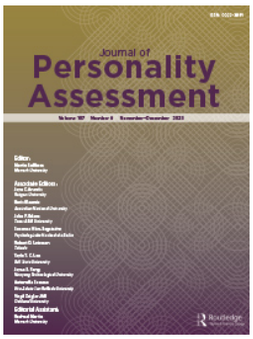

Journal of Personality Assessment

ISSN: 0022-3891 (Print) 1532-7752 (Online) Journal homepage: www.tandfonline.com/journals/hjpa20

A Population Wide Analysis of MCMI-IV Symptom Validity Scales Administered in the Veterans Affairs Healthcare System

Robert D. Shura, Jordan V. Rine, Paul B. Ingram, Ryan W. Schroeder, Patrick Armistead-Jehle & Luciano Giromini

To cite this article: Robert D. Shura, Jordan V. Rine, Paul B. Ingram, Ryan W. Schroeder, Patrick Armistead-Jehle & Luciano Giromini (03 Dec 2025): A Population Wide Analysis of MCMI-IV Symptom Validity Scales Administered in the Veterans Affairs Healthcare System, Journal of Personality Assessment, DOI: 10.1080/00223891.2025.2592958

To link to this article: https://doi.org/10.1080/00223891.2025.2592958

Published online: 03 Dec 2025.

JOURNAL OF PERSONALITY ASSESSMENT
https://doi.org/10.1080/00223891.2025.2592958

A Population Wide Analysis of MCMI-IV Symptom Validity Scales Administered in the Veterans Affairs Healthcare System

Robert D. Shura[1,2,3], Jordan V. Rine[1,3], Paul B. Ingram[4], Ryan W. Schroeder[5], Patrick Armistead-Jehle[6] and Luciano Giromini[7]

[1]Salisbury VA Health Care System, Salisbury, North Carolina; [2]VA Mid-Atlantic Mental Illness Research, Education, and Clinical Center, Durham, North Carolina; [3]Department of Neurology, Wake Forest School of Medicine, Winston-Salem, North Carolina; [4]Department of Psychological Sciences, Texas Tech University, Lubbock, Texas; [5]Robert J. Dole VA Medical Center, Wichita, Kansas; [6]Munson Army Health Center, Fort Leavenworth, Kansas; [7]Department of Psychology, University of Turin, Turin, Italy

ABSTRACT

The Millon Clinical Multiaxial Inventory-IV (MCMI-IV) is a psychological assessment tool commonly used in Veteran Affairs (VA) settings. However, no research has examined the MCMI-IV symptom validity scales in the veteran population, where high rates of response bias can occur. This study examined convergent validity of the MCMI-IV scales to the Minnesota Multiphasic Personality Inventory-2-Restructured Form (MMPI-2-RF) validity measures, identified base rates of invalid MCMI-IV validity scores in the veteran population, and explored alternative cutoff scores. All MCMI-IVs administered in the VA (04/2019−08/2024, N=6,466), using data from the Corporate Data Warehouse, were examined. MCMI-IV protocols were correlated with same day administrations of the MMPI-2-RF (n=1,401) using Spearman's correlations. Disclosure and Debasement positively correlated with overreporting validity scales and negatively correlated with underreporting scales on the MMPI-2-RF, while Desirability showed the opposite pattern (all $p < .001$). Additionally, the Inconsistency scale positively correlated with MMPI-2-RF non-content validity scales. Less than 1% of MCMI-IV of protocols met the test manual's criteria for invalidity, a significant departure from invalid rates reported on other measures administered to veterans. Diagnostic accuracy analyses suggested new cutoff scores, specifically that overreporting be identified by X Disclosure base rate score (BRS) ≥ 87 and Z Debasement BRS ≥ 84, and underreporting identified by X Disclosure BRS ≤ 49 and Y Desirability BRS ≥ 74. Results suggest that while the MCMI-IV validity indices measure intended constructs, more optimal cutoffs are presented for the veteran population.

ARTICLE HISTORY
Received 15 April 2025
Revised 7 November 2025
Accepted 9 November 2025

Psychologists have used muti-scale self-report inventories steadily throughout history as a staple of assessment. Based on one practice survey, the Minnesota Multiphasic Personality Inventory (MMPI) was the most frequently used (60.0% of respondents), followed by the Personality Assessment Inventory (PAI; 34.6% of respondents), and the Millon Clinical Multiaxial Inventory (MCMI; 13.3% of respondents) (Rabin et al., 2016). Although there are numerous differences across these three common broadband measures, a core aspect of interpretation across all of these instruments is that such begins by evaluating symptom validity scales measuring consistency of response and likelihood of response bias. It is of note, though, that the MCMI validity scales are employed far less than their MMPI and PAI counterparts, with the MCMI validity scales being used by only about 13% of North American neuropsychologists and validity assessment experts (Martin et al., 2025). Further, a recent review of the MCMI-IV validity scales in a special issue of *Psychological Injury and Law* resorted to using MCMI-III studies as proxies due to lack of research on the newer version of the MCMI (Choca & Pignolo, 2022), highlighting a need for research on these scales specifically.

The MCMI-IV contains five validity scales across non-content responding, overreporting, and underreporting. Two scales are non-content scales designed to detect lack of attending to item content or random response (Millon et al., 2015). V Invalidity contains items that are extremely unlikely to be endorsed, with a cutoff of raw ≥ 2 used to identify invalid protocols, and a score of 1 considered indeterminant. This measure (originally called Validity instead of Invalidity) was the only validity measure on the original MCMI and initially contained four items (Choca & Van Denburg, 1997). On the MCMI-IV, the scale contains three items, which do not overlap with any other scale. V Invalidity's minimal item count has been a historical weakness, with research on the MCMI-III demonstrating that it failed to identify nearly 50% of random responders, and based on probability theory, increasing the number of items on this scale would drastically reduce the number of random responders missed (Charter & Lopez, 2002). The introduction of the W scale in the MCMI-IV appears to be a direct response to this limitation. W Inconsistency, which is new to the MCMI-IV, is created from 25 pairs of items with the highest correlations to each other (coefficients are not listed in the manual).

CONTACT Robert D. Shura Robert.Shura2@va.gov W. G. (Bill) Salisbury VA Health Care System, 1601 Brenner Ave (11-M), Salisbury, NC 28144, USA.

The manual indicates that a cutoff score of raw ≥ 20 render a protocol invalid, though scores from 9 to 19 can be considered indeterminant. There is overlap present across W Inconsistency and other MCMI-IV scales.

The primary overreporting scale of the MCMI-IV is X Disclosure, which was formally added to the MCMI-II and altered over the course of versions. This scale is made up of 121 out of the 195 items on the MCMI-IV, and a raw score > 114 invalidates the protocol. However, as a Modifying Index, if X Disclosure is moderately elevated, it is used to adjust elevations on the 12 primary clinical personality scales (Millon et al., 2015). Given this scale contains 62% of all items on the MCMI-IV, it highly overlaps with other scales, and could be considered a measure of broad exaggerated psychopathology (i.e., an individual endorsed an extreme number of symptoms across a variety of symptom types). The second overreporting scale is Z Debasement, which was originally added to the MCMI-II (Choca & Van Denburg, 1997). The scale was reworked for the MCMI-IV and now contains 30 items, all of which are contained in at least one other scale. The manual vaguely describes the creation of the scale as involving individual item evaluation to identify items that reflected an overwhelmingly negative presentation. A BRS of > 74 reflects a mild elevation and a BRS > 84 a more significant elevation, but neither are indicated to invalidate the protocol formally, with increasing scores indicating increasing likelihood of negative response bias.

For underreporting, a low raw score < 7 on X Disclosure invalidates the protocol due to underreporting (Millon et al., 2015). This low of a score reflects that the respondent is presenting as free from psychopathology. Although not discussed in the manual, conceptually the possibility exists for those who are generally well adjusted to score in this range. Finally, Y Desirability is the kin to Z Debasement and was reworked similarly for the MCMI-IV; the current scale was created from 24 items identified as reflecting an overwhelmingly positive presentation. Like Z Debasement, no item is unique to this scale.

A primary and longstanding criticism of the MCMI's validity scales (except for V Invalidity) has been their extensive item overlap with the clinical personality and syndrome scales, a factor that can artificially inflate correlations and conflate genuine psychopathology with response style. Furthermore, scales like Y Desirability and Z Debasement have been historically questioned for their limited utility in detecting sophisticated response bias (Groth-Marnat, 2003). Finally, the X Disclosure scale has been shown to lead to false negatives for overreporting and underreporting in prior MCMI versions (Groth-Marnat & Wright, 2016). A key question for the MCMI-IV is whether its revisions have successfully mitigated these known issues.

Millon used base rate scores (BRSs; *not base rates*) to account for personality and mood disorder prevalence rate heterogeneity on the MCMI scales. These BRS are distinct from the more common T-score based scales employed by other broadband measures, such as the MMPI and PAI. In contrast to T-score-based approaches, which presume that a target construct is normally distributed, BRSs use the prevalence of a construct as an anchor for cutoff scores (Choca & Van Denburg, 1997). Prior to evaluating BRSs of personality and clinical scales on the MCMI, validity of an MCMI profile is first established based on non-content responding indicators (scales Validity [V] and Inconsistency [W]). If a protocol is valid, the Clinical Personality scales (scales 1 through 8B) and the Severe Personality scales (S through P) on the MCMI-IV are interpreted through the lens of the modifying scale score (scale Disclosure [X]), and BRS are adjusted up or down. Further, extreme scores on X Disclosure may invalidate the entire protocol if too high or too low (X Disclosure < 7 or $X > 114$, respectively). Finally, Desirability [Y], and Debasement [Z] provide further support for the presence of response bias, requiring clinical interpretation to rule out feigning or denial of psychopathology, but as noted prior, these scales do not formally invalidate the protocol.

MCMI validity scales and their recommended cutoffs have been criticized and alternative interpretive values have been recommended fairly consistently across the extant literature. For example, on the MCMI-III, Charter and Lopez (Charter & Lopez, 2002) suggested that 50% of truly random protocols would be considered valid based on V Validity < 2 when using binomial expansion, which was confirmed using 40,000 computer-generated protocols. Of note, it has been suggested that computer-based random response is distinct from human-based random response (Giromini et al., 2020). Additionally, Daubert and Metzler (2000) found that existing cutoffs for X Disclosure, Y Desirability, and Z Debasement scales (BRS ≥ 85 for X Disclosure and Z Debasement and BRS ≤ 35 for Y Desirability) were not optimized for classification in a simulation study using an outpatient sample of individuals with mood or schizophrenia spectrum disorders. They recommended changes based on diagnostic power (DP; a composite of sensitivity and specificity). These new cutoff suggestions were to decrease X Disclosure scale BRS cutoff from 85 (sensitivity/specificity; 61/81) to 80 (76/71), increase the Y Desirability scale BRS cutoff from 35 (58/76) to 39 (64/74), and decrease the Z Debasement scale BRS from 85 (55/79) to 81 (64/78), increasing sensitivity by 15%, 6%, and 9% and total DP by 3%, 2%, and 4%, respectively.

Schoenberg and colleagues (Schoenberg et al., 2003) likewise suggested changing some cutoffs to increase detection of response bias based on their simulation study of coached college educated simulators ($n = 111$) compared to psychiatric inpatients ($n = 181$). While they suggested keeping the X Disclosure BRS cutoff of ≥ 85 (sensitivity/specificity; 52/73), they recommended reducing the Y Desirability BRS cutoff to ≤ 25 (57/70) and reducing the Z Debasement BRS cutoff to ≥ 81 (59/66). Taken together these studies suggest that the cutoffs offered by Millon are not optimized to identify invalidity or biased responding styles. Nevertheless, despite expressed caution toward use of multi-axial scales in secondary gain scenarios (Sellbom & Bagby, 2008), such use continues. These studies are also reliant on simulation designs (albeit with clinical controls), which may limit generalizability. In short, research surrounding optimization of cutoff scores could assist in the MCMI-IV's utility in scenarios where validity and response bias are paramount.

Concurrently, there is evidence that detection of response bias is sample dependent (Boccaccini & Hart, 2018). For

instance, optimizing cutoffs to increase sensitivity and/or specificity for those feigning PTSD and disability when using the PAI has been recommended (Calhoun et al., 2000; Rogers et al., 1996, 2013; Thomas et al., 2012; Wooley & Rogers, 2015). Similarly, despite the MMPI-2-RF's validity scales' well-established history of sound reliability and validity in the general civilian population (Schroeder et al., 2012), the development of new cutoff scores to more accurately capture response bias has been suggested in veterans (Goodwin et al., 2013; Mason et al., 2013). This recommendation exists because veterans participate in post-military disability system (i.e., service connection process), leading to increased testing invalidity prevalence rates as higher clinical severity (e.g., overreporting) may be rewarded with higher compensation rates.

In veterans, research on other broadband instruments provide context with which to compare rates on the MCMI-IV. For example, on over 36,000 PAI protocols administered to veterans, rates of scale invalidity ranged from 3.1% to 4.5% for non-content response scales (Hong Randomness and Inconsistency), 9.3% to 29.9% for overreporting scales (Hong Malingering and Cognitive Bias Scale of Scales 2), and < 0.1% to 7.5% for underreporting scales (Positive Distortion Scale and Cashel's Discriminant Function)(Shura et al., 2025). Data on the MMPI-2-RF is generally comparable to that of the PAI, and in a large national sample demonstrated that between 0.4% (K-r) and 27.3% (RBS) of MMPI-2-RF symptom validity scales reached skyline elevations, typically due to potential overreporting (Ingram et al., 2020). Both of these had large sample sizes and converged on the following results: (1) Overreporting was far more common than other types of response bias; (2) invalidity rates were highly variable across clinical contexts; and (3) invalidity rates were higher when multiple scales were considered concurrently (as is typically done during interpretation). These studies work to emphasize the sample dependent nature of response bias. Unfortunately, the MCMI-IV has no research to date evaluating its validity scales outside of the test manual. While the MCMI-III validity scale research has been used as proxy, this research is also sparse and does not center on veteran populations (Aguerrevere et al., 2011; Lenny & Dear, 2009; Ruocco et al., 2008).

The primary aim of this study is to assess convergent validity between the symptom validity tests (SVTs) of the MCMI-IV and the SVTs embedded within the MMPI-2-RF. A secondary aim of this study is to evaluate base rates of invalidity on the MCMI-IV validity scales, which are interpreted in the context of prior SVT base rate studies of veterans. Although this study is primarily a psychometric and descriptive study, given the limited research base on the MCMI scales, we predicted that the following: (1) The five MCMI-IV validity scales will generally converge with their MMPI-2-RF counterparts such that non-content scales would most strongly correlate with one another, overreporting scales would most strongly correlate with one another, and underreporting scales would most strongly correlate with one another; (2) Rates of invalidity on the MCMI-IV will be highest with overreporting scales, compared to non-content and underreporting scales; and (3) Given results of prior research on MCMI SVTs, we predicted that alternate cutoff scores using diagnostic accuracy statistics might better identify invalid protocols based on MMPI-2-RF SVTs.

Method

Participants

The VA Corporate Data Warehouse (CDW) was used to identify all veterans in the VA system who completed the MCMI-IV within the Mental Health Assistant (MHA) and the new web-based MHA Web. MHA is the electronic interface of administrating, scoring, and reviewing results of numerous available questionnaires, including the MCMI-IV, and is embedded within the VA's electronic medical record platform. A total of 6,466 MCMI-IV protocols were pulled from 04/2019 through 08/2024. Basic demographic information is presented in Table 1. Most of the sample was male (76.5%), White (72.5%), and receiving some form of service-connected disability (85.7%). Mean age was 44.02 years and 37.6% were married. Demographics for a combined MCMI-IV/MMPI-2-RF subsample after removing non-content-invalid profiles ($n = 1,351$) are similar and also presented in Table 1. This subsample was included in the current study for a correlational comparison of validity scales, and for this group all MCMI-IV and MMPI-2-RF protocols were administered on the same day. The MCMI-IV and MMPI-2-RF protocols were administered across all VA medical centers and clinics in the US *via* MHA or MHA Web.

For the full MCMI-IV sample, 6,090 had primary stop code information, which is a VA system for identifying the type of clinic where services were provided; these codes are attached to clinics when they are built, and reflect the type of service (e.g., group psychotherapy versus individual assessment) or provider (e.g., psychologist or psychiatrist) for a given clinic a patient is scheduled into. The most common codes were 538—Psychological Testing ($n = 2,623$, 40.6%), 502—Mental Health Clinic Individual ($n = 2,050$, 31.7%), 562 - PTSD Individual ($n = 324$, 5.0%), 586 - Residential Rehabilitation Treatment Program ($n = 228$, 3.5%), and 674—Administrative Activities

Table 1. Demographics for full sample and sub-sample administered both test measures.

		Full Sample ($N = 6,466$)	MMPI Sub-Sample ($n = 1,351$)
Class	Variable	*M* (*SD*; range) or *n* (%)	*M* (*SD*; range) or *n* (%)
Age	Years	44.02 (13.45; 19–95)	42.79 (12.57; 19–80)
SC	Total %	78.28 (26.15; 0–100)	78.88 (26.00; 0–100)
Sex	Male	4,949 (76.5%)	1,029 (76.2%)
	Female	1,517 (23.5%)	322 (23.8%)
Race	White	4,690 (72.5%)	993 (73.5%)
	Black	910 (14.1%)	158 (11.7%)
	Asian	114 (1.8%)	17 (1.3%)
	Native	117 (1.8%)	19 (1.4%)
	Pacific	75 (1.2%)	19 (1.4%)
Ethnicity	Hispanic	574 (9.7%)	123 (9.1%)
	Not Hispanic	5,352 (90.3%)	1,228 (90.9%)
SC	Yes	5,543 (85.7%)	1,165 (86.2%)
	No	923 (14.3%)	186 (13.8%)

Note. SC: service connected.

($n = 216$, 3.3%). A total of 4,629 also had a secondary stop code assigned, with the most common in this sample being 510—Psychology Individual ($n = 2{,}901$, 44.9%). Of note, disability clinics are also identified by secondary stop codes, with only 52 (0.8%) participants in the current sample being assigned a 450 - Compensation & Pension code. In sum, the vast majority of the MCMI-IV protocols were administered in individual psychology clinics housed in mental health or devoted testing clinics, with less than 1% given for the explicit purpose of disability.

Measures

MCMI-IV

The MCMI-IV (Millon et al., 2015) is a multi-scale measure of personality styles and psychopathology that was developed to inform treatment planning in adult clinical populations. There are 195 true-false questions written at a fifth grade reading level, and the test takes approximately 30 min to complete. The 195 items of the MCMI-IV load onto the 25 substantive scales, which include 12 clinical personality, 3 severe personality pathology, 7 clinical syndrome, and 3 severe clinical syndrome scales. The MCMI-IV uses Millon's BRSs that reflect prevalence of disorders in clinical populations. These values translate skewed data into standardized scores. A BRS of 60 was set by Millon as the median raw score of a clinical sample, a BRS of 75 was set as the raw score of clients who met criteria for *DSM-5* diagnoses or demonstrated elevated levels of impairment/dysfunction, and a BRS of 85 was set as the raw score of clients who had more pronounced impairment than the levels in the BRS 75 group (Grossman & Amendolace, 2017). A maximum raw score on a scale will lead to a BRS of 115.

The five validity scales include two indicators of content-unrelated response style, V Validity and W Inconsistency, which measure random or careless responding, and three modifying indices, X Disclosure, Y Desirability, and Z Debasement, which reflect content-related response biases such as exaggeration or minimization of symptoms. V Validity solely consists of three items that are so infrequently endorsed that Millon suggests a single endorsement may reflect inattention or random/careless responding, while two or three endorsements are highly indicative of such responding (Groth-Marnat & Wright, 2016; Millon et al., 2015), thereby rendering the profile invalid. W Inconsistency is a set of 25 semantically related question dyads, with each pair answered dissimilarly increasing the scale's score. A score between 20 and 25 is considered indicative of random responding and is highly characteristic of an invalid profile.

X Disclosure reflects a continuum of the respondent's answering style, ranging from underreporting or under-representation of symptoms (raw of 21 to 60 [unlikely], 7 to 20 [possible], and less than six [highly likely]) to overreporting (raw of 21 to 60 [unlikely], 61–114 [possible], 115 and up [highly likely]). Invalid profiles are those with raw scores below 7 and above 114 on X Disclosure. Y Desirability measures defensiveness, and higher scores (BRS of 75 or greater) can indicate an attempt to appear overly altruistic, unusually moral, and without significant psychosocial difficulties. Z measures the degree to which an individual describes themselves as pathologically negative, and a high BRS suggest a self-deprecating nature, or it may be an indication of a "fake bad" profile. Of note, Groth-Marnat cautioned that Y Desirability and Z Debasement are "not particularly good" and "not particularly effective," concluding that that both should be "interpreted with caution" (Groth-Marnat, 2003). Further, although these are categorized as modifying indices, they are not actually used to correct clinical scales the way X Disclosure is used. Regarding overlap, V Validity items do not overlap with any other validity scale; Y Desirability and Z Debasement do not overlap with each other; and W Inconsistency and X Disclosure (in particular) overlap to some degree with each other and both Y Desirability and Z Debasement.

MMPI-2-RF

The MMPI-2-RF (Ben-Porath & Tellegen, 2008, 2011) is a substantially restructured version of the 567-item MMPI-2 comprised of 338 true-false items. These items are written at a fourth-and-a-half to fifth grade reading level and the measure takes approximately 30 min to complete. There are a total of 51 scales comprised of 42 substantive scales that measure clinical constructs and nine that measure symptom validity, including overreporting, underreporting, and invalid content responding. Variable Response Inconsistency (VRIN-r), made up of 53 item dyads, assesses random responding while True Response Inconsistency (TRIN-r), made of 26 item-dyads, assesses fixed-true and/or fixed-false responding. Together, VRIN-r and TRIN-r assess non-content responding. Five scales measure overreporting: Infrequent Responses scale (F-r) on this MMPI version includes 32 items that were endorsed by less than or equal to 10% of the normative population; Infrequent Psychopathology Responses (Fp-r) is made up of items endorsed by less than or equal to 20% of the psychiatric normative population; Infrequent Somatic Responses (Fs) was developed based on items rarely endorsed (< 25% across three samples) by those suffering from chronic pain and other physical/medical conditions; Symptom Validity (FBS-r) was developed incorporating items for "the detection of malingering in personal injury claims" (Lees-Haley et al., 1991); and the Response Bias Scale (RBS) is comprised of 28 items that were developed using empirical keying that successfully discriminated between those who produced valid and invalid scores on performance validity tests administered during disability claims (Gervais et al., 2007). These scales together assess overreporting, and unlike the MCMI-IV, there is relatively little item-level overlap between them (Burchett & Bagby, 2022).

The remaining two standard validity scales are the Uncommon Virtues (L-r) and Adjustment Validity (K-r) scales, both of which are underreporting scales and comprised of 14 items. L-r reflects the extent to which a respondent denies minor faults and short comings, whereas K-r reflects the extent to which a respondent presents as overly well adjusted. Research on the underreporting scales suggests difficulties with sensitivity (Keen et al., 2023), including in Veterans (Khazem et al., 2025). There are various interpretative cutoffs levels for each MMPI-2-RF validity

scale, with conservative ("skyline") cutoff scores for each validity scale set by Ben-Porath and Tellegen (2008/2011).

Procedures

The study was reviewed by the Salisbury VAHCS IRB and determined to be exempt. Data were obtained *via* a Data Access Request Tracker (DART) request; data were initially pulled using Microsoft SQL Server. All analyses were conducted using SAS Enterprise Guide and SPSS 29. All MCMI-IVs available in the system were obtained ($N = 6{,}466$). MMPI-2-RF data were also obtained. The MMPI-2-RF protocols were comprised of a mixture of two subsamples: all MMPI-2-RFs ($N = 103{,}114$) and all MMPI-2s administered ($N = 92{,}723$) in the VA. The MMPI-2s were then rescored to obtain MMPI-2-RF scale scores (Tarescavage et al., 2015). The final full MMPI-2-RF sample was comprised of 195,837 profiles. Not all MMPI-2-RF data were used in the current study, however. Rather, only those data that were obtained on the same day as the MCMI-IV were used, resulting in a final sample of 1,401 same-day MMPI-2-RF and MCMI-IV administrations. The data that support the findings of this study may be available from the corresponding author, RDS, upon reasonable request and based on VA regulations.

The 1,401 sample was used for analyses focused on non-content scales (V and W); otherwise, invalid non-content protocols were removed, leaving a sample of 1,351 used for other analyses. Spearman's rho correlations were calculated, given non-normality of the scores, across all five MCMI-IV SVTs and all nine MMPI-2-RF SVTs. Correlations were considered weak but practically significant if .20 or higher, moderate at .50 and higher, and strong at .80 and higher (Ferguson, 2009). Although the minimal bar of .20 is small in magnitude, the average correlation for variables on self-report measures has been noted to be .24 (Meyer et al., 2001), which is close to the .20 recommended by Ferguson. Base rate data were calculated for the full MCMI-IV sample for all symptom validity scales based on published data in the manual; several other cutoffs were also examined at the BRS cutoff of ≥ 75 (a mild elevation for the clinical scales) and BRS ≥ 85 (skyline elevations for clinical scales).

Diagnostic accuracy analyses were completed for all MCMI-IV scales predicting invalid groups based on MMPI-2-RF scales. AUC values are interpreted as follows: .70–.79 acceptable, .80–.89 excellent, ≥ .90 outstanding (Hosmer & Lemeshow, 2000). For non-content scales, the invalid criterion group was defined as producing invalid scores ($T \geq 80$) on either VRIN-r or TRIN-r. Similarly, the underreporting criterion group was defined as producing invalid scores on either L-r ($T \geq 80$) or K-r ($T \geq 70$). The overreporting criterion group was created in a manner based on (Roma et al., 2023). Invalidity due to overreporting was identified as the subgroup who produced invalid scores on 2 or more out of the five MMPI-2-RF overreporting scales at the following cutoff scores: F-r $T = 120$, Fp-r $T \geq 80$, Fs $T \geq 100$, FBS-r $T \geq 100$, and RBS $T \geq 100$. The valid subsample produced no invalid scores, whereas those with one MMPI-2-RF overreporting scale elevation were excluded from analyses, reducing the sub-sample to 1,114 subjects.

Results

Spearman rho correlations between MCMI-IV scales and MMPI-2-RF scales are presented in Table 2. Scale V Validity did not significantly correlate to any MMPI-2-RF scale, including TRIN-r and VRIN-r. However, W Inconsistency significantly correlated to both MMPI-2-RF non-content based scales with approximately small effects. For overreporting scales, X Disclosure and Z Debasement correlated with all five MMPI-2-RF overreporting scales at a generally moderate level (mean ρ: X Disclosure = .57, Z Debasement = .64), with similar effect ranges falling between .40 (X Disclosure and FBS-r) to .79 (Z and F-r). Similarly, Y Desirability (and X Disclosure negatively) were significantly correlated to L-r and K-r, although the magnitude was much lower than that of overreporting scales and ranged from .21 to −0.68. Thus, our first hypothesis that like scales would most strongly correlate with each other (i.e., overreporting MCMI-IV scales to overreporting MMPI-2-RF scales) was generally supported, with the exception of V Invalidity.

Base rates for invalidity across the five MCMI-IV SVTs are presented in Tables 3 and 4 and summarized across cutoffs in Table 5. At the recommended skyline cutoff scores per the test manual, nearly no participants were invalid across V Validity (< 1%), W Inconsistency (0%), or X Disclosure (for either direction < 1%). Across more liberal (i.e., lower) cutoff scores using BRS ≥ 75, over half the sample was invalid on X Disclosure. More conservative BRS ≥ 85 led to decreased invalidity rates on Y Desirability and Z Debasement, although underreporting ($Y = 3.5\%$) was significantly lower than overreporting ($Z = 23.4\%$). Our second hypothesis was partially supported as overreporting protocols were indeed more prevalent than underreporting profiles

Table 2. Correlations among MCMI-IV and MMPI-2-RF Validity Scales ($n = 1{,}351$).

		NC		OR					UR	
Scale Type	MCMI-IV Scale	VRIN-r	TRIN-r	F-r	Fp-r	Fs	FBS-r	RBS	L-r	K-r
NC	*V Invalidity raw*	0.02	0.00	0.02	−0.00	0.03	0.01	0.02	0.03	0.00
NC	*W Inconsistency raw*	0.22	0.09	−0.02	−0.04	0.01	−0.05	−0.10	−0.05	−0.03
OR/UR	*X Disclosure raw*	−0.03	0.01	0.72	0.61	0.58	0.40	0.55	−0.25	−0.68
UR	*Y Desirability BRS*	0.15	**0.08**	−0.49	−0.35	−0.27	−0.36	−0.46	0.22	0.43
OR	*Z Debasement BRS*	−0.13	0.02	0.79	0.57	0.58	0.60	0.66	−0.21	−0.63

Note. **Bold** $p < .05$; Grey $p < .001$. MCMI-IV: Millon Clinical Multiaxial Inventory-IV; MMPI-2-RF: Minnesota Multiphasic Personality Inventory-2-Restructured Form; BRS: base rate score; NC: non-content response scale; OR: overreporting scale; UR: underreporting scale; VRIN-r: Variable Response Inconsistency; TRIN-r: True Response Inconsistency; F-r: Infrequent Responses; Fp-r: Infrequent Psychopathology Responses; Fs: Infrequent Somatic Responses; FBS-r: Symptom Validity; RBS: Response Bias; L-r: Uncommon Virtues; K-r: Adjustment Validity.

Table 3. Raw score endorsement rates for V Invalidity, W Inconsistency, and X Disclosure.

Raw	V Invalidity (*n* = 6,466)	W Inconsistency (*n* = 6,465)	X Disclosure (*n* = 6,466)	Raw
0	6,355 (98.3%)	47 (0.7%)	4 (0.1%)	0
1	99 (1.5%)	190 (2.9%)	0 (0.0%)	1
2	9 (0.1%)	440 (6.8%)	0 (0.0%)	2
3	3 (<0.1%)	753 (11.6%)	0 (0.0%)	3
4		981 (15.2%)	2 (0.0%)	4
5		1061 (16.4%)	5 (0.1%)	5
6		1000 (15.5%)	0 (0.0%)	6
7		770 (11.9%)	3 (0.0%)	7
8		565 (8.7%)	2 (0.0%)	8
9		332 (5.1%)	2 (0.0%)	9
10		174 (2.7%)	5 (0.1%)	10
11		102 (1.6%)	6413 (99.2%)	11–97
12		33 (0.5%)	5 (0.1%)	98
13		12 (0.2%)	5 (0.1%)	99
14		1 (<0.1%)	6 (0.1%)	100
15		2 (<0.1%)	4 (0.1%)	101
16		1 (<0.1%)	1 (<0.1%)	102
17		0 (<0.1%)	4 (0.1%)	103
18		1 (<0.1%)	1 (<0.1%)	104
> 18		0 (<0.1%)	2 (<0.1%)	109
			1 (<0.1%)	111
			1 (<0.1%)	121

Table 4. Base rate score endorsement rates for Y Desirability and Z Debasement.

Base Rate Score	Y Desirability (*n* = 6,466)	Z Debasement (*n* = 6,466)
< 60	4,207 (65.1%)	1,402 (21.7%)
60	376 (5.8%)	193 (3.0%)
62		201 (3.1%)
63	345 (5.3%)	
64		184 (2.8%)
66	326 (5.0%)	231 (3.6%)
68		223 (3.4%)
69	245 (3.8%)	
70		225 (3.5%)
72	235 (3.6%)	260 (4.0%)
74		263 (4.1%)
75	200 (3.1%)	252 (3.9%)
77		270 (4.2%)
78	162 (2.5%)	
79		326 (5.0%)
80		315 (4.9%)
81	144 (2.2%)	315 (4.9%)
83		292 (4.5%)
85	111 (1.7%)	284 (4.4%)
88		259 (4.0%)
89	74 (1.1%)	
91		244 (3.8%)
93	27 (0.4%)	241 (3.7%)
95		207 (3.2%)
97	14 (0.2%)	157 (2.4%)
100		122 (1.9%)

when considering Y Desirability compared to Z Debasement; however, raw cutoff rates using V, W, and high versus low X were so low as to preclude meaningful comparison.

Diagnostic accuracy analyses first considered non-content response ($n = 1{,}401$) using V and W raw scores to predict invalidity on either VRIN-r or TRIN-r ($T \geq 80$). This resulted in an unacceptable AUC value of .502 for V Invalidity raw; a cutoff score of ≥ 1 achieved a specificity of .98 but at a sensitivity of only .02. Similarly, W Inconsistency raw resulted in an unacceptable AUC of .624; specificity surpassed .90 (.94) at a cutoff score of 10, with a sensitivity of .08. Given the poor outcomes, tables were not devoted to presenting additional ROC information on these scales.

Table 5. Invalidity rates based on various cutoff scores (*N* = 6,466).

Scale	Cutoff	Invalid *N* (%)	*M* (*SD*)	Min–Max
V Invalidity			0.02 (0.16)	0–3
	raw > 1	12 (0.2%)		
W Inconsistency			5.41 (2.38)	0–18
	raw > 19	0 (0.0%)		
X Disclosure			[a]54.11 (18.28)	0–121
	raw < 7	11 (0.2%)	[b]72.25 (16.85)	0–100
	raw > 114	1 (<0.1%)		
	BRS ≥ 75	3,490 (54.1%)		
	BRS ≥ 85	1,832 (28.4%)		
Y Desirability			47.20 (20.59)	0–97
	BRS ≥ 75	732 (11.3%)		
	BRS ≥ 85	226 (3.5%)		
Z Debasement			70.95 (18.90)	0–100
	BRS ≥ 75	3,284 (50.8%)		
	BRS ≥ 85	1,514 (23.4%)		

Note. BRS: Base Rate Score. [a]Raw scores; [b]BRSs.

For overreporting scales, after again excluding non-content invalid protocols leading to $n = 1{,}351$, BRSs were used for both X Disclosure and Z Debasement to predict the overreporting group. There were 231 indeterminate scores (invalid on only one overreporting scale), which were excluded leaving 1,114: of those, 320 (28.7%) were invalid and 794 (71.3%) were valid. Both scales performed well, reaching an excellent AUC for X Disclosure (.879) and outstanding AUC for Z Debasement (.908). Tables 6 and 7 present diagnostic accuracy results across various cutoff scores. For X Disclosure, a cutoff score of BRS ≥ 87 maximized sensitivity at .60 at specificity of .92. Per the test manual (Millon et al., 2015), a BRS for X corresponds to a raw score of 70—72, far lower than the recommended cutoff score of raw > 114 (which is equivalent of BRS 100). Z Debasement performed even better than X, despite the scale not being a primary protocol-invalidating validity scale. At BRS ≥ 84, Z Debasement achieved a sensitivity of .66 at specificity of .91. Using the full sample with these cutoff scores, 21.5% were invalid based on high X, 23.4% were invalid based on high Z, and 9.4% were invalid on both.

Underreporting scales performed better than MCMI-IV non-content scales, but poorer than overreporting scales. Using BRS for *low* X Disclosure predicting invalid scores on either L-r or K-r, AUC was acceptable at .798. BRS ≤ 49 (equivalent to raw 31) maximized sensitivity to .46 at specificity of .90 (Table 8). Y Desirability resulted in an AUC = .757, with a cutoff of BRS ≥ 74 resulting in sensitivity of .43 at specificity of .91 (Table 9). Using the full sample with these cutoff scores, 12.3% were invalid based on low X, 11.3% were invalid based on high Y, and 10.1% were invalid on both. Broadly speaking, results support adjusted cutoffs across overreporting and underreporting MCMI-IV scales.

Discussion

Broadband psychological assessment measures are a staple of psychological and neuropsychological assessment and included symptom validity scales are essential elements for interpretation. Although the MCMI-IV is a commonly-used broadband measure, no studies to date have examined the embedded symptom validity scales; this is the first study of these validity measures.

Table 6. Diagnostic Accuracy of the MCMI-IV X Disclosure BRS to the MMPI-2-RF Overreporting scales ($N = 1{,}114$).

					15%		30%		45%	
≥ cut	LR+	LR−	Sensitivity	Specificity	PPV	NPV	PPV	NPV	PPV	NPV
75	2.42	0.41	0.93	0.62	0.30	0.98	0.51	0.95	0.66	0.91
76	2.52	0.40	0.92	0.64	0.31	0.98	0.52	0.95	0.67	0.91
77	2.67	0.37	0.91	0.66	0.32	0.98	0.53	0.94	0.69	0.90
78	2.85	0.35	0.90	0.68	0.33	0.98	0.55	0.94	0.70	0.90
79	2.89	0.35	0.88	0.70	0.34	0.97	0.55	0.93	0.70	0.87
80	3.18	0.31	0.83	0.74	0.36	0.96	0.58	0.91	0.72	0.84
81	3.47	0.29	0.79	0.77	0.38	0.95	0.60	0.89	0.74	0.82
82	3.89	0.26	0.76	0.80	0.41	0.95	0.63	0.89	0.76	0.81
83	4.12	0.24	0.73	0.82	0.42	0.94	0.64	0.87	0.77	0.79
84	4.35	0.23	0.70	0.84	0.43	0.94	0.65	0.87	0.78	0.77
85	4.99	0.20	0.68	0.86	0.47	0.94	0.68	0.86	0.80	0.77
86	5.75	0.17	0.68	0.88	0.50	0.94	0.71	0.86	0.82	0.77
87	**7.79**	**0.13**	**0.60**	**0.92**	**0.58**	**0.93**	**0.77**	**0.84**	**0.86**	**0.74**
88	10.77	0.09	0.51	0.95	0.66	0.92	0.82	0.82	0.90	0.70
89	12.30	0.08	0.41	0.97	0.68	0.90	0.84	0.79	0.91	0.67
90	15.09	0.07	0.35	0.98	0.73	0.89	0.87	0.78	0.93	0.65
91	16.67	0.06	0.25	0.99	0.75	0.88	0.88	0.75	0.93	0.62
92	18.40	0.05	0.18	0.99	0.76	0.87	0.89	0.74	0.94	0.60
93	22.33	0.04	0.13	0.99	0.80	0.87	0.91	0.73	0.95	0.58
94	18.20	0.05	0.09	1.00	0.76	0.86	0.89	0.72	0.94	0.57
95	17.25	0.06	0.07	1.00	0.75	0.86	0.88	0.71	0.93	0.57

Note. MCMI-IV: Millon Clinical Multiaxial Inventory-IV; BRS: base rate score; MMPI-2-RF: Minnesota Multiphasic Personality Inventory-2-Restructured Form; LR: likelihood ratio; PPV: positive predictive value; NPV: negative predictive value. AUC: .879. **Bold** row indicates cutoff score with best sensitivity at specificity of ≥ .90. Valid $n = 794$; invalid $n = 320$.

Table 7. Diagnostic Accuracy of the MCMI-IV Z Debasement BRS to the MMPI-2-RF Overreporting scales ($N = 1{,}114$).

					15%		30%		45%	
≥ cut	LR+	LR−	Sensitivity	Specificity	PPV	NPV	PPV	NPV	PPV	NPV
75	2.78	0.36	0.95	0.66	0.33	0.99	0.54	0.97	0.69	0.94
76	3.00	0.33	0.93	0.69	0.35	0.98	0.56	0.96	0.71	0.92
78	3.43	0.29	0.90	0.74	0.38	0.98	0.60	0.95	0.74	0.90
80	3.98	0.25	0.88	0.78	0.41	0.97	0.63	0.94	0.76	0.88
81	5.05	0.20	0.83	0.84	0.47	0.96	0.68	0.92	0.81	0.86
82	5.90	0.17	0.74	0.87	0.51	0.95	0.72	0.89	0.83	0.81
84	**7.16**	**0.14**	**0.66**	**0.91**	**0.56**	**0.94**	**0.75**	**0.86**	**0.85**	**0.76**
87	9.18	0.11	0.60	0.94	0.62	0.93	0.80	0.84	0.88	0.74
90	11.24	0.09	0.51	0.96	0.66	0.92	0.83	0.82	0.90	0.70
92	16.36	0.06	0.41	0.98	0.74	0.90	0.88	0.79	0.93	0.67
94	26.73	0.04	0.29	0.99	0.83	0.89	0.92	0.77	0.96	0.63
96	51.50	0.02	0.21	1.00	0.90	0.88	0.96	0.75	0.98	0.61
99	31.33	0.03	0.09	1.00	0.85	0.86	0.93	0.72	0.96	0.57
101	2.78	0.36	0.00	1.00		0.85		0.70		0.55

Note. MCMI-IV: Millon Clinical Multiaxial Inventory-IV; BRS: base rate score; MMPI-2-RF: Minnesota Multiphasic Personality Inventory-2-Restructured Form; LR: likelihood ratio; PPV: positive predictive value; NPV: negative predictive value. AUC: .908. **Bold** row indicates cutoff score with best sensitivity at specificity of ≥ .90. Valid $n = 794$; invalid $n = 320$.

The first aim of this paper was to evaluate convergent validity of the MCMI-IV validity scales to the validity scales of the MMPI-2-RF, for the subsample who completed both measures on the same day ($n = 1{,}403$ for non-content scales, $n = 1{,}351$ for other scales after excluding non-content invalid protocols). Non-content correlations had mixed results: V Validity was not correlated to any MMPI-2-RF scale, including VRIN-r and TRIN-r, whereas W Inconsistency was correlated to both VRIN-r (ρ = .22) and TRIN-r (ρ = .09), as well as RBS, though in the negative direction ($\rho - 0.10$). W Inconsistency was created from 25 response pairs that were highly correlated, and if a respondent answers discrepantly across these pairs, W Inconsistency becomes more elevated. This design is similar to how VRIN-r is constructed, hence the correlation of $\rho = 0.22$ passing the recommended minimum effect of 0.2 (Ferguson, 2009), which supports the validity of W Inconsistency as a non-content scale. The lack of correlation with V Validity likely reflects the construction of this scale, which includes only three deliberately written items that would be very unlikely endorsed, as opposed to using paired-item discrepancy. This does not negate the fact, though, that elevated V Validity scores should invalidate the MCMI-IV due to content non-responsiveness.

Regarding underreporting scales, X Disclosure was negatively correlated with L-r (ρ = −0.25) and K-r (ρ = −0.68); hence, lower X Disclosure scores coinciding with higher L-r and K-r scores. Additionally, Y Desirability was positively correlated with L-r (ρ = .22) and K-r (ρ = .43), as expected. In addition, X Disclosure was positively correlated with MMPI-2-RF overreporting scales and Y Desirability was negatively correlated with those same MMPI-2-RF scales, as expected, further supporting that low X Disclosure and high Y Desirability are indicators of underreporting.

Finally, the five MMPI-2-RF overreporting scales were positively and moderately correlated with both X Disclosure (ρ = .40 [FBS-r] – .72 [F-r]) and Z Debasement (ρ = .57 [Fp-r] – .79 [F-r]). While the magnitudes of these correlations are on the lower range of expected

Table 8. Diagnostic Accuracy of the MCMI-IV X Disclosure BRS to the MMPI-2-RF Underreporting Scales (N = 1,347).

					15%		30%		45%	
≤ cut	LR+	LR−	Sensitivity	Specificity	PPV	NPV	PPV	NPV	PPV	NPV
40	3.98	0.25	0.18	0.96	0.41	0.87	0.63	0.73	0.76	0.59
42	5.21	0.19	0.25	0.95	0.48	0.88	0.69	0.75	0.81	0.61
44	5.02	0.20	0.29	0.94	0.47	0.88	0.68	0.76	0.80	0.62
45	4.14	0.24	0.29	0.93	0.42	0.88	0.64	0.75	0.77	0.61
46	4.91	0.20	0.39	0.92	0.46	0.90	0.68	0.78	0.80	0.65
47	4.93	0.20	0.43	0.91	0.47	0.90	0.68	0.79	0.80	0.66
48	4.99	0.20	0.46	0.91	0.47	0.91	0.68	0.80	0.80	0.67
49	**4.59**	**0.22**	**0.46**	**0.90**	**0.45**	**0.90**	**0.66**	**0.80**	**0.79**	**0.67**
51	4.26	0.23	0.46	0.89	0.43	0.90	0.65	0.80	0.78	0.67
53	4.00	0.25	0.50	0.88	0.41	0.91	0.63	0.80	0.77	0.68
55	3.94	0.25	0.54	0.86	0.41	0.91	0.63	0.81	0.76	0.69
57	3.99	0.25	0.57	0.86	0.41	0.92	0.63	0.82	0.77	0.71
59	3.76	0.27	0.57	0.85	0.40	0.92	0.62	0.82	0.75	0.71
60	3.83	0.26	0.64	0.83	0.40	0.93	0.62	0.84	0.76	0.74

Note. MCMI-IV: Millon Clinical Multiaxial Inventory-IV; BRS: base rate score; MMPI-2-RF: Minnesota Multiphasic Personality Inventory-2-Restructured Form; LR: likelihood ratio; PPV: positive predictive value; NPV: negative predictive value. AUC: .798. **Bold** row indicates cutoff score with best sensitivity at specificity of ≥ .90. Valid *n* = 1,319; invalid *n* = 28.

Table 9. Diagnostic Accuracy of the MCMI-IV Y Desirability BRS to the MMPI-2-RF Underreporting Scales (N = 1,351).

					15%		30%		45%	
≥ cut	LR+	LR−	Sensitivity	Specificity	PPV	NPV	PPV	NPV	PPV	NPV
23	1.17	0.86	1.00	0.14	0.17	1.00	0.33	1.00	0.49	1.00
28	1.23	0.82	0.96	0.21	0.18	0.97	0.34	0.93	0.50	0.88
33	1.35	0.74	0.93	0.31	0.19	0.96	0.37	0.91	0.53	0.84
38	1.42	0.70	0.86	0.40	0.20	0.94	0.38	0.87	0.54	0.77
43	1.56	0.64	0.82	0.47	0.22	0.94	0.40	0.86	0.56	0.76
48	1.74	0.58	0.79	0.55	0.23	0.94	0.43	0.86	0.59	0.76
53	1.94	0.52	0.75	0.61	0.25	0.93	0.45	0.85	0.61	0.75
58	2.11	0.47	0.68	0.68	0.27	0.92	0.47	0.83	0.63	0.72
62	2.47	0.40	0.64	0.74	0.30	0.92	0.51	0.83	0.67	0.72
65	2.73	0.37	0.57	0.79	0.33	0.91	0.54	0.81	0.69	0.69
68	3.41	0.29	0.54	0.84	0.38	0.91	0.59	0.81	0.74	0.69
71	4.03	0.25	0.46	0.89	0.42	0.90	0.63	0.79	0.77	0.67
74	**4.71**	**0.21**	**0.43**	**0.91**	**0.45**	**0.90**	**0.67**	**0.79**	**0.79**	**0.66**
77	5.33	0.19	0.36	0.93	0.48	0.89	0.70	0.77	0.81	0.64
80	4.76	0.21	0.21	0.96	0.46	0.87	0.67	0.74	0.80	0.60
83	6.88	0.15	0.18	0.97	0.55	0.87	0.75	0.73	0.85	0.59
87	14.92	0.07	0.18	0.99	0.72	0.87	0.86	0.74	0.92	0.60
91	18.00	0.06	0.04	1.00	0.76	0.85	0.89	0.71	0.94	0.56
94			0.00	1.00		0.85		0.70		0.55

Note. MCMI-IV: Millon Clinical Multiaxial Inventory-IV; BRS: base rate score; MMPI-2-RF: Minnesota Multiphasic Personality Inventory-2-Restructured Form; LR: likelihood ratio; PPV: positive predictive value; NPV: negative predictive value. AUC: .757. **Bold** row indicates cutoff score with best sensitivity at specificity of ≥ .90. Valid *n* = 1,323; invalid *n* = 28.

relationships (Schroeder et al., 2025), both X Disclosure and Z Debasement likely measure a similar underlying construct defined by non-bizarre yet infrequently endorsed items. Conceptually, X Disclosure contains 121 items, 62% of all MCMI-IV items, selected because of their implausible, contradictory, or overly rare nature in typical clinical populations. As such, X Disclosure likely reflects a measure of generalized symptom burden/severity, which may reflect broad exaggerated or misrepresented experiences. In contrast, Z Debasement was rationally derived from 30 items determined to reflect an overwhelmingly negative self-presentation, the difference which likely accounts for different correlation magnitude.

In sum, across scales, V Validity was unrelated to any MMPI-2-RF scale; W was significantly correlated to both TRIN-r and (in particular) VRIN-r; X Disclosure and Z Debasement were correlated with all five MMPI-2-RF overreporting scales, most strongly with F-r at moderate effects; and low X Disclosure and high Y Desirability were correlated with both L-r and K-r, the later with a moderate effect. These are promising results as they suggest that MCMI-IV validity indices generally measure the intended constructs.

Our second aim was to examine MCMI-IV invalidity rates in a veteran population using all MCMI-IV protocols administered in the VA MHA system (*N* = 6,466). The MCMI-IV manual recommends cutoff scores for the two non-content scales (V Validity and W Inconsistency), as well as both high and low cutoff scores for the modifying index X Disclosure. Our base rates of invalid protocols using these criteria were low (all < 1%).

Rates of non-content responding and underreporting are generally low in the veteran population, although not quite as low as our results. For example, Ingram and colleagues (Ingram et al., 2020) found rates of non-content responding on the MMPI-2-RF to also be low in a large sample of veterans (*N* = 17,640), with invalidity rates at 1.5% on VRIN-r and 3.6% on TRIN-r. To reach a comparable rate on the MCMI-IV, a cutoff score of W Inconsistency ≥ 11 raw yielded a rate of 2.4%, which falls in the middle of the VRIN-r/TRIN-r rates found by Ingram, potentially reflecting a better

cutoff to be further evaluated in future research. Similarly, rates of underreporting on the MMPI-2-RF were also low, at 2.3% on L-r and 0.4% on K-r. To reach a comparable base rate, a cutoff score of X Disclosure ≤ 11 raw would yield a comparable rate of 0.8% invalid due to underreporting. These different base rates support further examination *via* diagnostic accuracy analyses, our final focus. Regardless, the low rates of invalidity on underreporting scales may reflect lower base rates of underreporting in the veteran population more generally and/or difficulties in underreporting detection psychometrically (Khazem et al., 2025).

The low base rate of MCMI-IV overreporting could reflect generally lower rates of negative response bias in veterans than in non-veteran samples; however, this possibility seems unlikely given the base rates of invalidity on other broadband measures. Rates of overreporting psychopathology on the MMPI-2-RF (Ingram et al., 2020) ranged from 12.3% on Fp-r to 23.2% on F-r, which are drastically higher than our overreporting rate of 0.0% on X Disclosure (high). This discrepancy could suggest that X Disclosure is measuring a vastly different overreporting construct than the MMPI-2-RF scales, but this hypothesis seems very unlikely given the correlations with the MMPI-2-RF overreporting validity scales presented in Table 2. Conversely, another explanation is that the cutoff suggested for X Disclosure is not appropriate for the veteran population. This hypothesis aligns with prior research (Daubert & Metzler, 2000) suggesting using a BRS of 80 as a cutoff score as opposed to a raw score of X Disclosure > 114. In our veteran sample, using a BRS of X Disclosure ≥ 85 led to an invalid rate of 28.4%, whereas Ingram and colleague's found somewhat discrepant rates ranging from 5.2% (FBS-r) to 23.2% (F-r) on individual scales but a strikingly similar 24.6% invalidity rate when using 2+ SVT failures, as is often recommended. Millon is unclear in the MCMI-IV manual as to how cutoff scores were derived, although clearly, at least in the veteran population, X Disclosure > 114 appears inappropriately conservative.

We also examined the Y Desirability (underreporting) and Z Debasement (overreporting) scales. The manual does not provide cutoff scores that render the MCMI-IV formally invalid for these scales, although it recommends BRSs of 75 or higher as indicating positive or negative impression management, respectively. This cutoff led to relatively high base rates in our sample (11.3% on Y Desirability and 50.8% on Z); however, increasing the threshold to BRS ≥ 85 led to underreporting rates at 3.5% and overreporting rates at 23.4%. These rates essentially duplicate the MMPI-2-RF rates of underreporting on L-r (2.3%) and overreporting on F-r (23.2%) (Ingram et al., 2020). Future research should further validate these two under-utilized scales on the MCMI-IV, as these scales (although diminished in the literature) might be better indicators of response bias than X Disclosure. This possibility is further highlighted by the fact that X Disclosure is used as a modifying scale, along with Generalized Anxiety (A) and Major Depression (CC), similar to how K was used in the MMPI-2 to modify the clinical scales before that practice was abandoned due to its limited utility (Barthlow et al., 2002). Given the use of X Disclosure as a bi-directional scale modifier, Y Desirability and Z Debasement might show better utility for identifying protocol invalidity.

Given these differences in rates of invalidity using manual cutoff scores, we conducted ROC analyses for all five MCMI-IV scales (both directions for X). Neither non-content scales of V Invalidity nor W Inconsistency reached acceptable AUCs against VRIN-r or TRIN-r. Given the content of V Invalidity, retaining the manual cutoff score of raw > 1 is still recommended, given the high unlikelihood of endorsing two or three of those items. Similarly, retaining use of the manual cutoff score for W Inconsistency of raw > 19 as invalid is still suggested, given these are items that highly correlated with one another (though, correlations not in the manual), thus conceptually, to endorse them discrepantly would indicate a problem.

In contrast to V Invalidity and W Inconsistency, X Disclosure, Y Desirability, and Z Debasement all performed well as far as AUC results. However, as predicated, cutoff scores required modification to optimize classification properties, especially for X Disclosure. To identify overreporting, X Disclosure BRS ≥ 87 and Z Debasement BRS ≥ 84 might be considered. This second score is only 1 BRS point off from the skyline recommendation in the manual, which when combined with correlation results suggests that Z Debasement might function far better than previously thought. In contrast, the alternate score for X Disclosure was notably discrepant from the manual recommendation. Given the clinical personality scales are adjusted based on moderate X Disclosure elevations, we recommend considering the entire protocol invalid due to overreporting if both If X Disclosure BRS ≥ 87 and Z Debasement ≥ 84. This approach uses a criterion of two or more validity scales as a basis for invalidity, a practice now common to PVTs and of increasing interest with SVTs.

Like high X Disclosure elevations predict overreporting, low scores functioned well to predict underreporting, though again at a very different cutoff score of BRS ≤ 49. Similar to Z Debasement, Y Desirability predicted underreporting, though a lower elevation of BRS ≥ 74 faired the best. Akin to using two validity scores to determine overreporting, a similar approach might be considered in the identification of underreporting. More specifically, a greater degree of confidence can be asserted in those with both X Disclosure BRS ≤ 49 and Y Desirability BRS ≥ 74, whereas only one invalid score might be considered indeterminate.

This study is the first to evaluate the validity scales of the MCMI-IV in any manner. A strength of this study was the use of population-based data from VA, allowing for high sample size with direct application to the Veteran population. Additionally, the mixed nature of clinics the MCMI-IV was administered in allows for broad applicability across settings (e.g., outpatient testing clinicals, residential). However, there are several limitations to this study. First, due to the use of clinical data, available demographic and diagnostic data were limited. Second, the sample included Veterans who were mostly middle-aged, White males, most of whom were receiving service connected disability, and results might not generalize to populations outside of that demographic.

Lastly, these analyses reflect a validity by proxy relationship—however, convergence between MCMI and MMPI elevation rates does not imply equivalent classification accuracy, due to compounded systematic and random error as well as inherent error within SVT measurement. Additionally, results are limited by the accuracy of the criterion grouping, thus results of ROC analyses reflect prediction of MMPI-2-RF validity scales, not of response bias per se. Further research using multi-method assessment or more complex external criteria (e.g., malingering criteria [Sherman et al., 2020] in forensic patients) would be beneficial. Additional research on MCMI-IV validity scales could also rely on some of the recent innovations in scale development to potentially revise scales or even develop supplemental ones, for example using the Scale of Scales paradigm that was applied to the PAI (Boress et al., 2022). Nonetheless, this study provides results from an initial step toward evaluating the MCMI-IV symptom validity scales.

Disclosure statement

The views, opinions, and/or findings contained in this article are those of the authors and should not be construed as an official VA, DHA, DoD, or US Government position, policy, or decision unless so designated by other official documentation. The authors declare no conflicts of interest, financial or otherwise.

Funding

This work was supported by the Salisbury VA Health Care System and the VA Mid-Atlantic Mental Illness Research, Education, and Clinical Center (MIRECC). The material is also the result of work supported with resources and the use of facilities at Robert J Dole VAMC. There was no other funding supporting this study.

ORCID

Robert D. Shura http://orcid.org/0000-0002-9505-0080
Paul B. Ingram http://orcid.org/0000-0002-5409-4896
Ryan W. Schroeder http://orcid.org/0009-0008-9775-2562
Patrick Armistead-Jehle http://orcid.org/0000-0002-6784-1432
Luciano Giromini http://orcid.org/0000-0002-9540-4803

References

Aguerrevere, L. E., Greve, K. W., Bianchini, K. J., & Ord, J. S. (2011). Classification accuracy of the Millon Clinical Multiaxial Inventory–III modifier indices in the detection of malingering in traumatic brain injury. *Journal of Clinical and Experimental Neuropsychology*, *33*(5), 497–504. https://doi.org/10.1080/13803395.2010.535503

Barthlow, D. L., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., & McNulty, J. L. (2002). The appropriateness of the MMPI-2 K correction. *Assessment*, *9*(3), 219–229. https://doi.org/10.1177/1073191102009003001

Ben-Porath, Y. S., & Tellegen, A. (2008). *MMPI-2-RF: Manual for administration, scoring, and interpretation*. University of Minnesota Press, Pearson.

Boccaccini, M. T., & Hart, J. R. (2018). Response style on the personality assessment inventory and other multiscale inventories. In *Clinical assessment of malingering and deception* (4th ed., pp. 280–300). The Guilford Press. https://research.ebsco.com/linkprocessor/plink?id=d893e7a5-e6c2-3bcc-b2db-c9be4acaff9d

Boress, K., Gaasedelen, O. J., Croghan, A., Johnson, M. K., Caraher, K., Basso, M. R., & Whiteside, D. M. (2022). Validation of the Personality Assessment Inventory (PAI) scale of scales in a mixed clinical sample. *The Clinical Neuropsychologist*, *36*(7), 1844–1859. https://doi.org/10.1080/13854046.2021.1900400

Burchett, D., & Bagby, R. M. (2022). Assessing negative response bias: A review of the noncredible overreporting scales of the MMPI-2-RF and MMPI-3. *Psychological Injury and Law*, *15*(1), 22–36. https://doi.org/10.1007/s12207-021-09435-9

Calhoun, P. S., Earnst, K. S., Tucker, D. D., Kirby, A. C., & Beckham, J. C. (2000). Feigning combat-related posttraumatic stress disorder on the Personality Assessment Inventory. *Journal of Personality Assessment*, *75*(2), 338–350. https://doi.org/10.1207/S15327752JPA7502_11

Charter, R. A., & Lopez, M. N. (2002). Million Clinical Multiaxial Inventory (MCMI-III): The inability of the validity conditions to detect random responders. *Journal of Clinical Psychology*, *58*(12), 1615–1617. https://doi.org/10.1002/jclp.10073

Choca, J. P., & Pignolo, C. (2022). Assessing negative response bias with the Millon Clinical Multiaxial Inventory-IV (MCMI-IV): A review of the literature. *Psychological Injury and Law*, *15*(1), 48–55. https://doi.org/10.1007/s12207-022-09442-4

Choca, J. P., & Van Denburg, E. J. (1997). *Interpretive guide to the Millon Clinical Multiaxial Inventory* (2nd ed.). American Psychological Association.

Daubert, S. D., & Metzler, A. E. (2000). The detection of fake-bad and fake-good responding on the Millon Clinical Multiaxial Inventory III. *Psychological Assessment*, *12*(4), 418–424. https://doi.org/10.1037/1040-3590.12.4.418

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*(5), 532–538. https://doi.org/10.1037/a0015808

Gervais, R. O., Ben-Porath, Y. S., Wygant, D. B., & Green, P. (2007). Development and validation of a Response Bias Scale (RBS) for the MMPI-2. *Assessment*, *14*(2), 196–208. https://doi.org/10.1177/1073191106295861

Giromini, L., Viglione, D. J., Pignolo, C., & Zennaro, A. (2020). An Inventory of Problems–29 study on random responding using experimental feigners, honest controls, and computer-generated data. *Journal of Personality Assessment*, *102*(6), 731–742. https://doi.org/10.1080/00223891.2019.1639188

Goodwin, B. E., Sellbom, M., & Arbisi, P. A. (2013). Posttraumatic stress disorder in veterans: The utility of the MMPI–2–RF validity scales in detecting overreported symptoms. *Psychological Assessment*, *25*(3), 671–678. https://doi.org/10.1037/a0032214

Grossman, S., Amendolace, B. (2017). *Essentials of MCMI-IV Assessment*. Wiley. https://liblynxgateway.com/va?url=https%3a//search.ebscohost.com/login.aspx%3fdirect%3dtrue%26db%3dnlebk%26AN%3d1459279%26site%3dehost-live

Groth-Marnat, G. (2003). *Handbook of psychological assessment* (4th ed.). John Wiley & Sons, Inc.

Groth-Marnat, G., Wright, J. (2016). *Handbook of psychological assessment* (6th ed.). Wiley. Wiley.Com. https://www.wiley.com/en-us/Handbook+of+Psychological+Assessment%2C+6th+Edition-p-9781118960646

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). John Wiley & Sons.

Ingram, P. B., Tarescavage, A. M., Ben-Porath, Y. S., & Oehlert, M. E. (2020). Patterns of MMPI-2-Restructured Form (MMPI-2-RF) validity scale scores observed across Veteran Affairs settings. *Psychological Services*, *17*(3), 355–362. https://doi.org/10.1037/ser0000339

Keen, M. A., Lee, T. T. C., Pscheid, K., & Forbey, J. D. (2023). Examination of the generalizability of underreporting detected by the MMPI-2-RF in a correctional sample. *Assessment*, *30*(4), 1157–1167. https://doi.org/10.1177/10731911221089036

Khazem, L., Keen, M., Rodriguez, T. R., Ingram, P. B., Hay, J. M., Long, C. M., Bryan, C. J., & Anestis, J. C. (2025). Detecting Simulated Underreporting on the Minnesota Multiphasic Personality Inventory-3 (MMPI-3) in Veterans with past-month death/suicide ideation. *Suicide & Life-Threatening Behavior*, *55*(2), e13170. https://doi.org/10.1111/sltb.13170

Lees-Haley, P. R., English, L. T., & Glenn, W. J. (1991). A Fake Bad Scale on the MMPI-2 for personal injury claimants. *Psychological Reports*, *68*(1), 203–210. https://doi.org/10.2466/PR0.68.1.203-210

Lenny, P., & Dear, G. E. (2009). Faking good on the MCMI-III: Implications for child custody evaluations. *Journal of Personality Assessment*, *91*(6), 553–559. https://doi.org/10.1080/00223890903228505

Martin, P. K., Schroeder, R. W., & Odland, A. P. (2025). Neuropsychological validity assessment beliefs and practices: A survey of North American neuropsychologists and validity assessment experts. *Archives of Clinical Neuropsychology*, *40*(2), 201–223. https://doi.org/10.1093/arclin/acae102

Mason, L. H., Shandera-Ochsner, A. L., Williamson, K. D., Harp, J. P., Edmundson, M., Berry, D. T. R., & High, W. M.J. (2013). Accuracy of MMPI–2–RF validity scales for identifying feigned PTSD symptoms, random responding, and genuine PTSD. *Journal of Personality Assessment*, *95*(6), 585–593. https://doi.org/10.1080/00223891.2013.819512

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *The American Psychologist*, *56*(2), 128–165. https://doi.org/10.1037/0003-066X.56.2.128

Millon, T., Grossman, S., & Millon, C. (2015). *Millon Clinical Multiaxial Inventory-IV (MCMI-IV): Manual*. NCS Pearson, Inc.

Rabin, L. A., Paolillo, E., & Barr, W. B. (2016). Stability in test-usage practices of clinical neuropsychologists in the United States and Canada over a 10-year period: A follow-up survey of INS and NAN members. *Archives of Clinical Neuropsychology*, *31*(3), 206–230. https://doi.org/10.1093/arclin/acw007

Rogers, R., Gillard, N. D., Wooley, C. N., & Kelsey, K. R. (2013). Cross-validation of the PAI Negative Distortion Scale for feigned mental disorders: A research report. *Assessment*, *20*(1), 36–42. https://doi.org/10.1177/1073191112451493

Rogers, R., Sewell, K. W., Morey, L. C., & Ustad, K. L. (1996). Detection of feigned mental disorders on the Personality Assessment Inventory: A discriminant analysis. *Journal of Personality Assessment*, *67*(3), 629–640. https://doi.org/10.1207/s15327752jpa6703_15

Roma, P., Giromini, L., Sellbom, M., Cardinale, A., Ferracuti, S., & Mazza, C. (2023). The ecological validity of the IOP-29: A follow-up study using the MMPI-2-RF and the SIMS as criterion variables. *Psychological Assessment*, *35*(10), 868–879. https://doi.org/10.1037/pas0001273

Ruocco, A. C., Swirsky-Sacchetti, T., Chute, D. L., Mandel, S., Platek, S. M., & Zillmer, E. A. (2008). Distinguishing between neuropsychological malingering and exaggerated psychiatric symptoms in a neuropsychological setting. *The Clinical Neuropsychologist*, *22*(3), 547–564. https://doi.org/10.1080/13854040701336444

Schoenberg, M. R., Dorr, D., & Morgan, C. D. (2003). The ability of the Millon Clinical Multiaxial Inventory—Third Edition to detect malingering. *Psychological Assessment*, *15*(2), 198–204. https://doi.org/10.1037/1040-3590.15.2.198

Schroeder, R. W., Baade, L. E., Peck, C. P., VonDran, E. J., Brockman, C. J., Webster, B. K., & Heinrichs, R. J. (2012). Validation of MMPI-2-RF Validity Scales in criterion group neuropsychological samples. *The Clinical Neuropsychologist*, *26*(1), 129–146. https://doi.org/10.1080/13854046.2011.639314

Schroeder, R. W., Bieu, R. K., & Snodgrass, M. (2025). Comparing the Cognitive Bias Scale and Cognitive Bias Scale of Scales to other Personality Assessment Inventory validity scales for detecting non-credible memory dysfunction in a clinical veteran sample. *Journal of Clinical and Experimental Neuropsychology*, *47*(1-2), 12–25. https://doi.org/10.1080/13803395.2025.2464635

Sellbom, M., & Bagby, R. M. (2008). Response styles on multiscale inventories.In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (3rd ed., pp. 182–206). The Guilford Press.

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

Shura, R. D., Ingram, P. B., Schroeder, R. W., & Armistead-Jehle, P. (2025). Interpreting the Personality Assessment Inventory (PAI) validity scales: Leveraging population-level Veteran Affairs (VA) data from 2008 to 2024. *Psychological Assessment*. Advance online publication. https://doi.org/10.1037/pas0001403

Tarescavage, A. M., Alosco, M. L., Ben-Porath, Y. S., Wood, A., & Luna-Jones, L. (2015). Minnesota Multiphasic Personality Inventory-2-Restructured Form (MMPI-2-RF) scores generated from the MMPI-2 and MMPI-2-RF test booklets: Internal structure comparability in a sample of criminal defendants. *Assessment*, *22*(2), 188–197. https://doi.org/10.1177/1073191114537347

Thomas, K. M., Hopwood, C. J., Orlando, M. J., Weathers, F. W., & McDevitt-Murphy, M. E. (2012). Detecting feigned PTSD using the Personality Assessment Inventory. *Psychological Injury and Law*, *5*(3-4), 192–201. https://doi.org/10.1007/s12207-011-9111-6

Wooley, C. N., & Rogers, R. (2015). The effectiveness of the personality assessment inventory with feigned PTSD: An initial investigation of Resnick's model of malingering. *Assessment*, *22*(4), 449–458. https://doi.org/10.1177/1073191114552076