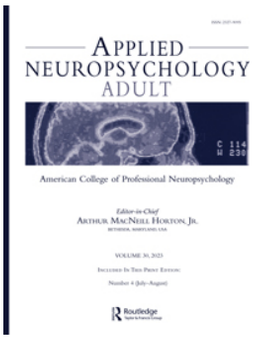

Applied Neuropsychology: Adult

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hapn21

Validity assessment of early retirement claimants: Symptom overreporting on the Beck Depression Inventory – II

Anselm B. M. Fuermaier, Brechje Dandachi-Fitzgerald & Johann Lehrner

To cite this article: Anselm B. M. Fuermaier, Brechje Dandachi-Fitzgerald & Johann Lehrner (2023): Validity assessment of early retirement claimants: Symptom overreporting on the Beck Depression Inventory – II, Applied Neuropsychology: Adult, DOI: 10.1080/23279095.2023.2206031

To link to this article: https://doi.org/10.1080/23279095.2023.2206031

APPLIED NEUROPSYCHOLOGY: ADULT
https://doi.org/10.1080/23279095.2023.2206031

Routledge
Taylor & Francis Group

Validity assessment of early retirement claimants: Symptom overreporting on the Beck Depression Inventory – II

Anselm B. M. Fuermaier[a], Brechje Dandachi-Fitzgerald[b,c], and Johann Lehrner[d]

[a]Department of Clinical and Developmental Neuropsychology, Faculty of Behavioral and Social Sciences, University of Groningen, Groningen, The Netherlands; [b]Department of Clinical Psychological Science, University of Maastricht, Maastricht, The Netherlands; [c]Faculty of Psychology, Open University, Heerlen, The Netherlands; [d]Department of Neurology, Medical University of Vienna, Vienna, Austria

ABSTRACT

Objectives: The Beck Depression Inventory-II (BDI-II) is a commonly used clinical measure; however, it contains no method to assess validity of self-report. The primary objective of this research was to examine the possibility of cut scores on the BDI-II indicating possible invalid symptom report in forensic neuropsychological evaluations. Secondary objectives were to explore the utility of education specific cut scores and the effects of the criterion for invalid symptom report.

Methods: Two hundred and seventeen early retirement claimants (age range 19–64 years) presenting for forensic neuropsychological examination were considered for this study. Invalid symptom report was determined based on two independent self-report symptom validity tests. Further, all individuals completed the BDI-II as part of their routine assessment battery.

Results: Individuals with invalid symptom report (30.9%) showed significantly higher BDI-II scores compared to individuals passing symptom validity assessment. ROC analysis supports the utility of the BDI-II to differentiate valid from invalid symptom report, AUC = 0.822, $SE = 0.032$, $p < .001$, 95%-CI = 0.760–0.884. A BDI-II cut score of 38 points reached a desired level of 0.90 specificity with 0.58 sensitivity. Secondary analysis indicated that the recommended cut score may vary depending on the educational level of the examinee. Further, results seem to be largely robust against the chosen criterion for invalid symptom report.

Conclusion: The BDI-II appears to be a useful adjunct embedded validity indicator in forensic neuropsychological evaluations.

KEYWORDS

Beck Depression Inventory; embedded validity testing; retirement claimants; symptom validity; validity assessment

The assessment of possible symptom overreporting (i.e., symptom validity) and cognitive underperformance (i.e., performance validity) with proven validity measures are essential features of any assessment context, including clinical and rehabilitation settings (e.g., Carone & Bush, 2018; McWhirter et al., 2020; Roor et al., 2023) and forensic evaluations (e.g., Bush et al., 2014; Sherman et al., 2020; Sweet et al., 2021). An extensive body of research was dedicated to the development and evaluation of self-report symptom validity tests (SVTs), which help to evaluate whether the symptom report can be trusted with a sufficient degree of confidence. Performance validity research has recommended the use of multiple validity indicators in order to sample validity continuously throughout an assessment and across clinical domains (see for example Fuermaier et al., 2023; Rhoads et al., 2021; Soble, 2021; Sweet et al., 2021). Freestanding validity indicators are known for their high diagnostic accuracy in distinguishing credible from noncredible performance, although the additional time required to administer stand-alone measures is not always feasible. Embedded validity indicators (as originally proposed for performance validity tests, PVT) have the potential to make a crucial contribution, because they have the advantage that they automatically sample various clinical domains as they are derived from routinely administered measures. If the performance is deemed valid, the routine measure provides information on a clinically relevant construct. Diagnostic accuracy of embedded validity assessment can be improved by aggregating multiple indicators. Although these recommendations were primarily derived from performance validity research, the conclusion to apply multiple SVTs of different test principles in a clinical examination appears evident (e.g., see Sherman et al., 2020, for criteria of malingered neurocognitive dysfunction).

Depressive symptoms play a vital role in various clinical populations and, thus, are commonly assessed across settings. In this respect, the Beck Depression Inventory – II (BDI-II, Beck et al., 1996) is one of the most widely accepted and commonly applied self-report measures for depressive symptoms. At first, an earlier version of the BDI-II was introduced by Beck et al. (1961), and, after revision, resulted in the BDI-II in a later version (Beck et al., 1996). Since its first introduction, the BDI has been popular for the assessment of depressive symptoms across settings, including psychiatric patients (Camara et al., 2000; Piotrowski, 1996; Steer et al., 1999, 2000) but also non-clinical populations

CONTACT Johann Lehrner johann.lehrner@meduniwien.ac.at Department of Neurology, Medical University of Vienna, Vienna, Austria.

(Steer et al., 1986). Further, a survey among clinical neuropsychologists engaging in forensic assessments found that the BDI is one of the most prominent and commonly used self-report symptom measures in forensic neuropsychology (LaDuke et al., 2018). Among 77 board certified forensic psychologists, 49 participants (64%) endorsed frequent use of the BDI-II for the assessment of mood.

Given that the BDI-II is a commonly used measure it is a weakness that it contains no assessment of symptom validity. For example, Lees-Haley (1989) instructed 52 undergraduate psychology students to complete a previous version of the BDI under simulated conditions. In this study, students were instructed to imagine that they were exposed to a toxic substance from a hazardous waste site, and to complete the BDI as if they suffered from psychological reactions. Of those 52 individuals, 96% successfully feigned depression and 58% were able to feign extreme levels of depression on the BDI. Further, it has been found that the most severe levels of genuine depression are reflected by scores of 40 (or 50), and any BDI-II scores larger than 40 may indicate symptom exaggeration (Groth-Marnat & Wright, 2016, Merten et al., 2020). Merten et al. (2020) did further examinations on this cut score and demonstrated clear association between exceeding a cut score of 40 on the BDI-II and failing two SVTs in a large sample of 537 psychosomatic rehabilitation patients. Of those 50 individuals who obtained a BDI-II score above 40, 84% were positive on either the SIMS (Structured Inventory of Malingered Symptomatology, Smith & Burger, 1997), the SRSI (Self-Report Symptom Inventory, Merten et al., 2019), or both. Moreover, of those 339 individuals who scored below the recommended cut scores on both SRSI and SIMS, the vast majority (98%) did not exceed a BDI-II score of 40, giving strong evidence for the association between high BDI-II scores and invalid symptom report in this referral context.

The present report aims to provide BDI-II cut scores for symptom validity testing in forensic evaluations. An archival data set of early retirement claimants presented for forensic assessment are considered for this purpose. The majority of the sample was diagnosed with a mental or behavioral disorder (ICD-10 F diagnoses), including mood disorders in a sizeable proportion of patients. As the primary objective, we examine the utility of the BDI-II to distinguish between valid and invalid symptom report. In this analysis, we chose a conservative criterion for invalid symptom report by positive results on two independent freestanding SVTs. Further, secondary analysis addresses whether different cut scores should be suggested based on level of school education and vocabulary skills, as well as the effect of the chosen criterion for invalid symptom report on the BDI-II's diagnostic utility. The results may provide clinicians an additional source of embedded validity testing based on an already existing and widely distributed clinical measure.

Methods

The objectives of the present study are examined based on archival data reported in earlier research on neuropsychological performance tests as embedded validity indicators (Czornik et al., 2022; Fuermaier et al., 2023). As the procedure and methodology of the assessment protocol is described in detail elsewhere, the present report provides only a brief description that is relevant for the comprehension of this study. Please refer to Czornik et al. (2022), and Fuermaier et al. (2023) for more detail and comprehensive background. The study protocol was approved by the Ethical Committee of the Medical University of Vienna and was conducted in accordance with the Declaration of Helsinki (registry number 2231/2020).

Participants and procedure

The sample contained data of patients presented to private neuropsychological office in Vienna, Austria. All patients were referred by a general court or a pension insurance agency for a neuropsychological assessment because of claimed early retirement due to significant cognitive impairment. Assessments were conducted in German language. All participants were fluent in German. A systematic report of the diagnostic status of individuals cannot be provided, because diagnoses are not communicated to the examiners on a regular basis in the context of Austrian pension insurance referrals. The vast majority of patients had a confirmed ICD-10 chapter *F* (mental and behavioral disorders) diagnosis, a probable *F* diagnosis, or the claim of an *F* diagnosis. Among those a variety of different diagnoses are likely present, including depression (about one third), adjustment disorder and chronic fatigue (about one fourth), anxiety (about one tenth), somatoform disorders (in about 5% of cases), and a smaller proportion of diverse diagnoses including substance abuse. Psychopathological syndromes of psychoses, delusions, confusional states, amnestic syndromes, or dementia, were in a small minority.

Two hundred and twenty individuals were considered for inclusion in this study. Participants with severe memory impairment ($n = 3$, due to possible dementia or intellectual disability) were excluded based on imaging data, clinical judgment and performance on the Word Memory Test (WMT, Green, 2003), leaving a sample of 217 individuals. The WMT is a stand-alone memory based PVT, that includes profile scores (i.e. difference between easy and hard items) that may indicate the presence of genuine severe cognitive impairment. Performance validity data (based on WMT performance) are not considered in this study as this topic goes beyond the scope of this research and has been presented in earlier work on a largely overlapping sample of the same referral context (see Czornik et al., 2022; Fuermaier et al., 2023). Symptom validity was assessed with two independent freestanding SVTs (i.e., SIMS and SRSI, see material section for details). Derived from international consensus and current practice standards in performance validity testing (Jennette et al., 2022; Rhoads et al., 2021; Schroeder et al., 2019; Soble et al., 2020; Sweet et al., 2021), we chose a conservative criterion for invalid symptom report for our primary analysis, i.e., if patients scored above the recommended cut scores on both SVTs. Symptom reports of

Table 1. Descriptive information and group comparisons of individuals with valid and invalid symptom report.

Construct	Total ($N = 217$)	Valid ($n = 150$)	Invalid ($n = 67$)	t/X^2	p	Cohen's d	d – 95%-CI
Age (in years)	47.8 + 9.7	48.8 + 9.5	45.7 + 9.9	2.170	0.031	0.32	0.03–0.61
Sex (f/m)	113/104	75/75	38/29	0.837	0.360		
Education (years)	11.3 + 3.7	12.0 + 4.1	9.9 + 2.0	3.944	<.001	0.58	0.29–0.88
Vocabulary skills	98.0 ± 14.2	102.3 ± 12.9	88.3 ± 12.1	7.491	<.001	1.10	0.79–1.40
Depression (BDI-II)	27.9 ± 13.1	23.5 ± 11.6	37.8 ± 10.7	8.581	<.001	1.26	0.95–1.57

Note. Education is indicated in years of school education including university/college (not reported by one individual); Vocabulary skills is assessed with the WTS-IQ on an IQ scale; BDI-II: Beck Depression Inventory – II.

those patients who passed at least one of the two validity tests were considered valid. Table 1 presents BDI-II scores as well as descriptive information. Mean age for the entire group was 47.8 years ($SD = 9.7$), ranging from 19 to 64 years. School education had a mean of 11.3 years ($SD = 3.7$), ranging 8–23 years, whereas vocabulary skills ranged from 71 to 133, with a mean of 98.0 points ($SD = 14.2$). Table 1 further presents descriptive information of the group passing or failing symptom validity assessment (SVA).

Materials

Descriptive information

Descriptive information (i.e. age, gender, years of schooling) was obtained from all individuals, including a short assessment of vocabulary skills with the WST-IQ (Schmidt & Metzler, 1992). The WST-IQ is a vocabulary recognition test consisting of 40 items (one target and 5 distractors per item), and depicts the vocabulary skills of the participant on an IQ-scale.

Symptom validity

The *Structured Inventory of Malingered Symptomatology* (SIMS, Smith & Burger, 1997) is a questionnaire consisting of 75 dichotomous Yes/No-items. The total scale with a maximum score of 75 points can be divided in five subscales, i.e. neurological impairment, amnestic disorders, psychosis, low intelligence, and affective disorders, each with a maximum scores of 15 points. The items relate to atypical, extreme, or bizarre symptoms that appear to correspond to broad psychopathological domains. For comprehensive reviews and meta-analyses of the SIMS please refer to van Impelen et al. (2014) and Shura et al. (2022). In the current study, we applied the recommended cut score of >23 for use in forensic situations with multiple SVTs (Shura et al., 2022; van Impelen et al., 2014).

The *Self-Report Symptom Inventory* (SRSI, Merten et al., 2016, 2019, 2022) is an SVT with 107 dichotomous True/False-items. Items can be subdivided into five subscales of potentially genuine symptoms (cognitive symptoms, depressive symptoms, pain symptoms, nonspecific somatic symptoms, and posttraumatic stress disorder/anxiety symptoms) and five subscales of pseudosymptoms describing atypical, bizarre or extreme symptom claims (cognitive, motor neurological, sensory neurological, pain, and mental pseudosymptoms). The SRSI yields ten subscale scores, a total genuine symptoms score, and a total pseudosymptoms score. The manual of the SRSI reports an extensive body research supporting its validation. Test-retest reliability is good for the pseudosymptom scale (0.91). Internal consistency (Cronbach's alpha) was reported to be excellent (ranging from 0.92 to 0.95). In this study, we applied the standard use cut score > 9 on the pseudosymptom scale (Specificity = 0.96; Sensitivity = 0.62; Merten et al., 2022).

Depressive symptoms (BDI-II)

The *Beck Depression Inventory–II* (BDI-II; Beck et al., 1996, 2006) contains 21 items on the experiences of depressive symptoms within the last two weeks. Each item is rated on a four-point scale (0–3), which are summed up to obtain the total score (0–63). Symptom self-reports allow a classification of severity of depressive symptoms, ranging from minimal to severe. Internal consistency (Cronbach's alpha) of the BDI-II was reported to be excellent (0.92–0.93; Beck et al., 2006). For use in psychiatric patients, any BDI-II score equal or higher than 19 indicates clinically relevant depressive symptoms (von Glischinski et al., 2019).

Statistical analysis

In the primary analysis, descriptive information and BDI-II scores were presented for the total group and separately for individuals passing and failing SVA. Groups were compared by statistical tests (indicating presence of effects) and calculating effect sizes (indicating magnitude of findings). Further, we computed a receiver operating characteristic (ROC) analysis to determine the utility of the BDI-II to differentiate valid from invalid symptom report. A classification table will be given (Table 2) to present sensitivity and specificity for a range of BDI-II cut scores. Positive predictive power and negative predictive power will be given for various hypothetical base rates, ranging from 15 to 60%.

In the secondary analysis, we explored the effects of school education, vocabulary skills, and SVT failure criterion. Classification statistics per BDI-II cut scores to detect invalid symptom report will be presented separately for individuals below or above average education (groups determined by median split) and for individuals below or above average vocabulary skills (split by the normative mean of 100). Further, ROC analysis and classification statistics as computed for the primary analysis will be repeated for various SVT failure criteria, that is, positive results on (1) the SIMS, (2) the SRSI, and (3) at least one of the two SVTs.

Table 2. Classification accuracy of BDI-II scores for various cut scores and hypothetical base rates of invalid symptom report.

	Sensitivity	Specificity	BR 15%		20%		30%		40%		50%		60%	
			PPV	NPV	PPV	NPV	PPV	NPV	PPV	NPV	PPV	NPV	PPV	NPV
BDI-II ≥														
30	0.82	0.65	0.29	0.95	0.37	0.94	0.50	0.89	0.61	0.84	0.70	0.78	0.78	0.71
34	0.70	0.80	0.38	0.94	0.47	0.91	0.60	0.86	0.70	0.80	0.78	0.73	0.84	0.64
36	0.60	0.86	0.43	0.92	0.52	0.90	0.65	0.83	0.74	0.76	0.81	0.68	0.87	0.59
37	0.58	0.89	0.48	0.92	0.57	0.90	0.69	0.83	0.78	0.76	0.84	0.68	0.89	0.59
38	**0.58**	**0.90**	0.51	0.92	0.59	0.90	0.71	0.83	0.80	0.76	0.85	0.68	0.90	0.59
39	0.52	0.92	0.53	0.92	0.62	0.89	0.74	0.82	0.81	0.74	0.87	0.66	0.91	0.56
40	0.51	0.95	0.64	0.92	0.72	0.89	0.81	0.82	0.87	0.74	0.91	0.66	0.94	0.56
45	0.28	0.98	0.71	0.89	0.78	0.84	0.86	0.76	0.90	0.67	0.93	0.58	0.96	0.48

Note. BR: base rate; *PPV*: positive predictive value; NPV: negative predictive value; in bold: cut score with specificity = 90%.

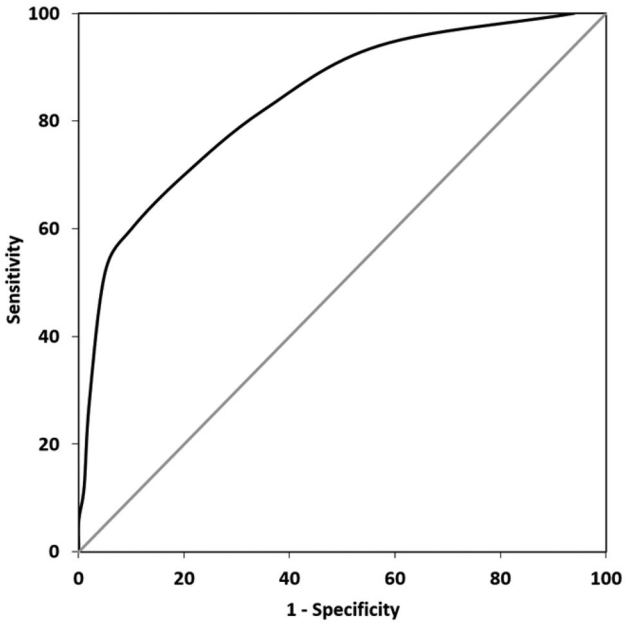

Figure 1. Receiver operating characteristic (ROC) curve indicating diagnostic accuracy of BDI-II scores in distinguishing individuals with valid ($n = 150$) from invalid ($n = 67$) symptom report (two SVT failure criterion).

Results

Primary analysis

Of the consecutive sample of 217 individuals entering data analysis, 150 (69.1%) were interpreted as valid based on previously described cut scores based on stand alone SVT performance and 67 (30.9%) were invalid. Descriptive information and BDI-II scores of all individuals are presented in Table 1. Compared to the group with valid symptom report, the group with invalid symptom report was significantly younger (small effect), less educated (medium effect), obtained a lower score on vocabulary skills (large effect) and indicated higher levels of depressive symptoms on the BDI-II (large effect). ROC analysis demonstrated significant classification accuracy of the BDI-II in distinguishing valid from invalid symptom report, i.e., AUC = 0.822, $SE = 0.032$, $p < .001$, 95%-CI = 0.760–0.884. Figure 1 presents a graphical depiction of the classification accuracy derived from the ROC analysis. Classification accuracies in terms of sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV), are presented in Table 2.

Secondary analysis

Effects of school education and vocabulary skills on the choice of BDI-II cut scores for indicating invalid symptom report are presented in Table 3. For individuals with below average school education (years of schooling), higher levels of sensitivity can be reached based on a desired specificity of at least 90%. An effect in opposite direction emerged for education level as determined by the test for vocabulary skills. For individuals with ≤9 years of education, a BDI-II cut score ≥38 had the highest sensitivity (0.59) while maintaining a specificity of 0.90. For individuals with > 9 years of education, a BDI-II cut score of ≥39 resulted in a sensitivity of 0.45, with a specificity of 0.91. For individuals with below average vocabulary skills (≤100), a BDI-II cut score of ≥39 resulted in a sensitivity of 0.49 while maintaining a specificity of 0.91. Finally, for individuals with above average vocabulary skills (>100), it is a BDI-II cut score ≥36 that corresponds to a sensitivity of 0.75 and a specificity of .90. Although results are exploratory in nature, it could be hypothesized that lower BDI-II cut scores can be recommended for individuals with lower formal education and higher vocabulary skills to reach higher levels of sensitivity while keeping chances of false positives below or at 10%.

Moreover, Table 4 depicts the BDI-II's utility to detect invalid symptom report for alternative SVT failure criteria, that is, failing the SRSI, SIMS, at least one of the two SVTs, and, as a comparison to this additional analysis, failing both SIMS and SRSI as it is presented in the primary analysis. Higher SVT failure rates are observed for the SRSI (48.9%) compared to the SIMS (36.4%), and, as expected, the largest failure rate is observed if invalid symptom report is defined by failing at least one of the two SVTs (54.4%). The BDI-II's classification accuracy does not vary largely depending the SVT failure criterion, with AUC's ranging from 80.8% (SIMS) to 83.1% (≥1 SVT failure). BDI-II cut scores reaching the often desired 90% specificity level varied between ≥34 (for ≥1 SVT failure) and ≥38 (for SIMS, as well as 2 SVT failures).

Discussion

In this report, we evaluated the BDI-II as an embedded validity indicator of symptom self-reports in forensic evaluations. Previous research suggested scores above 40 may

Table 3. Classification accuracy of the BDI-II to detect invalid symptom report (i.e., failing two SVTs), per education level and vocabulary skills.

	Education <9 years ($n = 121$)		Education >9 years ($n = 95$)		Vocabulary skills <100 ($n = 122$)		Vocabulary skills >100 ($n = 95$)	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
BDI-II ≥								
30	0.85	0.64	0.75	0.65	0.82	0.56	0.81	0.72
34	0.74	0.77	0.60	0.83	0.69	0.73	0.75	0.86
36	0.61	0.85	0.55	0.87	0.55	0.82	**0.75**	**0.90**
37	0.59	0.88	0.55	0.89	0.53	0.84	0.75	0.92
38	**0.59**	**0.91**	0.55	0.89	0.53	0.87	0.75	0.92
39	0.54	0.93	**0.45**	**0.91**	**0.49**	**0.91**	0.63	0.92
40	0.54	0.95	0.40	0.95	0.47	0.94	0.63	0.95
45	0.39	0.96	0.05	1.00	0.29	0.93	0.25	0.99

Note. Years of education is missing for one individual. Recommended cut scores resulting in at least 90% specificity are indicated in bold.

indicate symptom exaggeration (Groth-Marnat & Wright, 2016, Merten et al., 2020). The present study comports with previous findings and adds evidence for its use in forensic evaluations with a cut score of ≥38 achieving a specificity level of 0.90 (Boone, 2021; Larrabee, 2008) with a sensitivity of 0.58. With a cut score of ≥40, specificity rises to 0.95, while sensitivity reaches still 0.51.

In the light of the chosen criterion for invalid symptom report of both SIMS and SRSI failure, extreme scores on the BDI-II in forensic populations seem to be associated with the endorsement not only of pronounced levels of genuine symptoms, but also with the endorsement of bizarre, extreme, or rarely occurring symptoms (see conceptualization of the SIRS and SIMS as freestanding SVTs). Due to the popularity of the BDI-II across settings, the findings of this study give clinicians in forensic settings a ready-to-apply new opportunity in their repertoire to evaluate whether the symptom report given by an examinee can be trusted and whether claimed symptomatology is credible. Of note, individuals failing SVA in our study differed from individuals passing SVA not only in BD-II scores, but were also slightly younger (small effect), less educated (medium effect) and showed lower vocabulary skills (large effect). In secondary and exploratory analysis, we underscored the relevance of these effects by differentiating between levels of school education and vocabulary skills. Although results are subject for independent replication on larger samples, our additional analysis tentatively suggests education specific BDI-II cut scores. We observed slightly lower optimal BDI-II cut scores for individuals with lower formal education and higher vocabulary skills in order to reach high levels of sensitivity while keeping chances for false positives low (e.g., often desired below or at 10%). The effects of vocabulary skills and school education in opposite directions may be support for the assumption of nonequivalence of test performance and real life functioning, however, future hypothesis driven studies need to address this effect more thoroughly. In this context, it must be considered that the present findings of school education and vocabulary skills do not reflect causal relationships and do not necessarily mean that cut scores should be adjusted based on school education and/or vocabulary skills. Potential third variables, e.g. a careless response style, or compliance to instructions, may affect both test performance (including the test for vocabulary skills) and symptom reports (on all measures, including SVTs and the BDI-II), and may thus contribute to the observed effects. Further, secondary analysis demonstrated that the BDI-II's classification accuracy is largely robust against the chosen criterion of invalid symptom report, as evidenced by similar *Area Under the Curves* for the various test criteria (i.e., positive results on SIRS, SIMS, ≥1 SVT, or 2 SVTs). The recommended BDI-II cut score, however, may differ per chosen SVT failure criterion. For example, a lower cut score of ≥34 is indicated based on the SVT failure criterion of at least one SVT, assuming a specificity of at least 90% is envisaged.

The present study on individuals from a forensic context suggests BDI-II cut scores in a similar range than the ones

Table 4. Classification accuracy of the BDI-II to detect invalid symptom report per SVT failure criterion.

AUC (*p*)	SRSI (pass/fail =111/106) 0.818 (<.001)		SIMS (pass/fail = 138/79) 0.808 (<.001)		≥1 SVT (pass/fail = 99/118) 0.831 (<.001)		2 SVTs (pass/fail =150/67) 0.822 (<.001)	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
BDI-II ≥								
30	0.73	0.72	0.79	0.67	0.71	0.76	0.82	0.65
34	0.59	0.87	0.66	0.82	**0.58**	**0.91**	0.70	0.80
36	**0.48**	**0.91**	0.54	0.87	0.46	0.93	0.60	0.86
37	0.46	0.94	**0.53**	**0.90**	0.44	0.94	0.58	0.89
38	0.44	0.94	0.53	0.91	0.42	0.94	**0.58**	**0.90**
39	0.40	0.95	0.47	0.93	0.37	0.97	0.52	0.92
40	0.38	0.98	0.44	0.95	0.35	0.99	0.51	0.95
45	0.20	0.99	0.25	1.00	0.19	1.00	0.28	0.98

Note. SRSI = Self-Report Symptom Inventory; SIMS = Structured Inventory of Malingered Symptomatology.
Recommended cut scores resulting in at least 90% specificity are indicated in bold.

discussed by Merten et al. (2020) on a psychosomatic rehabilitation sample. A comparison of both studies finds several striking similarities, for example the use of SRSI and SIMS as SVTs, the assessment in a German speaking environment, roughly equal distribution of gender, age range from 19 to mid/end sixty, and a similar composition of diagnostic groups, including depression, adjustment disorder, chronic fatigue, anxiety, and somatoform disorders. However, it must be noted that it was not the aim of the Merten et al. (2020) study to determine optimal BDI-II cut scores for the differentiation between valid and invalid symptom report. Instead, the earlier proposed BDI-II score of 40 was explored regarding its association with positive SRSI and/or SIMS results. Thus, a comparison of a range of BDI-II cut scores regarding their classification accuracy was not given, and the choice of an optimal cut score cannot be concluded from the study of Merten et al. (2020).

A strength of the present study is the consideration and comparison of different SVA failure criteria based on SIMS and SRSI results. The SIMS is presumably the most widely used SVT in forensic practice (Dandachi-FitzGerald et al., 2013; Martin et al., 2015). The SRSI, in turn, was developed recently for a setting that resembles closely the one of the present study; i.e., the detection of invalid symptom report of individuals presenting with 'soft' psychopathology in forensic and clinical context.

As a limitation to the generalization of our findings, one must note that we examined the BDI-II as an embedded validity indicator in a specific sample of early retirement claimants undergoing forensic neuropsychological assessment. While these findings are valuable and add to the empirical knowledge base, they should not be extrapolated to other populations or settings. Further studies with a more diverse range of conditions and settings are needed to fully assess the diagnostic quality of the BDI-II as a validity indicator. Further, it must be considered that embedded validity indicators are more sensitive to genuine psychopathology compared to freestanding SVTs, and, thus, bear the risk of confusing genuine psychopathology with invalid symptom report (Erdodi & Lichtenstein, 2017). For this reason, aggregating multiple indicators of symptom validity, each with sufficiently high levels of specificity, is of great relevance especially in the use of embedded validity indicators (e.g., Erdodi & Lichtenstein, 2017). Aggregating multiple symptom validity indicators is also relevant in the light of rather low sensitivity of the BDI-II, which does not justify the use of the BDI-II as a sole measure of symptom validity. Although the likelihood of false positives is low with commonly claimed specificity levels of at least 90%, the use of multiple independent SVTs could reduce the likelihood of false negatives. Further, additional sources of information, such as positive results on PVTs, presence of external incentives, or marked discrepancy, would allow a thorough investigation of the underlying motivation for invalid data (e.g., for malingered neurocognitive dysfunction, see Sherman et al., 2020).

Funding

The author(s) reported there is no funding associated with the work featured in this article.

ORCID

Anselm B. M. Fuermaier http://orcid.org/0000-0002-2331-0840
Brechje Dandachi-Fitzgerald http://orcid.org/0000-0002-8984-8192
Johann Lehrner http://orcid.org/0000-0001-8270-9272

References

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II*. Psychological Corporation.

Beck, A. T., Steer, R. A., & Brown, G. K. (2006). *BDI–II. Beck-Depressions-Inventar manual* (2nd edition). Harcourt Test Services.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*, 561–571. https://doi.org/10.1001/archpsyc.1961.01710120031004

Boone, K. B. (2021). *Assessment of feigned cognitive impairment: A neuropsychological perspective*. Guilford Press.

Bush, S. S., Heilbronner, R. L., & Ruff, R. M. (2014). Psychological assessment of symptom and performance validity, response bias, and malingering: Official position of the Association for Scientific Advancement in Psychological Injury and Law. *Psychological Injury and Law*, *7*(3), 197–205. https://doi.org/10.1007/s12207-014-9198-7.

Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology*, *31*(2), 141–154. https://doi.org/10.1037/0735-7028.31.2.141

Carone, D. A., & Bush, S. S. (2018). *Validity assessment in rehabilitation psychology and settings*. Oxford University Press.

Czornik, M., Seidl, D., Tavakoli, S., Merten, T., & Lehrner, J. (2022). Motor reaction times as an embedded measure of performance validity: A study with a sample of Austrian early retirement claimants. *Psychological Injury and Law*, *15*(2), 200–212. https://doi.org/10.1007/s12207-021-09431-z

Dandachi-FitzGerald, B., Ponds, R., & Merten, T. (2013). Symptom validity and neuropsychological assessment: A survey of practices and beliefs of neuropsychologists in six European countries. *Archives of Clinical Neuropsychology*, *28*(8), 771–783. https://doi.org/10.1093/arclin/act073.

Erdodi, L. A., & Lichtenstein, J. D. (2017). Invalid before impaired: An emerging paradox of embedded validity indicators. *The Clinical Neuropsychologist*, *31*(6-7), 1029–1046. https://doi.org/10.1080/13854046.2017.1323119.

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

Green, P. (2003). *Green's Word Memory Test. User's manual*. Green's Publishing.

Groth-Marnat, G., & Wright, J. (2016). *Handbook of psychological assessment*. John Wiley & Sons.

Jennette, K. J., Williams, C. P., Resch, Z. J., Ovsiew, G. P., Durkin, N. M., O'Rourke, J. J. F., Marceaux, J. C., Critchfield, E. A., & Soble, J. R. (2022). Assessment of differential neurocognitive performance based on the number of performance validity tests failures: A cross-validation study across multiple mixed clinical samples. *The Clinical Neuropsychologist*, *36*(7), 1915–1932. https://doi.org/10.1080/138540462021.1900398.

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

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

Lees-Haley, P. R. (1989). Malingering traumatic mental disorder on the Beck Depression Inventory: Cancerphobia and toxic exposure. *Psychological Reports*, *65*(2), 623–626. https://doi.org/10.2466/pr0.1989.65.2.623.

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

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

Merten, T., Dandachi-FitzGerald, B., Boskovic, I., Puente-López, E., & Merckelbach, H. (2022). The Self-Report Symptom Inventory. *Psychological Injury and Law*, *15*(1), 94–103. https://doi.org/10.1007/s12207-021-09434-w

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

Merten, T., Kaminski, A., & Pfeiffer, W. (2020). Prevalence of overreporting on symptom validity tests in a large sample of psychosomatic rehabilitation inpatients. *The Clinical Neuropsychologist*, *34*(5), 1004–1024. https://doi.org/10.1080/13854046.2019.1694073.

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

Piotrowski, C. (1996). Use of the Beck Depression Inventory in clinical practice. *Psychological Reports*, *79*(3 Pt 1), 873–874. 10.2466/pr0.1996.79.3.873.

Rhoads, T., Neale, A. C., Resch, Z. J., Cohen, C. D., Keezer, R. D., Cerny, B. M., Jennette, K. J., Ovsiew, G. P., & Soble, J. R. (2021). J.R. (2021). Psychometric implications of failure on one performance validity test: A cross-validation study to inform criterion group definition. *Journal of Clinical and Experimental Neuropsychology*, *43*(5), 437–448. https://doi.org/10.1080/13803395.2021.1945540.

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

Schmidt, K.-H., & Metzler, P. (1992). *Wortschatztest WST*. Beltz.

Schroeder, R. W., Martin, P. K., Heinrichs, R. J., & Baade, L. E. (2019). Research methods in performance validity testing studies: Criterion grouping approach impacts study outcomes. *The Clinical Neuropsychologist*, *33*(3), 466–477. https://doi.org/10.1080/13854 046.2018.1484517.

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

Shura, R. D., Ord, A. S., & Worthen, M. D. (2022). Structured Inventory of malingered symptomatology: A psychometric review. *Psychological Injury and Law*, *15*(1), 64–78. https://doi.org/10.1007/s12207-021-09432-y

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

Soble, J. R. (2021). Future directions in performance validity assessment to optimize detection of invalid neuropsychological test performance: Special issue introduction. *Psychological Injury and Law*, *14*(4), 227–231. https://doi.org/10.1007/s12207-021-09425-x.

Soble, J. R., Alverson, W. A., Phillips, J. I., Critchfield, E. A., Fullen, C., O'Rourke, J. J. F., Messerly, J., Highsmith, J. M., Bailey, K. C., Webber, T. A., & Marceaux, J. C. (2020). Strength in numbers or quality over quantity? Examining the importance of criterion measure selection to define validity groups in performance validity test (PVT) research. *Psychological Injury and Law*, *13*(1), 44–56. https://doi.org/10.1007/s12207-019-09370-w

Steer, R. A., Ball, R., Ranieri, W. F., & Beck, A. T. (1999). Dimensions of the Beck Depression Inventory–II in clinically depressed outpatients. *Journal of Clinical Psychology*, *55*(1), 117–128. https://doi.org/10.1002/(sici)1097-4679(199901)55:1<117::aid-jclp12>3.0.co;2-a

Steer, R. A., Beck, A. T., & Garrison, B. (1986). Applications of the Beck Depression Inventory. In N. Sartorius & T.A. Ban (Eds.), *Assessment of depression* (pp. 121–142). World Health Organization.

Steer, R. A., Rissmiller, D. J., & Beck, A. T. (2000). Use of Beck Depression Inventory-II with depressed geriatric inpatients. *Behaviour Research and Therapy*, *38*(3), 311–318. https://doi.org/10.1016/S0005-7967(99)00068-6

Sweet, J. J., Heilbronner, R. L., Morgan, J. E., Larrabee, G. J., Rohling, M. L., Boone, K. B., Kirkwood, M. W., Schroeder, R. W., & Suhr, J. A. & Conference Participants. (2021). American Academy of Clinical Neuropsychology (AACN) 2021 consensus statement on validity assessment: Update of the 2009 AACN consensus conference statement on neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist*, *35*(6), 1053–1106. https://doi.org/10.1080/13854046.2021.1896036.

van 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 Glischinski, M., von Brachel, R., & Hirschfeld, G. (2019). How depressed is "depressed"? A systematic review and diagnostic meta-analysis of optimal cut points for the Beck Depression Inventory revised (BDI-II). *Quality of Life Research*, *28*(5), 1111–1118. https://doi.org/10.1007/s11136-018-2050-x.