Psychological Injury and Law (2026) 19:19
https://doi.org/10.1007/s12207-026-09567-w

Criteria-Based Validity Assessment in Legal Cases Involving Pension and Accident Insurance

N. Klett[1,2,4] · R. Dohrenbusch[3] · E. M. Siegmann[2] · A. Schütz[4] · J. Kornhuber[2] · J. Röhner[4] · F. Keller[5] · E. S. Capito[1] · T. Grömer[1,2]

Received: 3 March 2026 / Accepted: 11 May 2026

Abstract
Background: Medical expert witness assessments (MEWAs) evaluate case validity on the basis of both psychometric and non-psychometric modalities. Although symptom validity tests (SVTs) and performance validity tests (PVTs) are widely used, many thresholds were developed in analogue and known-groups validation contexts, and their transferability to MEWAs remains uncertain. Prior research validated a Criteria-Based Validity Assessment (CVA) as a multimodal framework for assessing case plausibility, but it remained unclear whether CVA and psychometric thresholds perform similarly across different legal contexts, such as pension insurance versus accident insurance evaluations. Objective: This study aimed to investigate differences in CVA, SVT, and PVT results across pension and accident insurance cases. Methods: A total of 721 MEWAs (572 pension; 149 accident) were analyzed. CVA criteria were rated by trained raters, and response biases were assessed psychometrically using the SIMS and the ASTM. Two-component beta-binomial mixture models were applied to derive CVA plausibility thresholds. Results: Mixture modeling reliably identified a distinct bimodal distribution of conspicuous CVA criteria counts, interpreted as plausible and implausible subgroups in both pension and accident cases, with ≥4 conspicuous CVA criteria representing the optimal plausibility threshold. Pension claimants showed significantly higher SIMS scores and lower ASTM scores than accident claimants, independent of case plausibility. Conclusion: CVA can be used to assess validity information collected from multiple data sources (longitudinal and cross-sectional) and to assess the validity of health-related claims across different legal settings. CVA, however, was not a tool for measuring context-specific response biases. Cases with ≤2 conspicuous CVA criteria may be valid, cases with three conspicuous criteria require individual examination because feigning is possible but not certain, and ≥4 conspicuous criteria supported a conservative classification of CVA-defined case implausibility.

Keywords Criteria-Based Validity Assessment · Medical expert witness assessment · Pension insurance · Accident insurance · Symptom validity test · Performance validity test

Introduction

Medical expert witness assessments (MEWAs) are primarily conducted by physicians and are standard practice in German litigation. They are independent medical evaluations, with the majority concerning pension or accident insurance cases. External incentives in this context usually include financial compensation due to a reduced capacity to work regarding pension insurance cases concerning the number of hours per day a person could still work and under which functional restrictions. A reduction in earning capacity represents the incentive in accident insurance cases primarily for the purpose of determining entitlement to an accident-related pension. Only causally accident-related symptoms and functional impairments are considered in this context. The most common disorders are mental disorders, followed by oncological diseases as well as musculoskeletal and connective tissue disorders in pension insurance cases (Deutsche Rentenversicherung Bund, 2024), while accident insurance cases mostly consist of injuries caused by

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

Extended author information available on the last page of the article

Published online: 06 June 2026

physical influences (e.g. injuries of the extremities, head, and spine) and shock resulting from emotional trauma or psychological factors (DGUV, 2025). These work-related accidents can lead to pain and psychological disorders such as post-traumatic stress disorder (PTSD) or adjustment disorders.

Feigning in forensic assessments is a widely discussed topic and studies have demonstrated its high prevalence in these settings. Sherman et al. (2020) report base rates of feigning of up to 40% in personal injury and disability and up to 60% in social security cases. An investigation of German workers' compensation cases revealed invalid response behavior in approximately 45–48% of cases (Merten et al., 2010; Stevens et al., 2008) while data specifically regarding German pension insurance cases are lacking. Martin and Schroeder (2020) report only small differences in base rates regarding clinical patients having workers' compensation involved (33.5%) or patients considering disability (25%). These data are based solely on the professional experience of the medical expert, however.

In MEWAs, response behavior can be assessed using a multimodal criteria set and psychometric validity tests (VTs). This Criteria-Based Validity Assessment (CVA) (AWMF, 2019) covers discrepancies in symptom reporting (C1), observable behavior and clinical findings (C2), documented medical history including the individual course and development of disorders (C3), daily functioning (C4), treatment use (C5), psychometric test results (C6), and, if applicable, laboratory verification of medication intake (C7). This type of assessment is in agreement with the Official Position of the American Academy of Clinical Neuropsychology (Chafetz et al., 2015) regarding the integration of multiple, independent validity indicators as the core principle of neuropsychological assessments (Chafetz, 2011). It is important to note that CVA is composed of qualitative (C1-C5, C7) and quantitative (C6) criteria.

Within CVA, VTs constitute an objective complement to the expert's clinical judgment, and their thresholds accordingly affect the evaluation of cases. However, existing symptom validity tests (SVTs) and performance validity tests (PVTs) thresholds, as measurements of self-report validity and cognitive performance validity, respectively, originate from studies in which participants are trained to feign (Young et al. 2025a, b) symptoms, which often include cross-validation with a forensic sample. For example, the commonly used Structured Inventory of Malingered Symptomatology (SIMS) validation (Cima et al., 2003) combines an analogue design (instructed student simulators vs. honest controls) with a forensic psychiatric sample instructed to respond honestly, thereby calibrating thresholds against both feigned responding and genuine–yet predominantly severe–psychopathology. The original Amsterdam Short-Term Memory Test (ASTM) (Schagen et al., 1997) validation was based on a small, highly controlled sample comprising patients with closed head injury (CHI), healthy controls, and an analogue feigning group. Notably, the feigned deficit group consisted of relatives of CHI patients, who were explicitly instructed to simulate cognitive impairment based on their knowledge of the affected individuals. However, similar to the SIMS, the resulting thresholds are calibrated within a specific experimental and clinical context, namely analogue feigning based on instructed simulation and a relatively homogeneous neurological patient group. The present study extends this work by examining whether these established thresholds are appropriately calibrated for a fundamentally different population, namely MEWA claimants. In this context, the focus is not on re-establishing validity per se, but on evaluating the performance of existing thresholds in a diagnostically heterogeneous, real-world medicolegal setting characterized by differing incentive structures.

Additionally, SVTs, for example the SIMS, can contain items referring to genuine instead of feigned symptoms, and as the disease burden increases, the diagnostic accuracy of the SIMS decreases (van Impelen et al., 2014; Wertz et al., 2021). Thus, the generalizability of the results of the original validation to MEWAs remains a topic of debate (Bush et al., 2005; Walczyk et al., 2018), particularly in a German MEWA population, in which affective disorders are highly prevalent (von Kardorff et al., 2020). While SVTs and PVTs are valuable tools in MEWAs, their specific thresholds for this context have remained largely undefined. Nevertheless, there are findings from personal injury and disability-claim settings (Wisdom et al., 2010) as well as recent evidence (Klett et al., 2026) suggesting that a threshold around 24 points on the SIMS may be more appropriate in the context of real-world settings as opposed to the validation settings. However, these studies did not distinguish between different case types, such as disability pension claims and accident cases involving causality assessments.

Beyond this contextual limitation, methodological work has highlighted that single PVTs vary across sensory modalities, cognitive domains, and detection mechanisms, prompting recommendations to combine multiple tests and use of multivariate models (Rai et al., 2023). Several authors emphasize that SVT and PVT results should not stand alone but be integrated with other validity indicators (Bush et al., 2005), highlighting the need for a multimodal approach to case validity. Furthermore, the practice of evaluating SVT and PVT results alongside other validity indicators is regarded as standard practice and can be found in more widely known malingering frameworks (e.g., the MND criteria by Sherman et al. (2020) and places the CVA system alongside these already empirically validated systems. Importantly, empirical work demonstrated that

feigning leads to multiple, partly dissociable response patterns, underscoring the rationale for multi-criterion validity assessment frameworks (Röhner et al., 2022).

Previous studies primarily validated psychometric tests in relation to specific pathological disorders. Yet MEWAs do not involve homogeneous populations, but evaluate distinct claimant groups depending on the underlying legal question. For instance, populations assessed after work-related accidents tend to be younger than those in pension insurance evaluations. Response patterns also differ: disability claimants frequently report a wide range of symptoms (Brongers et al., 2022), whereas such diffuse presentations are less useful in accident cases since they could, in theory, undermine causal attribution. Additionally, pain, fibromyalgia, and chronic fatigue–disorders usually present in pension insurance cases–are described to present the highest base rates of probable feigning in MEWAs, apart from mild head injury (Mittenberg et al., 2002). Moreover, while several studies demonstrated high diagnostic accuracy of SVT and PVT measures in German-speaking early retirement claimants, these studies also warned not to extrapolate these results to other forensic contexts due to observed higher SVT/PVT failure rates or differences in severity of cognitive dysfunction between forensic contexts (Fuermaier et al., 2023, 2025; Teßmann et al., 2025). This variability presupposes the need for context-specific interpretation of validity criteria and VT results, and highlights the limitations of applying uniform psychometric thresholds across forensic settings. Furthermore, some studies suggest that there is a tendency to overreport symptoms in the German Statutory Pension Insurance (GPI) system, driven by concerns about the seriousness with which complaints are considered or as an expression of hopelessness in the workplace (Kobelt-Pönicke et al., 2020; Kobelt-Pönicke & Walter, 2020), indicating another source of conspicuous SVT/PVT results.

Data from a large sample of MEWAs within the GPI system revealed a bimodal distribution of cases based on non-psychometric CVA, which was interpreted as representing plausible and implausible components (Klett et al., 2026). Furthermore, CVA showed convergent validity with an SVT and PVT, reinforcing its role as a multimodal tool and providing evidence that VTs yield meaningful results in real cases. However, it remained unclear whether psychometric thresholds and validity parameters generalize to other MEWA contexts, such as accident insurance claims, limiting CVA's generalizability. Although that study examined non-psychometric CVA together with one SVT and PVT, it did not address whether their relationship differs between case types, raising questions about the consistency and fairness of validity determinations across legal contexts. Although this study reported a sensitivity and specificity analysis of CVA, this calculation was based on the results of a single SVT as a point of reference, limiting the informativeness of these values regarding the aim of multimodality in MEWA contexts. Inherently, in forensic assessments the truth about invalid or valid response behavior is known at no point, not even retrospectively after the litigation is resolved. This differs from experimental studies, in which participants are trained to simulate and the true extent of invalid response behavior is known to the researchers. This means that in MEWAs a reference standard constitutes only an approximation. It is a circumstance which cannot be clarified. Consequently, classical diagnostic accuracy indices (e.g., sensitivity, specificity, or predictive values) should be interpreted with caution, as they reflect agreement with an imperfect proxy rather than true classification accuracy.

The qualitative nature of CVA also offers the possibility for inconsistencies when evaluating conditions which are prone to intrinsic symptom fluctuations or day-to-day variability, such as chronic pain, fatigue, or fibromyalgia. In theory, CVA counters this possibility by not relying solely on the apparent physiological observations (criterion C2) but on multiple independent criteria. This is a beneficial factor of CVA in terms of incremental validity since it acts as an evaluation guideline combining quantitative (SVT/PVT) and multiple qualitative criteria, but this is also a factor which has not been empirically validated and presents another limitation of CVA.

In this study, we aimed to investigate two primary objectives. First, we examined whether a bimodal plausibility model, previously developed in the context of pension insurance cases, generalizes to MEWAs conducted in accident insurance settings. Second, we investigated the integration of criterion C6 (psychometric test results) into the CVA framework, addressing a limitation of prior work in which C6 had to be excluded to avoid circularity. Based on theoretical considerations regarding differing incentive structures and evaluative goals across legal contexts, we formulated the following hypotheses:

1. Context differences in response validity: We hypothesized that SVT and PVT results differ between pension and accident insurance cases. Specifically, pension insurance cases–where global invalidity may represent a primary incentive–were expected to show higher rates of invalid responding compared to accident insurance cases, in which establishing a causal link to a specific impairment is more central.
2. Differences in CVA base rates: We expected higher base rates of invalid response behavior, as indicated by CVA classifications, in pension insurance cases relative to accident insurance cases.

To address these aims, we analyzed real-world MEWA data to evaluate group differences and to assess whether context-specific patterns of response validity emerge. Furthermore, we explored the integration of criterion C6 into the CVA framework as a step toward a more comprehensive validity assessment approach.

Materials and Methods

Data and Subjects

The dataset used in this study is based on previous research (Klett et al., 2026) in which 466 pension insurance MEWAs were investigated. In this study, the previously eligible 933 MEWAs were supplemented with 207 additional MEWAs conducted in 2024 as well as cases related to accident insurance, resulting in 1,140 eligible MEWAs. Out of the 1,140 MEWAs screened, 725 were deemed relevant after screening, compared with the 466 previously analyzed MEWAs. All cases regarding the German statutory pension insurance and German statutory accident insurance qualified for further analysis. Pre-procedural pension insurance MEWAs were not included in this study because they do not include a full psychiatric and neurological assessment. Four additional MEWAs were excluded for reasons detailed in Fig. 1. The final dataset comprised 721 cases, including 572 related to the German Statutory Pension Insurance (GPI) and 149 concerning accident insurance cases. Although these data overlap with the previous study, 255 previously unanalyzed MEWAs, some of which were accident insurance cases, were added to this final dataset. Beyond the expanded sample size, the present study differs from our previous work in that it investigates differences in mean psychometric scores as well as base rates of invalid response behavior between pension and accident insurance cases, which was not feasible before.

Psychometric tests, including cognitive tests and SVT/PVT, were administered independently before or after a full psychiatric and neurological assessment interview by the medical expert and were evaluated by trained psychologists under supervision. The test results were typically not available to the medical expert prior to the assessment interview. The CVA ratings were not included in the assessment but were evaluated during the finalization of the expert witness report with all of the information available to the medical expert.

The age of the assessed claimants ranged from 19 to 72 years ($M = 53.31$, $SD = 8.67$). The demographic composition of the sample divided into pension and accident insurance cases is depicted in Table 1.

Psychometric Tests

The German versions of the Structured Inventory of Malingered Symptomatology (Cima et al., 2003) and the Amsterdam Short-Term Memory Test (Schagen et al., 1997) were administered in most of the MEWAs. These tests, serving as an SVT and PVT, respectively, have been empirically validated and described elsewhere (Klett et al., 2026). Missing

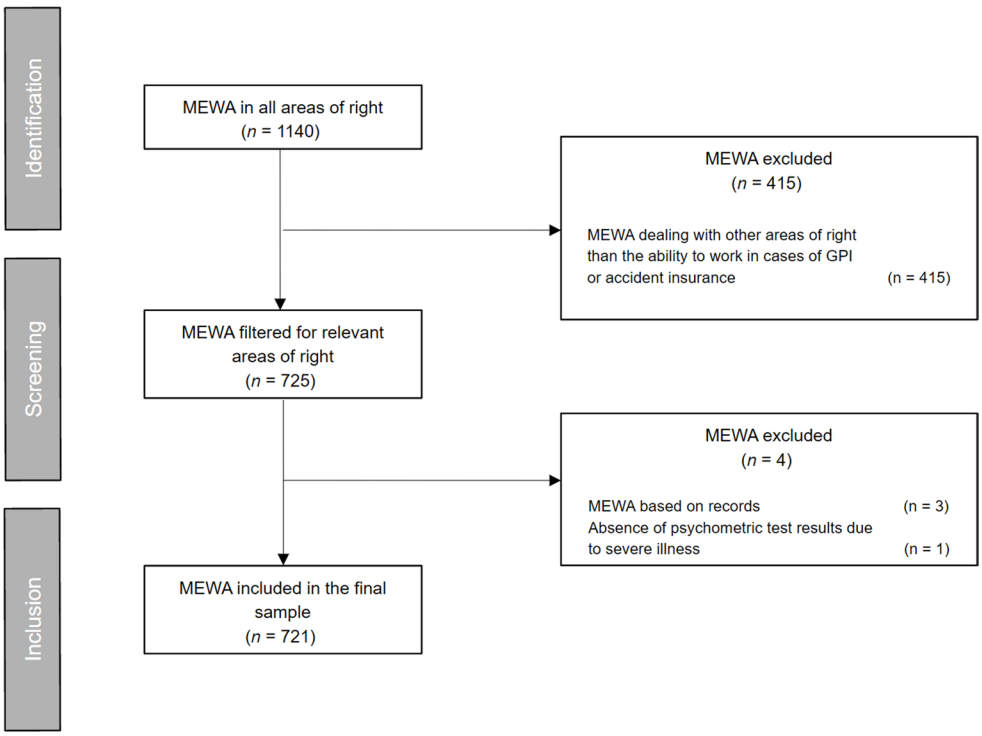

Fig. 1 Flowchart of Screened, Excluded and Included MEWAs

Table 1 Demographic composition of the sample

Group	*n* (RF)	Age (M; SD)
All data (*n* = 721)		
Female	366 (50.76%)	53.01 (8.76)
Male	355 (49.24%)	53.62 (8.54)
Pension Insurance (*n* = 572)		
Female	310 (54.20%)	54.03 (7.29)
Male	262 (45.80%)	54.27 (7.11)
Accident Insurance (*n* = 149)		
Female	56 (37.58%)	47.34 (13.13)
Male	93 (62.42%)	51.76 (11.53)

Note. *M* Mean, *SD* Standard Deviation, *n* Absolute Number, *RF* Relative Frequency. This table shows the demographic composition of the sample regarding age and gender in pension insurance and accident insurance cases

psychometric test data were imputed using Multivariate Imputation via Chained Equations (MICE; van Buuren & Groothuis-Oudshoorn, 2011).

Criteria-Based Validity Assessment (CVA) and Case Plausibility

The CVA classification system (AWMF, 2019) is an expert-consensus, multimodal approach for assessing case validity in MEWAs, and previous work provided preliminary empirical support for the CVA structure and convergent validity with psychometric validity indicators (Klett et al., 2026). As a framework for evaluating response behavior, it defines seven clearly separable criteria (C1–C7) that cover distinct information domains: the subject's report, behavioral/physiological observations, medical history and course of disease, daily activities, therapeutic engagement, psychometric findings, and–where applicable–laboratory/blood serum evidence of medication intake. A description of these criteria is provided in Table 2. CVA primarily aims to provide a general, integrated assessment of the plausibility of different data sources, while SVTs focus on the targeted control of formal and content-related response distortions in a survey situation. PVTs, on the other hand, focus on the identification of atypical performance patterns or performance levels in comparison to individuals with genuine cognitive impairment (Sweet et al., 2021). SVTs and PVTs are integrated in criterion C6. Thus, CVA represents a mixture of qualitative and quantitative criteria, as a complex and multifaceted set of assessment and evaluation rules for information processing. As a tool for measuring behavioral patterns, CVA does not directly prove feigning or malingering, but can only indicate validity issues or raise validity concerns.

To extract CVA criteria from the MEWAs, the rating instructions provided in supplementary material A were used. Each criterion was coded as 0 (inconspicuous/no discrepancies), 1 (conspicuous/discrepancies present), or 2 (not evaluated), based on predefined key phrases within the MEWAs (e.g. 'the clinical findings were bland' or 'the assessed person often answered vague and gave evasive answers to questions'). Not evaluated criteria were imputed using MICE. CVA criteria were extracted by two independent raters and inter-rater agreement for the CVA ratings had been examined and had demonstrated an acceptable level of consistency (Klett et al., 2026).

Table 2 Criteria for Criteria-Based Validity Assessment (CVA) in MEWAs

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

Case plausibility can be operationalized as a binary variable based on the number of conspicuous non-psychometric CVA criteria. Using this approach, cases can be classified as plausible or implausible, with CVA-defined implausibility indicating possible feigning of symptoms. It is important to note that this binarization of cases into plausible and implausible categories only refers to an approximation of feigning in MEWAs, since the extent of feigning cannot be known. It solely acts as an approximation based on a multimodal validity framework. Since no official thresholds exist for categorizing cases as plausible or implausible, potential thresholds need to be explored before further analysis can be done. In the following, threshold analysis was conducted using a two-component beta-binomial mixture modeling approach due to the bimodal nature of the data.

Statistical Analysis

Data preprocessing and statistical analyses were conducted in R using RStudio (version 4.4.0, RStudio Team, 2015). Given the nature of real-world data, approximately 75.87%

of cases contained at least one missing data point. These missing data were imputed using Multivariate Imputation via Chained Equations (van Buuren & Groothuis-Oudshoorn, 2011). Following established recommendations (White et al., 2011), 76 datasets were imputed over 20 iterations. Convergence and consistency were assessed throughout the imputation process and found to be satisfactory.

To explore SVT/PVT results in MEWAs, individual linear regression models were fitted to examine the relationship between SIMS/ASTM scores and case type. These models simultaneously controlled for case plausibility, age, and gender. Cohen's d was calculated as an effect size for group differences (Cohen, 1988). Given an α-level of 0.05, 0.80 power, and a sample size of 721, a sensitivity analysis revealed that effects as small as $f^2 = 0.01$ would be detectable by the linear regression models. Regarding the SIMS model, Q-Q plots revealed only minor deviations from normality. Regarding the ASTM model, Q-Q plots indicated a significant deviation from a normal distribution, likely due to the strong negative skewness in ASTM scores. A subsequent Box-Cox transformation with an optimal lambda of 2 resulted in a distribution that more closely approximated normality. A Levene's test and 'Residuals vs. Fitted' plots indicated heteroskedasticity. Thus, robust standard errors and p-values were computed. The same plot indicated linearity. Outliers were identified based on studentized residuals (Pardoe, 2012; Yan & Su, 2009). Several imputed datasets contained influential cases, necessitating a sensitivity analysis. For that purpose, and to maintain consistent group sizes for multivariate analyses, all cases identified as outliers across imputations were removed from each imputed dataset. 25 (3.47%) cases were removed from all of the datasets regarding the SIMS and 14 (1.94%) cases regarding the ASTM. Results were computed for models before and after removal of outliers. Multicollinearity (Kutner et al., 2005) was not found to be a concern in any of the models (Table S3, supplementary material B).

Similarly, the hypothesis regarding the association between case plausibility and case type was investigated using a logistic regression model, with case plausibility serving as the dependent variable. Since multiple imputation was applied, logistic regression assumptions were examined using the 15th, 20th, 25th, 30th, and 35th datasets from the imputed data. The Box-Tidwell method (Box & Tidwell, 1962) (Table S4, supplementary material B) confirmed the assumption of linearity. No problematic multicollinearity (Table S5, supplementary material B) or outliers were identified through studentized residuals, leverage values, and Cook's distance (Heiberger & Holland, 2015; Huber, 1981; Pardoe, 2012; Yan & Su, 2009).

To avoid the issue of circularity and artificial correlations when investigating the effect of case type on SVT and PVT scores while controlling for case plausibility, case plausibility was calculated without criterion C6 (the psychometric test results).

Results

CVA Threshold Approximation

The evaluation of CVA criteria is a sequential process in which the medical expert witness judges each criterion as conspicuous or not. This corresponds to a series of Bernoulli trials and, assuming homogeneous probabilities, yields a binomial distribution for the number of conspicuous criteria with pC+ denoting the probability that a given criterion is conspicuous. In real-world datasets, however, this probability is expected to vary between cases, so that a distribution of pC+ (with variance Var(pC+), expressible as a standard deviation SDpC+) is more realistic than a single fixed value. This overdispersion is naturally captured by a beta-binomial distribution, which is defined by three parameters: the mean probability pC+, its dispersion (SDpC+), and the number of assessed criteria. This framework allows a unified interpretation of the discrete counts of conspicuous criteria when six or seven CVA criteria are evaluated.

In the data, the distribution of conspicuous CVA criteria was visibly bimodal (Fig. 2), suggesting that the sample can be described as a mixture of two overlapping subgroups, corresponding to plausible and implausible cases. The data were, therefore, modeled using a two-component beta-binomial mixture model, which was considered more adequate than a one- or three-component model to represent the evaluation process and likely to yield a superior fit. The assumption of two subgroups was driven by the clearly bimodal pattern observed in both pension and accident insurance cases, and supported empirically by Akaike Information Criterion (AIC) comparisons, which clearly favored the two-component over the one-component model.

The resulting two-component, five-parameter beta-binomial mixture model used (i) pP, the probability that a CVA criterion is conspicuous in plausible cases, (ii) pIP, the corresponding probability in implausible cases, (iii) SDpP and (iv) SDpIP as dispersion parameters, and (v) the mixing ratio between plausible and implausible cases. Model fitting was performed in RStudio (RStudio Team, 2015) using maximum-likelihood estimation of the five-parameter two-component beta-binomial mixture (two mean probabilities, two dispersion parameters, and one mixing proportion). Goodness-of-Fit was evaluated using deviance, Pearson's χ^2, sum of squared errors (SSE), and information criteria (AIC, BIC). Model-based classification thresholds (CVA≥t indicating implausibility) were derived from posterior

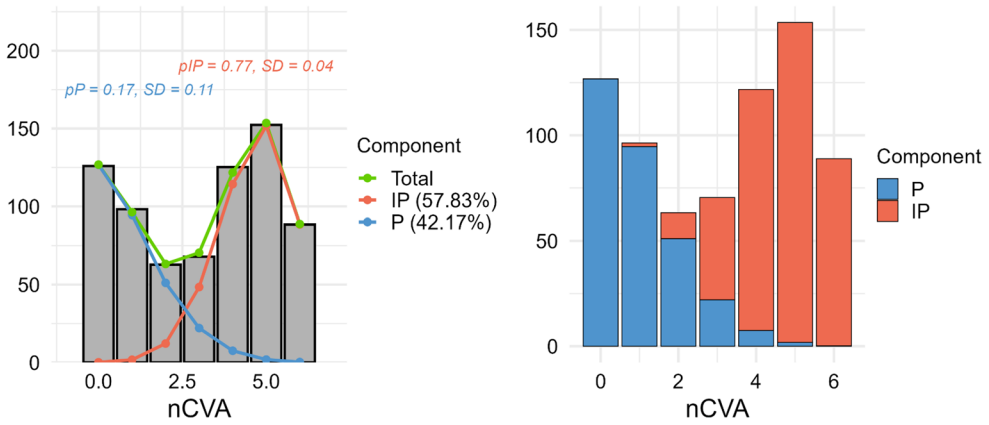

Fig. 2 Fitted and Observed Frequencies of Conspicuous Non-Psychometric CVA Criteria (Full MEWA Dataset) Note. nCVA = number of conspicuous CVA criteria; P = plausible component; IP = implausible component. Blue = plausible component (P); red = implausible component (IP); green = combined mixture fit; grey bars = observed frequencies. The left panel shows the fitted frequencies for the plausible (P) and implausible (IP) mixture components and the resulting combined fit. The right panel shows the corresponding expected frequencies produced by the mixture model as stacked bars. The strong correspondence between fitted and observed values demonstrates that the beta-binomial mixture model reliably supports a two-component representation of the empirical distribution with high accuracy

Table 3 Parameter estimates of the two-component beta-binomial mixture models fitted to real-world non-psychometric CVA data

Case plausibility	*p* (conspicuous)	SD	Relative frequency
All data ($n=721$)			
Plausible (P)	0.17	0.11	43.35%
Implausible (IP)	0.79	0.04	56.65%
Pension Insurance ($n=572$)			
Plausible (P)	0.19	0.12	39.71%
Implausible (IP)	0.77	0.04	60.29%
Accident Insurance ($n=149$)			
Plausible (P)	0.13	0.10	56.11%
Implausible (IP)	0.87	0.03	43.89%

Note. p (conspicuous) = subgroup-specific probability that a CVA criterion is rated conspicuous; SD = dispersion of this probability; Relative Frequency = proportion of plausible (P) and implausible (IP) cases. Values are maximum-likelihood estimates obtained by fitting two-component beta-binomial mixture models to the empirical distribution of conspicuous non-psychometric CVA criteria in the imputed MEWA dataset. These estimates form the empirical basis for defining plausibility thresholds

component probabilities by maximizing the F1-score. To account for potential differences between pension and accident insurance cases, we estimated separate models for each case type as well as for the combined dataset. Table 3 summarizes the parameter estimates of the mixture models and supports the assumption of two distinct subgroups. Across all subgroups, plausible cases (P) show a lower probability (approximately 17 ± 11% (Mean, SD) that a non-psychometric CVA criterion is rated conspicuous, whereas implausible cases (IP) consistently show a much higher probability (approximately 79 ± 4% (Mean, SD).

Table 4 summarizes the Goodness-of-Fit statistics, indicating that the empirical distribution of non-psychometric CVA criteria is accurately reproduced by the two-component beta-binomial mixture model. Figure 2 illustrates this by showing the fitted frequencies of the plausible and implausible components and their contribution to the overall distribution. Separate models for accident and pension

Table 4 Goodness-of-fit results from two-component beta-binomial mixture models capturing plausible and implausible subgroups in real-world MEWAs (non-psychometric CVA criteria)

Data	-logL	Deviance	Pearson χ^2	SSE	AIC	BIC
All data ($n=721$)	-1369.30	0.20	0.20	16.89	2748.61	2771.51
Pension Insurance ($n=572$)	-1086.05	0.51	0.50	36.12	2182.10	2203.85
Accident Insurance ($n=149$)	-267.94	0.22	0.22	4.18	545.87	560.89

Note. $-logL$ negative log-likelihood, Deviance and Pearson's χ^2 = discrepancy measures between observed and model-predicted frequencies; *SSE* sum of squared errors, AIC/BIC = penalized likelihood criteria for comparing model fit and complexity. The table reports Goodness-of-Fit statistics for the two-component beta-binomial mixture models applied to the imputed distributions of non-psychometric CVA criteria in the full dataset and in pension and accident insurance cases. The very low Deviance, Pearson's χ^2, and SSE indicate excellent agreement between the fitted mixture model and the empirical data

insurance cases are presented in supplementary material (Figure S1 and S2, supplementary material B) and exhibit comparable patterns.

A subsequent threshold analysis was performed to determine the optimal threshold for classifying cases as implausible using the F1-score, with detailed results provided in Table S1 in the supplementary material. Although the F1-score favored a threshold of ≥3 conspicuous criteria, a specificity-focused approach indicated ≥4 conspicuous criteria as the preferred threshold to minimize false positives. Thus, cases with ≥4 non-psychometric CVA criteria can be classified as implausible, irrespective of case type.

Descriptive Data

Using the optimal ≥4 threshold, 340.57 (47.24%) cases were rated as plausible and 380.43 (52.76%) cases were rated as implausible after imputation of missing psychometric test data and CVA criteria. The decimal values are a product of imputation and its subsequent uncertainty.

SIMS scores ranged from 0 to 64, with higher scores indicating higher levels of feigning of a mental disorder, while ASTM scores ranged from 38 to 90, with lower scores indicating higher probabilities of invalid cognitive test results. The number of non-conspicuous CVA criteria ranged from 0 to 6. Table 5 shows the results regarding descriptive analysis of mean SIMS, ASTM, and mean number of conspicuous CVA criteria.

Table S2 shows the absolute and relative frequencies of CVA criteria across case type and case plausibility. Since this analysis is of marginal importance to the research question, the table and text can be found in supplementary material B and C.

SIMS

The initial linear model (Table 6), conducted without the removal of outliers, revealed a significant effect of case type, case plausibility, and age on SIMS scores. Cohen's d was calculated using group means and a pooled standard deviation. Case type showed a medium effect (d = 0.48),

Table 5 Descriptive characteristics of SIMS, ASTM, and the number of conspicuous CVA criteria by case type, case plausibility, and gender

Condition (%)	SIMS		ASTM		Conspicuous CVA (max. 6)	
	Mean	SD	Mean	SD	Mean	SD
Case Type						
Pension (79.33%)	21.59	9.66	78.72	10.13	3.24	2.04
Accident (20.67%)	16.98	9.10	81.69	9.52	2.73	2.39
Case Plausibility						
Plausible (46.88%)	17.26	8.38	82.40	7.78	1.17	1.15
Implausible (53.13%)	23.66	9.84	76.58	11.05	4.89	0.89
Gender						
Female (50.69%)	20.47	9.13	79.61	9.90	3.18	2.11
Male (49.31%)	20.81	10.31	79.05	10.26	3.09	2.14
Case Type*Plausibility						
Pension*Plausible (45.81%)	18.32	8.54	81.72	8.03	1.33	1.19
Pension*Implausible (54.19%)	24.40	9.68	76.13	11.00	4.89	0.81
Accident*Plausible (51.00%)	13.53	6.59	84.78	6.33	0.60	0.76
Accident*Implausible (49.00%)	20.59	9.95	78.44	11.11	4.94	1.14
Case Type*Gender						
Pension*Female (54.20%)	21.44	9.05	78.64	10.16	3.29	2.06
Pension*Male (45.80%)	21.76	10.34	78.81	10.12	3.19	2.02
Accident*Female (37.16%)	15.08	7.56	84.96	5.91	2.55	2.28
Accident*Male (62.84%)	18.12	9.77	79.73	10.69	2.84	2.45
Case Plausibility*Gender						
Female*Plausible (49.72%)	17.38	8.51	82.51	7.80	1.21	1.18
Male*Plausible (50.28%)	17.13	8.25	82.30	7.77	1.13	1.12
Female*Implausible (51.55%)	23.18	8.79	77.06	10.79	4.89	0.87
Male*Implausible (48.45%)	24.17	10.84	76.08	11.32	4.90	0.91
Total	20.64	9.72	79.33	10.08	3.14	2.12

Note. *SIMS* Structured Inventory of Malingered Symptomatology, *ASTM* Amsterdam Short-Term Memory Test, *CVA* Criteria-Based Validity Assessment. All descriptive values refer to the fully imputed dataset; fractional case counts reflect the probabilistic nature of multiple imputation. Percentages in parentheses indicate subgroup proportions. The table summarizes descriptive statistics for SIMS and ASTM scores and for the number of conspicuous non-psychometric CVA criteria across subgroups defined by case type, case plausibility, and gender

Table 6 Effects of case type, case plausibility, age, and gender in linear regression models predicting SIMS scores (with and without outliers)

Predictor	Estimate	Robust SE	Robust *p*-Value
SIMS (with outliers)			
Intercept	12.16	1.94	0.000 ***
Case Type	-4.01	0.77	0.000 ***
Case Plausibility	6.34	0.67	0.000 ***
Age	0.10	0.04	0.010 *
Gender	0.78	0.68	0.302
SIMS (without outliers)			
Intercept	11.99	1.86	0.000 ***
Case Type	-3.98	0.69	0.000 ***
Case Plausibility	5.76	0.61	0.000 ***
Age	0.11	0.03	0.004 **
Gender	0.16	0.63	0.780

Note. *SE* Standard Error. This table summarizes regression estimates examining how case type, case plausibility, age, and gender relate to SIMS scores. The Intercept denotes the model-predicted SIMS score for the reference categories when all predictors are held at baseline. The lower section reports estimates from models in which all outlier cases (identified across imputations) were removed

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

indicating that mean SIMS scores were higher in pension insurance cases ($M = 21.59$; $SD = 9.66$) than in accident insurance cases ($M = 16.98$; $SD = 9.10$). Case plausibility demonstrated a large effect ($d = 0.70$), with implausible cases exhibiting higher mean SIMS scores ($M = 23.66$; $SD = 9.84$) than plausible cases ($M = 17.26$; $SD = 8.38$). Age showed only a small effect on SIMS scores with meaningful differences only emerging with large age differences. No significant differences in SIMS scores were found based on gender.

The sensitivity analysis, conducted after removing outliers, revealed slight changes in the model. The effects of case type ($d = 0.55$) (Fig. 3) and case plausibility ($d = 0.70$) became slightly larger. Additionally, age remained a minimal predictor, with each unit increase in age associated with a 0.104-unit increase in SIMS scores, independent of case type and case plausibility. The sensitivity analysis suggests that the inclusion of outliers in the initial model led to a small underestimation of the observed effects.

ASTM

The linear model (Table 7) revealed a significant effect of case type, case plausibility, and age. Pension insurance cases showed significantly lower ASTM scores ($M = 78.72$; $SD = 10.13$, $d = -0.32$) than accident insurance cases ($M = 81.69$; $SD = 9.52$). Plausible cases scored significantly higher in the ASTM ($M = 82.40$; $SD = 7.78$, $d = 0.61$) than implausible cases ($M = 76.58$; $SD = 11.05$). A significant effect of age on ASTM scores was found as well. No significant effect of gender on ASTM scores was found.

After removing outliers, the results changed only slightly. The effect of case type turned slightly larger ($d = -0.37$) (Fig. 4), while the effect of case plausibility did not change significantly ($d = 0.61$). In conclusion, the removal of outliers revealed that the model including outliers underestimated the effects of case type slightly.

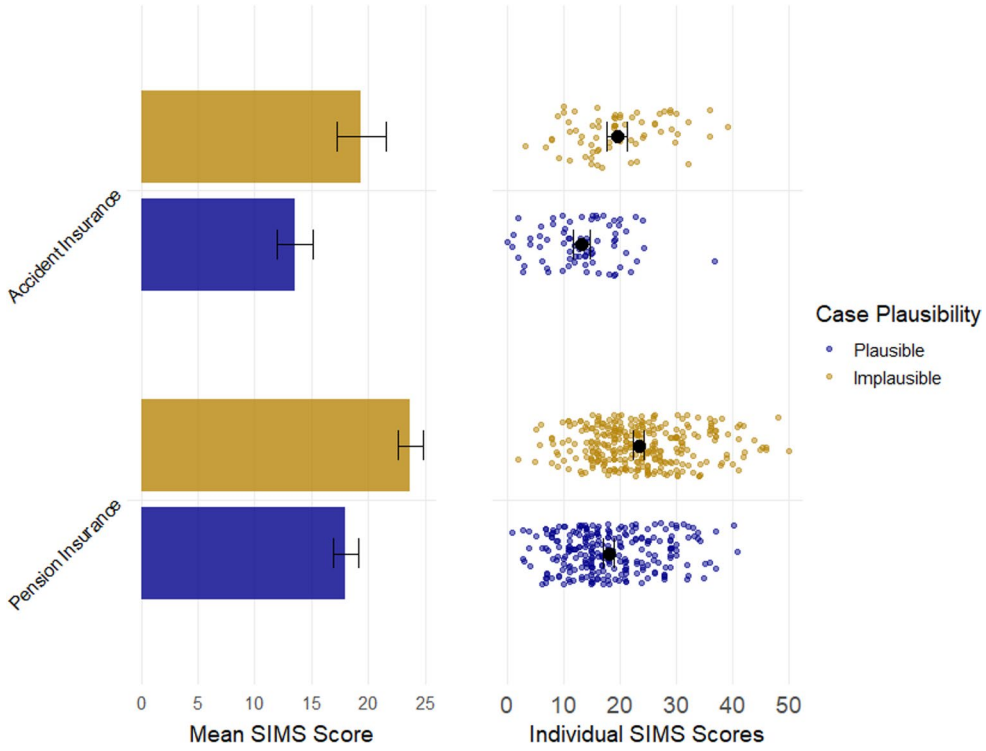

Fig. 3 SIMS Score Distributions by Case Type and Case Plausibility Note. SIMS = Structured Inventory of Malingered Symptomatology. Panels: The left panel displays pooled mean SIMS scores by case type and case plausibility in the imputed data after removal of outliers (studentized residuals > ±3) according to Rubin's rules; black bars indicate 95% confidence intervals. The right panel shows individual SIMS scores from the randomly selected dataset 12 for better visibility, together with the mean (black dot) and its 95% confidence interval (black bars). These visualizations illustrate the score distributions underlying the regression analyses

Table 7 Effects of case type, case plausibility, age, and gender in linear regression models predicting ASTM scores (with and without outliers)

Predictor	Estimate	Robust SE	Robust *p*-value
ASTM (with outliers)			
Intercept	7485.00	289.62	0.000 ***
Case Type	393.90	124.93	0.003 **
Case Plausibility	-859.03	103.07	0.000 ***
Age	-12.32	5.23	0.038 *
Gender	-125.14	104.03	0.272
ASTM (without outliers)			
Intercept	7464.89	285.63	0.000 ***
Case Type	434.15	118.82	0.001 **
Case Plausibility	-815.98	100.00	0.000 ***
Age	-11.83	5.14	0.038 *
Gender	-117.93	101.23	0.280

Note. *SE* Standard Error. This table summarizes regression estimates examining how case type, case plausibility, age, and gender relate to ASTM performance. Due to non-normal residuals, a Box–Cox transformation ($\lambda = 2$) was applied to the ASTM scores; all coefficients refer to this transformed scale. The Intercept represents the expected transformed ASTM score for the reference groups with predictors set to baseline. The lower section shows model estimates after exclusion of all identified outliers across imputations

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

Table 8 Effects of case type, age and gender in the logistic regression model examining CVA-derived case plausibility

Predictor	Estimate	SE	OR	95% CI for OR	*P*-value
Constant	0.39	0.79	1.47	[0.31; 7.08]	0.626
Case Type	−0.20	0.28	0.82	[0.47; 1.43]	0.476
Age	−0.00	0.01	1.00	[0.97; 1.02]	0.770
Gender	−0.02	0.20	0.98	[0.65; 1.46]	0.913

Note. *SE* Standard Error, *OR* odds ratio, *CI* Confidence Interval. Case plausibility was defined using the CVA threshold of ≥ 4 conspicuous non-psychometric criteria. The table shows logistic regression coefficients assessing whether case type, age, or gender are associated with this CVA-based implausibility classification. The Intercept represents the baseline log-odds of implausibility for the reference categories. No predictor reached statistical significance, consistent with CVA plausibility being independent of demographic variables and case type

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

Case Plausibility

Table 8 shows the results of the logistic regression model using case plausibility as the dependent variable to explore differences regarding case type while controlling for possible effects of age and gender. The D3 statistic was used to compare the imputed model with an intercept-only model ($D3(3, 765.91) = 0.20$, $p = .896$, $riv = 1.157$), revealing that the model does not perform significantly better than an intercept-only model. This shows that these predictors are not well suited to predict case plausibility. The Hosmer-Lemeshow test showed a good fit ($F(8, 1206.37) = 0.47$, $p = .881$). The mean Nagelkerke's R^2 across all analyses stood at 0.01, reflecting a low proportion of explained variance, demonstrating weak predictive power of the covariates in the model.

None of the predictors reached statistical significance, indicating that neither case type, age, nor gender was associated with case plausibility. The odds that a case is rated as implausible, therefore, do not differ between pension and accident insurance cases or between male and female claimants.

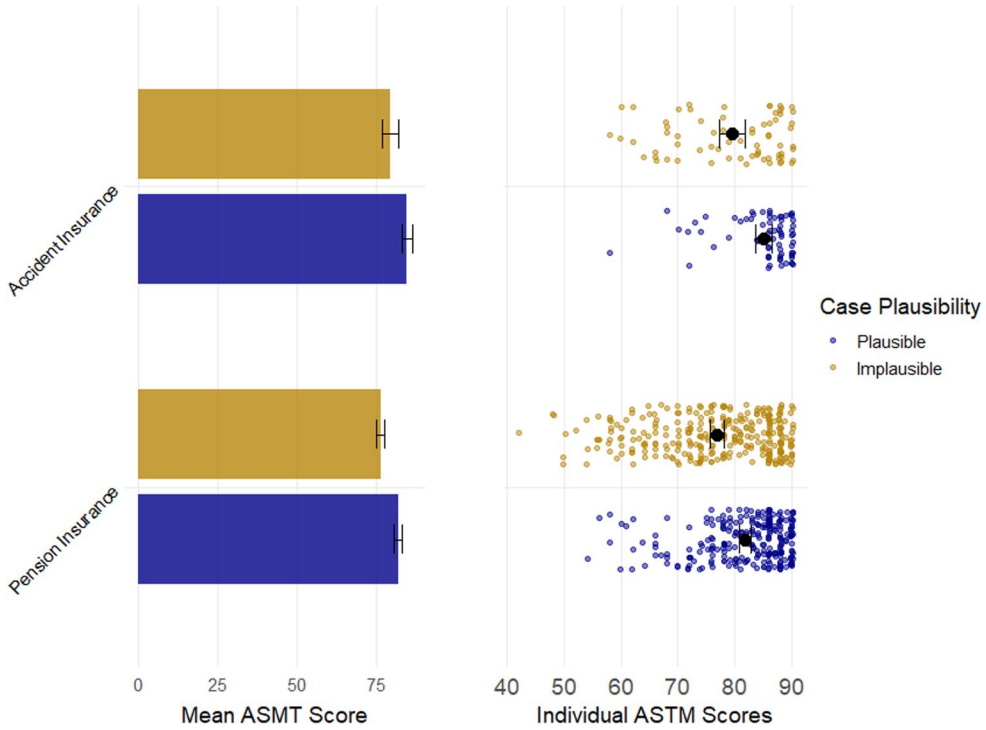

Fig. 4 ASTM Score Distributions by Case Type and Case Plausibility Note. ASTM = Amsterdam Short-Term Memory Test. Panels: The left panel shows the pooled mean ASTM scores across case type and case plausibility in the imputed data after removal of outliers (studentized residuals > ±3) according to Rubin's rules; black bars indicate 95% confidence intervals. The right panel presents the corresponding individual ASTM scores from the randomly selected dataset 12 for better visibility, with the mean (black dot) and its 95% confidence interval overlaid. Together, the panels illustrate the distributional structure underlying the regression analyses

Integrating Psychometric Test Results into CVA

The first validation study of CVA was limited by the necessity of excluding criterion C6 and using it as an external validation benchmark. Because CVA is an imperfect reference proxy rather than a diagnostic gold standard, an ROC analysis reported in supplementary material C was treated as exploratory and CVA-specific. Its purpose was to examine how criterion C6 could be incorporated into CVA under a conservative specificity-focused approach, not to establish generally applicable SIMS or ASTM thresholds. Because CVA data were limited and SIMS items may overlap with genuine psychopathology, the possibility of misclassification rather than upward-adjusted thresholds should be considered. These thresholds are not generally verified for every MEWA context and should be handled accordingly until CVA proves its diagnostic accuracy.

Using these exploratory CVA-specific thresholds, criterion C6 was calculated from conspicuous SIMS or ASTM results in this dataset. Applying the same two-component beta-binomial mixture model described in the methods section yielded similar results in terms of parameter estimations and Goodness-of-Fit. The specificity-focused approach again indicated an optimal CVA threshold of ≥4 conspicuous criteria. Tables regarding parameters, Goodness-of-Fit and thresholds are provided in the supplementary material (Table S6, S7, and S8, supplementary material B). Figure 5 depicts this fit for the whole dataset. A separate model for pension or accident insurance cases showed similar results and can be found in the supplementary material (Figures S3 and S4, supplementary material B) as well.

As a final step, comparing the two-component beta-binomial model and a one-component model revealed a significantly better fit of the two-component model ($\Delta AIC = 56.22$; $\Delta BIC = 42.48$). Comparing a two-component model using three parameters with a model using five parameters revealed a better fit of the five-parameter model regarding the AIC ($\Delta AIC = 3.16$, ΔBIC = -6.00) and Goodness-of-Fit statistics (Table S9, supplementary material B), making a two-component beta-binomial model with five parameters the optimal modeling choice for this kind of data.

This threshold is practical if all seven criteria are accurately evaluated during the MEWA process, but criterion C7 (laboratory results) is often left unevaluated (288 missing values). Naturally, this happens when a claimant reports that they do not take any medication, making a blood serum test obsolete and explaining a high proportion of missing C7 evaluations. To avoid relying on extrapolation of thresholds regarding case plausibility in MEWAs with missing criterion C7, an additional two-component beta-binomial mixture model was fitted using six CVA criteria, this time excluding criterion C7, i.e., laboratory assessment of the medication taken (Fig.6). Parameter estimations and Goodness-of-Fit statistics did not significantly deviate from the previous models, still identifying a threshold of ≥4 conspicuous CVA criteria as the optimal threshold across pension and accident insurance cases. Tables and figures are provided in the supplementary material (Table S10, S11, and S12, supplementary material B).

Procedural and Pre-Procedural Accident Insurance Cases

We further explored whether there were differences between procedural (MEWAs for court proceedings) and pre-procedural (administrative) accident insurance MEWAs because GPI cases solely contained court-ordered procedural MEWAs. This is due to the fact that pre-procedural

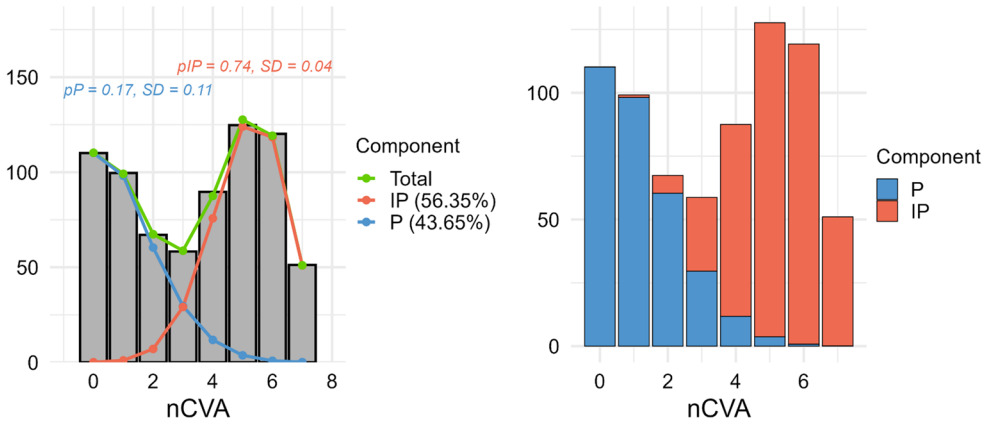

Fig. 5 Two-Component Beta-Binomial Mixture Fit Using all Seven CVA Criteria (Full MEWA Dataset) Note. nCVA = number of conspicuous CVA criteria; P = plausible component; IP = implausible component. Blue = plausible component; red = implausible component; green = summed mixture fit. The left panel presents the fitted frequencies for the plausible and implausible mixture components and their combined curve. The right panel shows the corresponding expected frequencies produced by the mixture model as stacked bars. The figure illustrates that even when all seven CVA criteria are included, the two-component mixture model aligns very closely with the data and maintains a stable separation between the latent components

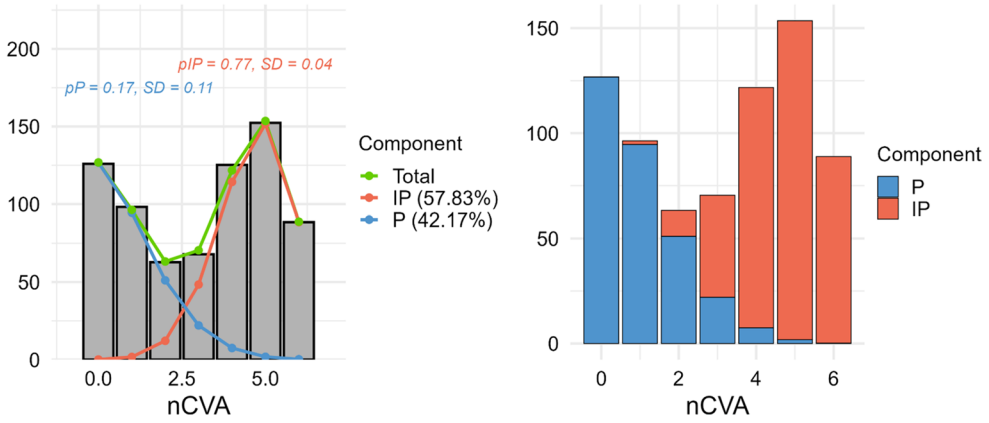

Fig. 6 Two-Component Beta-Binomial Mixture Fit Using the Six Non-Laboratory CVA Criteria (Full MEWA Dataset) Note. nCVA = number of conspicuous CVA criteria; P = plausible component; IP = implausible component. Panels: The left panel shows the fitted frequencies for the plausible (blue) and implausible (red) components of the two-component beta-binomial mixture model using all non-laboratory CVA criteria, together with the combined fit (green). The percentages in the legend indicate the estimated proportions of cases assigned to each component. The right panel displays the corresponding model-based expected frequencies as stacked bars

MEWAs for GPI have a special short format without extensive evaluation and typically without VTs and CVA. These assessments will hopefully be added in the future by the pension insurance administration and will then be accessible for analysis. Nonetheless, even though identical in format and content, the group of accident insurance cases is heterogeneous according to their placement in different stages in judicial instances, whereas GPI cases were more homogeneous in that respect. Although they seemed very similar, pre-procedural and procedural accident insurance cases had to be checked for differences. This analysis of only procedural accident and pension insurance cases can be found in supplementary material D.

An additional exploration of only procedural and pre-procedural MEWAs regarding the German statutory accident insurance (GAI) can be found in supplementary material E.

Discussion

This work addresses a central gap in the forensic assessment literature: despite the widespread use of CVA in German MEWAs, empirical evidence for its structure and thresholds has been limited. Our analyses of more than 700 real-world assessments show a consistent two-component pattern, and a threshold of ≥4 conspicuous criteria emerged as a conservative CVA-derived indicator of implausible response behavior across pension and accident insurance cases.

SIMS and ASTM scores differed across legal contexts, with higher SIMS and lower ASTM scores in pension assessments. These differences should not be interpreted as direct evidence of greater invalidity. They may reflect response bias, but also could be an expression of symptom burden, diagnostic composition, incentive structure, or a combination of these factors. Because no differences were found in CVA-derived plausibility ratings between pension and accident insurance cases, the findings can be interpreted in two ways: CVA may capture a stable multimodal case-level pattern across legal contexts, or it may be insufficiently sensitive to some context-specific differences in invalid responding. Conservative clinical coding of qualitative inconsistencies, incorporation bias–since the medical expert evaluates each CVA criterion with knowledge of each other criterion–and common judgment errors such as halo effects could contribute to apparent stability and must be considered when interpreting the results. In this case, a comparison of CVA complemented by accident-specific validity indicators, such as a pre-post-assessment of functional capability with regard to the accident, could highlight limitations of CVA and should be investigated in future work.

Our first aim was to revisit the plausibility threshold based on the six non-psychometric CVA criteria and assess whether it can be applied beyond the field of pension cases. The two-component beta-binomial mixture model–capturing plausible and implausible subgroups–showed an excellent fit to the data and supported the expected latent structure. This pattern appeared broadly similar across pension and accident insurance cases, with no meaningful differences in model performance or threshold characteristics. Using a specificity-focused approach to reduce false-positive classifications (Bianchini et al. 2005; Young et al. 2025a, b), a threshold of ≥4 conspicuous non-psychometric CVA criteria emerged as the conservative choice. Thus, across differing legal questions, CVA provides a preliminary empirical basis for plausibility classification when using non-psychometric criteria alone.

In practical application, while not measuring person-related but case-related features, these thresholds can be interpreted as a graded CVA-derived plausibility signal:

cases with ≤2 conspicuous CVA criteria may be plausible. However, negative validity findings cannot be interpreted as 'good effort' but simply as lack of evidence for malingering (Chafetz, 2022). Cases with three conspicuous criteria require individual examination because distortion is possible but not certain, and ≥4 conspicuous criteria indicate CVA-derived implausibility of the case. CVA does not equate implausibility with feigning, and it does not exclude genuine illness. Some individuals exhibit feigning while simultaneously coping with substantial or causally relevant disorders. This highlights the importance of formulating expert reports in a non-pejorative, respectful tone, even when CVA shows conspicuousness. CVA supports, but never replaces, the nuanced clinical judgment required in MEWAs.

Further analysis of multivariate differences in psychometric test scores revealed higher mean SIMS and lower mean ASTM scores in pension insurance cases compared to accident insurance cases, independent of case plausibility, age, or gender. We consider this relevant because it indicates that the nature of the legal case already influences psychometric outcomes as a measure of observable response behavior. One possible explanation is that pension insurance claimants may have a stronger incentive to report the full breadth of physical and mental struggles, consistent with evidence that disability claimants often report multiple problems (Brongers et al., 2022), which is the optimal strategy, considering that more illnesses/disorders collectively lead to a lower capacity to work. Accident insurance cases, by contrast, tend to focus on specific disorders caused by an accident. Reporting a wide range of symptoms could undermine the causal attribution of the accident to its consequences. This interpretation is consistent with research on response distortion, suggesting that feigning is implemented via qualitatively different strategies that are shaped by situational incentives and constraints (Röhner et al., 2025). Diagnostic composition and symptom severity were not available in the present dataset and remain an alternative explanation of these patterns.

While claimants in pension insurance cases can show symptom underreporting, particularly for disorders which are not socially acceptable or disease-specific (Rogers & Bender, 2018), symptom underreporting is, in theory, more prevalent in accident insurance cases, in which pre-accident disorders act like an obstacle regarding causal attribution. No study to date has made a direct comparison between accident and pension insurance cases with regard to the prevalence of symptom underreporting. Moreover, SIMS items may be endorsed because they are perceived as reflecting impaired concentration, memory problems, depressive symptoms, or other genuine complaints, which could contribute to higher scores when claimants report symptoms broadly.

Regarding the ASTM, a test primarily assessing non-credible performance levels, performing well is not the strategically optimal choice for claimants. High effort can be perceived as a risk and, therefore, be avoided. Along with a fear of negative consequences from optimal performance (Shefer et al., 2016), ASTM scores can fall below common thresholds and practitioners could assume feigning, even when no intentional symptom overreporting is present. Consequently, the goal of reporting many disorders, without conscious severe symptom overreporting, is a typical incentive in pension insurance claims and could lead to false positives regarding SVT/PVT results.

Consistent with validity-test theory, CVA-derived implausible cases showed higher mean SIMS scores and lower mean ASTM scores than CVA-derived plausible cases. This provides further evidence of convergence between CVA and psychometric validity indicators. However, because CVA is not an established diagnostic gold standard and the dataset partly overlaps with the previous validation study (Klett et al., 2026), these findings should be interpreted as support for CVA structure, not as definitive diagnostic accuracy evidence. Additionally, the odds of a case being rated as implausible in terms of CVA did not differ between pension and accident insurance cases, which may reflect either cross-context stability or reduced sensitivity to context-specific base-rate differences.

Using the exploratory CVA-specific thresholds, criterion C6 could be incorporated into the present CVA model. Transferring the two-component beta-binomial mixture model approach to CVA using all seven criteria yielded similar results, and a threshold of ≥4 conspicuous criteria remained the conservative CVA-derived plausibility threshold.

If the threshold of ≥4 conspicuous criteria appears unchanged after adding SVT/PVT results, the question arises whether psychometric testing is still needed. Our findings support its continued relevance. SVTs and PVTs provide an additional modality that cannot be derived from clinical impressions or file-based inconsistencies alone. They may identify response patterns not captured by qualitative CVA criteria, while the qualitative criteria may contextualize isolated psychometric findings. Thus, psychometric testing should be integrated into CVA as one source of validity evidence and interpreted alongside clinical, behavioral, file-based, and laboratory information.

Limitations

This study used data from real court cases; therefore, the actual extent and motivation of invalid response behavior cannot be known. CVA was used as an imperfect reference proxy rather than as a diagnostic gold standard.

Consequently, ROC-derived SIMS and ASTM thresholds must be interpreted as CVA-specific exploratory estimates. They do not establish new generally applicable SIMS or ASTM thresholds. Misclassification within CVA components, conservative clinical coding of qualitative inconsistencies, and incorporation bias may have influenced the observed thresholds and group differences.

Diagnostic composition and symptom severity were not systematically modeled. Therefore, differences in SIMS and ASTM scores may reflect symptom burden or diagnostic composition rather than invalid responding or legal context alone. This is a central limitation of the present analyses and should be addressed in future work.

This study provided further evidence of CVA stability across two legal questions. Assumptions about applicability to other legal questions, such as residence permit proceedings or assessment of job suitability, cannot be made at this point. Further research about CVA in other contexts and legal fields is necessary to identify potential obstacles or limits of this system.

This dataset consisted of real-world data and therefore contained substantial missing data. Although multiple imputation, the current state-of-the-art method, was used to handle missing data, the high proportion of cases with at least one missing value remains a limitation. In particular, 'not evaluated' CVA criteria may not be equivalent to randomly missing information, because non-evaluation can reflect case-specific procedural or clinical circumstances. Additionally, because the study dealt with real forensic cases, the study design was retrospective in nature and did not adhere to the conventions of a randomized controlled trial.

Conclusion

This study provides further empirical support for the structure of CVA in real-world MEWAs. The core pattern remained broadly consistent across pension and accident insurance cases and when criterion C6 or C7 was varied, with ≥4 conspicuous criteria emerging as a conservative indicator of invalid response behavior.

At the same time, observed differences in SIMS and ASTM scores between case types highlight the need for cautious, context-sensitive interpretation of psychometric validity tests. The ROC-based SIMS/ASTM analyses are exploratory and CVA-specific only; they do not establish independent new thresholds. SVTs and PVTs offer an independent modality within the CVA framework and add information that cannot be obtained from behavioral observations or file-based inconsistencies alone. Practitioners should therefore assess as many CVA criteria as possible and integrate psychometric results with clinical, behavioral, file-based, and laboratory information in order to minimize false-positive classifications. Age and gender showed little influence on CVA, SIMS, or ASTM results, suggesting limited relevance for validity assessment.

In sum, these findings indicate that CVA may serve as a structured multimodal reference framework within MEWAs, provided that psychometric data, clinical judgment, and the limitations of CVA as an inherently imperfect reference proxy are explicitly considered.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12207-026-09567-w.

Declarations

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

References

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

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: Official Journal of the North American Spine Society*, *5*(4), 404–417. https://doi.org/10.1016/j.spinee.2004.11.016

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

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. (2022). Deception is different: Negative validity test findings do not provide evidence for good effort. *The Clinical Neuropsychologist*, *36*(6), 1244–1264. https://doi.org/10.1080/13854046.2020.1840633

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. https://doi.org/10.1007/s00115-002-1438-5

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Deutsche Rentenversicherung Bund (2024). *Rentenversicherung in Zeitreihen*. https://www.deutscherentenversicherung.de/SharedDocs/Downloads/DE/Statistiken-und-Berichte/statistikpublikationen/rv_in_zeitreihen.pdf

DGUV (2025). *Arbeitsunfallgeschehen 2024*. https://publikationen.dguv.de/statistiken/arbeitsunfallgeschehen/5157/arbeitsunfallgeschehen-2024

Fuermaier, A. B. M., Dandachi-Fitzgerald, B., & Lehrner, J. (2023). Attention performance as an embedded validity indicator in the cognitive assessment of early retirement claimants. *Psychological Injury and Law*, *16*(1), 36–48. https://doi.org/10.1007/s12207-022-09468-8

Fuermaier, A. B. M., Dandachi-Fitzgerald, B., & Lehrner, J. (2025). Validity assessment of early retirement claimants: Symptom overreporting on the beck depression inventory - II. *Applied Neuropsychology Adult*, *32*(3), 712–718. https://doi.org/10.1080/23279095.2023.2206031

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

Klett, N., Dohrenbusch, R., Fischer, A., Geiger, T., Keller, F., Kornhuber, J., Littke, O., Schütz, A., Siegmann, E. M., Käfferlein, W., Grömer, T., & Capito, E. (2026). Criteria-based validity assessment in legal cases of claimed reduced work capacity. *Psychological Injury and Law*, *19*(1). https://doi.org/10.1007/s12207-026-09557-y

Kobelt-Pönicke, A., & Walter, F. (2020). Beschwerdenvalidierung in der sozialmedizinischen Begutachtung. *Zeitschrift für Psychiatrie Psychologie und Psychotherapie*, *68*(2). https://doi.org/10.1024/1661-4747/a000405

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

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (Fifth Edition). McGraw-Hill.

Martin, P. K., & Schroeder, R. W. (2020). Base rates of invalid test performance across clinical non-forensic contexts and settings. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *35*(6), 717–725. https://doi.org/10.1093/arclin/acaa017

Merten, T., Krahl, G., Krahl, C., & Freytag, H. W. (2010). [Base-rate estimates for negative response bias in a workers' compensation claim sample]. *Versicherungsmedizin / Herausgegeben Von Verband Der Lebensversicherungs-Unternehmen E.V. Und Verband Der Privaten Krankenversicherung E.V*, *62*(3), 126–131.

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://doi.org/10.1076/jcen.24.8.1094.8379

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

Rai, J. K., Gervais, R. O., & Erdodi, L. A. (2023). A large-scale investigation of the classification accuracy of various performance validity tests in a medical-legal setting. *Psychology & Neuroscience*, *16*(3), 225–243. https://doi.org/10.1037/pne0000320

Röhner, J., Schütz, A., & Ziegler, M. (2025). Faking in self-report personality scales: A qualitative analysis and taxonomy of the behaviors that constitute faking strategies. *International Journal of Selection and Assessment*, *33*(1), e12513. https://doi.org/10.1111/ijsa.12513

Röhner, J., Thoss, P., & Schütz, A. (2022). Lying on the dissection table: Anatomizing faked responses. *Behavior Research Methods*, *54*(6), 2878–2904. https://doi.org/10.3758/s13428-021-01770-8

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

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

Shefer, G., Henderson, C., Frost-Gaskin, M., & Pacitti, R. (2016). Only making things worse: A qualitative study of the impact of wrongly removing disability benefits from people with mental illness. *Community Mental Health Journal*, *52*(7), 834–841. https://doi.org/10.1007/s10597-016-0012-8

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

Stevens, A., Friedel, E., Mehren, G., & Merten, T. (2008). Malingering and uncooperativeness in psychiatric and psychological assessment: Prevalence and effects in a German sample of claimants. *Psychiatry Research*, *157*(1–3), 191–200. https://doi.org/10.1016/j.psychres.2007.01.003

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. (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

Teßmann, J., Fuermaier, A. B. M., Dandachi-FitzGerald, B., & Lehrner, J. (2025). The utility of an attention-based performance validity test in a sample of Austrian early retirement claimants. *Psychological Injury and Law*, *18*(4), 284–299. https://doi.org/10.1007/s12207-025-09544-9

van 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

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.

Walczyk, J. J., Sewell, N., & DiBenedetto, M. B. (2018). A review of approaches to detecting malingering in forensic contexts and promising cognitive load-inducing lie detection techniques. *Frontiers in Psychiatry*, *9*, 700. https://doi.org/10.3389/fpsyt.2018.00700

Wertz, M., Mader, E., Nedopil, N., Schiltz, K., & Yundina, E. (2021). Antwortverzerrung oder Symptombelastung? Beschwerdeschilderung von psychiatrischen Patienten und sozialmedizinischen Begutachtungsprobanden. *Der Nervenarzt*, *92*(11), 1163–1171. https://doi.org/10.1007/s00115-020-01041-5

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

Yan, X., & Su, X. (2009). *Linear Regression Analysis: Theory and Computing*.World Scientific. https://doi.org/10.1142/6986

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., 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

Authors and Affiliations

N. Klett[1,2,4] · R. Dohrenbusch[3] · E. M. Siegmann[2] · A. Schütz[4] · J. Kornhuber[2] · J. Röhner[4] · F. Keller[5] · E. S. Capito[1] · T. Grömer[1,2]

✉ N. Klett
noah.klett@gmx.de

R. Dohrenbusch
r.dohrenbusch@uni-bonn.de

E. M. Siegmann
eva-maria.siegmann@uk-erlangen.de

A. Schütz
astrid.schuetz@uni-bamberg.de

J. Kornhuber
johannes.kornhuber@uk-erlangen.de

J. Röhner
jessica.roehner@uni-bamberg.de

F. Keller
fritz.keller@live.de

E. S. Capito
praxis.capito@t-online.de

T. Grömer
tejagroemer@gmail.com

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

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

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

4 University of Bamberg, 96047 Bamberg, Germany

5 Thüringer Landessozialgericht, 99092 Erfurt, Germany