limited.

Another limitation of this investigation is the use of partially incomplete datasets. Unlike prospective research designs, achieving complete assessment of all CVA criteria for each case was not possible in this retrospective study. Nevertheless, most of the six possible criteria (plus psychometric testing as a seventh criterion) were evaluated, suggesting that CVA considerations, in conjunction with psychometric measures, largely were accounted for. Importantly, among the subset of 133 fully evaluated cases, the data showed a comparable structure: non-psychometric validity criteria remained bimodally distributed, SIMS scores rose alongside non-psychometric criteria and ASTM scores declined (see supplementary material). While multiple imputation is recognized as a validated technique (Wulff & Ejlskov, 2017), we also analyzed the complete dataset and the subset of fully evaluated cases to minimize any error resulting from the inherent uncertainty of imputation.

We chose not to focus exclusively on the fully evaluated cases, because it cannot be assumed with certainty that MEWA, in which all seven criteria were assessed, are entirely comparable to other cases. There may be reasons why the expert employed such a comprehensive approach in

those instances. For example, this may have occurred when the psychometric test result stood isolated (whether inconspicuous or conspicuous). In such circumstances, applying CVA in detail is particularly important, which could mean that fully evaluated cases contain more instances of this type. Since MEWA do not disclose the underlying reasons for complete assessments, this cannot be conclusively determined.

Conclusion

In this study, we provided an empirical basis for the application of CVA in forensic contexts and clarified how its findings can be interpreted in practice. Our results demonstrated that CVA showed convergent validity with both SVT and PVT, supporting its potential as an independent and complementary tool in legal evaluations. In forensic casework, CVA can, therefore, enhance the assessment of response behavior beyond traditional psychometric testing. While our analyses examined SVT, PVT and non-psychometric CVA criteria separately to establish validity evidence, in real-world forensic evaluations, integrating all seven validity criteria is likely to produce the most dependable results. Ultimately, these findings promote a more refined interpretation of both psychometric test data and non-psychometric CVA indicators, thereby strengthening the evidentiary basis for validity determinations in MEWA.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12207-026-09557-y.

Acknowledgements This research was financially supported by the DGNB-Research Promotion Award 2020 awarded by the German Society for Neuroscientific Assessment (DGNB). The authors would like to express their gratitude for this support.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Competing interests The authors have no financial or non-financial competing interests to disclose.

References

Allen, L. M., & Green, P. (2001). Declining CARB failure rates over 6 years of testing: What's wrong with this picture? *Archives of Clinical Neuropsychology, 16*(8), 846–856. https://doi.org/10.1093/arclin/16.8.846

Aparcero, M., Picard, E. H., Nijdam-Jones, A., & Rosenfeld, B. (2023). Comparing the ability of MMPI-2 and MMPI-2-RF validity scales to detect feigning: A meta-analysis. *Assessment, 30*(3), 744–760. https://doi.org/10.1177/10731911211067535

AWMF (2019). *S2k-Leitlinie zur Begutachtung psychischer und psychosomatischer Störungen*. https://register.awmf.org/de/leitlinien/detail/051-029

AWMF (2023). *S2k-Leitlinie Ärztliche Begutachtung von Menschen mit chronischen Schmerzen*. https://register.awmf.org/de/leitlinien/detail/187-006

Bajpai, S., Bajpai, R., & Chaturvedi, H. (2015). Evaluation of inter-rater agreement and inter-rater reliability for observational data: An overview of concepts and methods. *Journal of the Indian Academy of Applied Psychology, 41*, 20–27.

Bianchini, K. J., Greve, K. W., & Glynn, G. (2005). On the diagnosis of malingered pain-related disability: Lessons from cognitive malingering research. *The Spine Journal, 5*(4), 404–417. https://doi.org/10.1016/j.spinee.2004.11.016

Boe, E. W., & Evald, L. (2022). Symptom and performance validity in neuropsychological assessments of outpatients 15–30 years of age. *Brain Injury*. https://doi.org/10.1080/02699052.2022.2158222

Boskovic, I., Giromini, L., Katsouri, A., Tsvetanova, E., Fonse, J., & Merckelbach, H. (2024). The spectrum of response bias in trauma reports: Overreporting, underreporting, and mixed presentation. *Psychological Injury and Law, 17*(2), 117–128. https://doi.org/10.1007/s12207-024-09503-w

Box, G. E. P., & Tidwell, P. W. (1962). Transformation of the independent variables. *Technometrics, 4*(4), 531–550. https://doi.org/10.1080/00401706.1962.10490038

Brongers, K. A., Hoekstra, T., Roelofs, P. D. D. M., & Brouwer, S. (2022). Prevalence, types, and combinations of multiple problems among recipients of work disability benefits. *Disability and Rehabilitation, 44*(16), 4303–4310. https://doi.org/10.1080/09638288.2021.1900931

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

Buuren, S., & Groothuis-Oudshoorn, C. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*. https://doi.org/10.18637/jss.v045.i03

Chafetz, M. (2011). Reducing the probability of false positives in malingering detection of social security disability claimants. *The Clinical Neuropsychologist*, *25*(7), 1239–1252. https://doi.org/10.1080/13854046.2011.586785

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

Cima, M., Hollnack, S., Kremer, K., Knauer, E., Schellbach-Matties, R., Klein, B., & Merckelbach, H. (2003). The German version of the structured inventory of malingered symptomatology: SIMS. *Der Nervenarzt*, *74*, 977–986.

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

Grundsatz- und Querschnittsabteilung Finanzen und Statistik. (2023). *Rentenversicherung in Zahlen 2023*. Deutsche Rentenversicherung Bund.

Hallgren, K. A. (2012). Computing Inter-Rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34.

Heiberger, R. M., & Holland, B. (2015). *Statistical Analysis and Data Display: An Intermediate Course with Examples in R*. Springer. https://doi.org/10.1007/978-1-4939-2122-5

Huber, P. J. (1981). Regression. *Robust statistics* (pp. 153–198). John Wiley & Sons, Ltd. https://doi.org/10.1002/0471725250.ch7

Kobelt-Pönicke, A., & Walter, F. (2020). Beschwerdenvalidierung in der sozialmedizinischen begutachtung. *Zeitschrift für Psychiatrie Psychologie Und Psychotherapie*. https://doi.org/10.1024/1661-4747/a000405. https://econtent.hogrefe.com/doi/

Kobelt-Pönicke, A., Walter, F., & Riemann, M. (2020). Führt Das bewusstsein Moralischer grundwerte Zu einem authentischeren Antwortverhalten in beschwerdenvalidierungstests? *Zeitschrift Für Psychiatrie Psychologie Und Psychotherapie*, *68*(2), 106–112. https://doi.org/10.1024/1661-4747/a000409

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology*. SAGE.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (Fifth Edition). McGraw-Hill. https://thuvienso.hoasen.edu.vn/handle/123456789/9564

Lang, P. A., Thomas, L., & Lidbury, B. A. (2024). Psychopathology and the validity of Gastrointestinal symptom reporting as revealed through cluster analyses of MMPI-2-RF results. *Digestive Diseases and Sciences*, *69*(11), 4063–4071. https://doi.org/10.1007/s10620-024-08629-w

Lombard, M., Snyder-Duch, J., & Bracken, C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, *28*, 587–604. https://doi.org/10.1111/j.1468-2958.2002.tb00826.x

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods*. https://doi.org/10.1177/1609406919899220

Pardoe, I. (2012). Applied Regression Modeling: A Business Approach. *Applied Regression Modeling: A Business Approach*. https://doi.org/10.1002/9781118274415.ch6

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

Röhner, J., Schröder-Abé, M., & Schütz, A. (2013). What do fakers actually do to fake the IAT? An investigation of faking strategies under different faking conditions. *Journal of Research in Personality*, *47*, 330–338. https://doi.org/10.1016/j.jrp.2013.02.009

RStudio Team (2015). *RStudio: Integrated Development Environment for R* [Computer software]. http://www.rstudio.com/

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*(1), 43–51. https://doi.org/10.1080/01688639708403835

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

The MathWorks Inc (2024). *Optimization Toolbox (R2024b)* [Computer software]. https://www.mathworks.com

van Impelen, A., Merckelbach, H., Jelicic, M., & Merten, T. (2014). The Structured Inventory of Malingered Symptomatology (SIMS): A systematic review and meta-analysis. *The Clinical Neuropsychologist, 28*(8), 1336–1365. https://doi.org/10.1080/13854046.2014.984763

von Kardorff, E., Klaus, S., & Meschnig, A. (2020). *Wege Psychisch Kranker in die EM-Rente und Rückkehrperspektiven Aus der EM-Rente in arbeit: Ansatzpunkte Zu frühzeitiger intervention in biografische und krankheitsbezogene Verlaufskurven (WEMRE)*. Humboldt-University.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*(4), 377–399. https://doi.org/10.1002/sim.4067

Wisdom, N. M., Callahan, J. L., & Shaw, T. G. (2010). Diagnostic utility of the Structured Inventory of Malingered Symptomatology to detect malingering in a forensic sample. *Archives of Clinical Neuropsychology, 25*(2), 118–125. https://doi.org/10.1093/arclin/acp110

Wulff, J., & Ejlskov, L. (2017). Multiple imputation by chained equations in praxis: Guidelines and review. *Electronic Journal of Business Research Methods*, *15*, 2017–2058.

Yan, X., & Su, X. (2009). *Linear Regression Analysis: Theory and Computing*.

Young, G., Erdodi, L., Giromini, L., & Rogers, R. (2025a). Malingering-related assessments in psychological injury: Performance validity tests (PVTs), symptom validity tests (SVTs), and invalid response set. *Psychological Injury and Law, 18*(1), 19–34. https://doi.org/10.1007/s12207-024-09523-6

Young, G., Foote, W. E., Kerig, P. K., Mailis, A., Brovko, J., Kohutis, E. A., McCall, S., Hapidou, E. G., Fokas, K. F., & Goodman-Delahunty, J. (2020). Introducing psychological injury and law. *Psychological Injury and Law*, *13*(4), 452–463. https://doi.org/10.1007/s12207-020-09396-5

Young, G., Giromini, L., Erdodi, L., & Rogers, R. (2025b). Invalid response set and malingering-related assessments in psychological injury: Definitions and a hierarchy of terms. *Psychological Injury and Law, 18*(1), 3–18. https://doi.org/10.1007/s12207-025-09529-8

Youngjohn, J. (1995). Confirmed attorney coaching prior to neuropsychological evaluation. *Assessment*, *2*, 279–283. https://doi.org/10.1177/1073191195002003007