Journal of Psychopathology and Behavioral Assessment (2025) 47:61
https://doi.org/10.1007/s10862-025-10233-9

A Meta-Analysis of the Personality Assessment Inventory (PAI) Over-Reporting Scales and Supplemental Indicators

Tristan T. Herring[1] · **Arianna D. Albertorio**[1] · **Keegan J. Diehl**[1] · **Paul B. Ingram**[2]

Accepted: 11 June 2025

Abstract

The Personality Assessment Inventory (PAI) is a widely used broadband personality assessment embedded with validity scales capturing over-reported pathology. This meta-analysis examines the utility of the PAI over-reporting scales as measured by mean differences on standard (NIM, MAL, RDF) and supplemental (NDS, MFI, HMI, CBS, CB-SOS) scales. These comparisons are made across simulation and criterion studies to compare scale efficacy and effectiveness, respectively. 6,451 participants across 43 studies were analyzed using a series of random and fixed effect meta-analyses. We calculated general (e.g., detection effectiveness) and specific (e.g., simulation vs criterion) effect sizes as well as summarized classification statistics and other contextual information (e.g., criterion groups) observed across the literature. Results demonstrate moderate to large effect sizes across most standard ($g = .99$-1.50) and supplementary scales ($g = .84$-1.81), consistent with expected ranges. Compared to criterion studies ($g = .17$-$.92$), simulation designs ($g = 1.16$-2.27) were more effective ($g_{differences} = .57 - 1.60$). Published studies also produced lower effects than unpublished ($g_{differences} = -.21$—$-.59$). Our findings generally support the efficacy of the PAI's over-reporting scales, as well as their effectiveness in non-simulation (e.g., criterion-based) designs. However, RDF does not effectively measure over-reporting in criterion groups and should not be used for those decisions at present. Implications and future directions for the PAI and over-reporting are discussed.

Keywords Validity · PAI · Over-reporting · Personality · Meta Analysis

Content-based response bias detection is the assessment process focused on capturing potential symptom misrepresentation. Unsurprisingly, numerous professional agencies guide practitioners to use multiple well-validated measures of responding styles (Sweet et al., 2021). Over-reporting, a type of content-based responding, is particularly impactful and problematic across assessment contexts (Denning & Shura, 2017; DeViva & Bloem, 2003; Garriga, 2007; Gouvier et al., 2003; McDermott & Feldman, 2007; Morris et al., 2022). For example, over-reporting is estimated to occur in roughly half of disability claims and cost $20.02 billion (Chafetz & Underhill, 2013). These real-world impacts underscore forensic evaluations' use of evidence-based assessments to detect potential symptom misrepresentation (Medoff, 2003; Rotgers & Barrett, 1996).

Young's (2019) negative biased responding model posits that distorted responding results from intersecting dimensions of intentionality and deception. Within this framework, *conscious intent* and *degree of deception* can range from unconscious and low deception (e.g., somatization) to *conscious* and *high deception* (e.g., malingering), with a wide array of possible dimensional configurations that conceptualize distinct negative response bias tendencies. Theories of content-based invalid responding also conceptualize over-reporting in other ways (e.g., Halligan et al., 2003; LoPiccolo et al., 1999), including by specific domain or across distinct endorsement strategy (e.g., Rogers & Bender, 2018). Despite different underlying assumptions, these theories all conclude that detection of feigned presentation is a challenging area of work, one which rarely offers clear and precise decisional capacity. Accordingly, practice guidelines have concluded that the use of multiple well-validated measures

✉ Tristan T. Herring
tristantherring@gmail.com

✉ Paul B. Ingram
pbingram@gmail.com

[1] Texas Tech University, Lubbock, TX, USA

[2] Department of Psychological Sciences, Texas Tech University, 2810 18Th Street, Box 42501, Lubbock, TX 79409, USA

Published online: 17 July 2025

of response sets (e.g., Sweet et al., 2021) is needed to ensure accurate classification.

Research on content-based response styles has made considerate progress. For example, efforts to differentiate performance and symptom validity tests (PVT and SVT, respectively; Larrabee, 2012) into distinct symptom domains (e.g., somatic *or* psychiatric focused feigning; see Burchette & Bagby, 2022) have formed an integral part of the current response validity paradigm. Further, recommendations have been established and developed over time to more explicitly detect cognitive, somatic, and psychiatric malingering as separate constructs (see Sherman et al., 2020; Slick et al., 1999). The general acceptance of these standards as part of professional consensus updates (Sweet et al., 2021) and common approaches to interpretive practice (Rogers & Bender, 2018) underscore the importance of accurate over-reporting detection broadly (Woody, 2016).

It is also worth noting that the response bias detection standards described above have faced cogent and clear criticisms for their lack of scientific rigor evident in both statistical and methodological issues (Leonhard, 2023a, 2023b, 2023c). The result of those concerns is, Leonhard argues, a more limited predictive and clinical utility within validity testing, which must be addressed to advance validity determinations. While a full review and discussion with Leonhard's concerns (and the responses to them, e.g., Bush, 2023; Young & Erdodi, 2024) is outside the scope of the paper, the discussion generated by those papers further underscores the still developing nature of response bias detection. Accordingly, development and evaluation of validity assessment tools used in evidence-based assessment remains an ongoing need (Arbisi & Beck, 2016; Sleep et al., 2015; Wright et al., 2022), as does the integration of multimethod assessment methods in testing (e.g., PVT and SVTs; Burchett & Bagby, 2014; Burchett & Bagby, 2022).

The Personality Assessment Inventory (Morey, 1991, 2007) is a common personality and psychopathology measure meeting those evidence-based standards (see Charles et al., 2022; Kurtz & Pintarelli, 2024), which sees wide clinical and forensic utility and use (e.g., Bow et al., 2006; Gardner et al., 2015; Meaux et al., 2022). Moreover, recent doctoral training research also found that the PAI is the most frequently trained structured self-report measure of emotional and behavioral functioning in health service psychology (Ingram et al., 2022). This use also includes in applied clinical experiences (e.g., practicum), the strongest predictor of assessment-related competency and career aspiration (Bergquist et al., 2023). The PAI's popularity amassed literature across a variety of substantive areas, including assessment of somatic and physical health issues (Hoyt & Walter, 2022; Karlin et al., 2005; Oltmanns et al., 2020), therapy process and treatment outcome relationships (Anestis et al., 2021; Cersosimo et al., 2022; Nevid et al., 2020), risk and level of service and care need (Kim et al., 2021; Sinclair et al., 2014), and other psychiatric and neuropsychological symptoms and symptom topography (Aikman & Souheaver, 2010; McCredie et al., 2023, Ingram et al., 2021).

The PAI has a substantial, and growing research base, for its over-reporting validity scales. Additionally, researchers have continuously developed and refined other supplemental indices targeting neuropsychological and forensic settings (e.g., Boccaccini & Hart, 2018; Fokas & Brovko, 2020; Kurtz et al., 2023; Meaux et al., 2022). A meta-analytic review is warranted to determine scale performance across contexts (e.g., clinical vs simulation) and aid clinicians in assessment interpretation. However, the last meta-analysis of the PAI over-reporting scales was conducted nearly 15 years ago (Hawes & Boccaccini, 2009). Not only has research grown on existing measures at that time, but several new scales have since been developed and cross-replicated numerous times.

Hawes and Boccaccini's (2009) PAI meta-analysis included only the three standard over-reporting scales, drawn from a total of 26 studies. They generally found moderate to large differences between those identified as feigners and honest controls on the Negative Impression Management (NIM; d's range = 0.75 to 1.56; $k = 23$) and the Malingering Index (MAL; d's range = 0.94 to 1.27; $k = 16$) scales. Rogers' Discriminant Function (RDF) group differences magnitude ranged considerably ($d_{range} = 0.31$ to 1.69; $k = 7$), and its performance was notably worse for non-simulation designs (e.g., criterion-based designs). While differences between criterion and simulation groups are not surprising or uncommon (Rogers & Bender, 2018), the range of effects (including negative effect sizes) seen in non-simulation studies led Hawes and Boccaccini to caution against RDF's clinical utility. The between-condition differences (e.g., simulation and criterion groups, coached and uncoached, or sample demographics) on evaluated PAI scales (only NIM, MAL, and RDF), were broadly consistent with historic patterns across similar broadband measures (e.g., Ingram & Ternes, 2016; Rogers et al., 2003; Sharf et al., 2017). These consistent moderate to large effect sizes in over-reporting measures have shifted interpretive guidelines (see Rogers & Bender, 2018).[1]

[1] The effect range terms were defined as .75 moderate, 1.25 as large, and 1.50 as very large (Rogers & Bender, 2018).

Since Hawes and Boccaccini's (2009) PAI over-reporting meta-analyses, several new PAI validity scales were developed, growing literature on existing scales through cross-validations (e.g., Morris et al., 2022) and through new methodological and content-based coverage (e.g., Boress et al., 2022a, 2022b; Gaasedelen et al., 2019). The new scales are reviewed briefly here in their scope and design qualities. Readers are referred to Kurtz and McCredie (2022) for a recent, comprehensive review of recently developed validity scales for the PAI. The Cognitive Bias Scale (CBS; Gaasedelen et al., 2019) was developed using a bootstrapping method to predict performance validity test failure in a neuropsychological sample (see Burchette & Bagby, 2022 for a detailed discussion of this validation procedure). The Cognitive Bias Scale of Scales (CB-SOS) were also developed using a bootstrapping method predicting performance validity failure, but unlike the CBS, CB-SOS uses multiple clinical scales, subscales, and weighted scales (Boress et al., 2022b). Likewise, Gaines et al. (2013) developed the Multiscale Feigning Index using the average of multiple clinical scales in correctional settings to predict failure of the Structured Interview of Reported Symptoms (SIRS; Rogers et al., 1992), a common stand-alone SVT. The Negative-Distortion Scale (NDS) was also validated using the SIRS and uses an indiscriminate endorsement approach in forensic contexts. Finally, the Hong Malingering Index (HMI) used a weighted average of multiple scales selected via stepwise discriminant function analysis the South Korean version of the PAI (Hong & Kim, 2001).

In each case, these new scales (i.e., CBS, CB-SOS, MFI, NDS) have also evidenced numerous independent replications and cross-validations, providing strong support for the effect sizes observed. These between-group effects (e.g., pass/fail performance validity tests, or simulation/control) generally fall between 0.75 to 1.50 as an *over-reporting standard effect range*, roughly spanning the moderate to large range proposed by Rogers et al. (2003), with variations in performance as a function of established moderating factors (Ingram & Ternes, 2016), such as evaluative context or clinical sample (e.g., CBS performance across samples; Armistead-Jehle et al., 2020; Boress et al., 2022a, 2022b; Gaasedelen et al., 2019; Morris et al., 2022; Pignolo et al., 2023; Shura et al., 2023; Tylicki et al., 2021). Accordingly, these scales are currently labeled as incremental indicators, which "*may eventually be useful additions to the standard scoring protocol of the PAI*" if their validity and predictive evidence remain strong (McCredie & Morey, 2018, p.1298).

Current Study

This study provides an updated review and meta-analytic evaluation of the Personality Assessment Inventory (PAI)'s overreporting symptom validity measures. Using random and fixed effect meta-analyses, we examine PAI over-reporting scale effectiveness across study-coded criterion groups developed based on historic moderating influences (e.g., published/unpublished, simulation/criterion design, coached and uncoached simulation design). We also compare the magnitude of effects observed in the current study with those of Hawes and Boccaccini (2009), examining the consistency of standard effect ranges on the PAI's validity scale performance over time. In addition to NIM, MAL, and RDS, this study includes the following scales in its review: Cognitive Bias Scale (CBS), Multiscale Feigning Index (MFI), all versions of Cognitive Bias Scale of Scales (CB-SOS), and Negative Distortion Scale (NDS). Consistent with trends we have observed across studies during our review of past research, we hypothesize (1) each of the PAI validity scales' group difference magnitudes will generally fall within the standard effect range ($g_{range} = 0.75$ to 1.50), (2) simulation studies will outperform criterion studies by between half and full standard deviation ($g = 0.50$ to 1.00), and (3) coached simulation designs will more closely approximate criterion groups by a moderate magnitude of effect (approximately 0.30 effect difference). Standard deviations are expected to be between 1.5 and 2.0 times the normative sample in over-reporting known group and simulation designs.

Methods

General Approach

Meta-analyses synthesize data across published and unpublished studies to estimate a general effect. Generally, this approach falls within two categories: fixed effects and random effects. Fixed effect analyses assume one true effect in the population (e.g., all groups produce similar test score patterns), whereas random effect analyses assume that the population's true effect size differs based on moderators (e.g., sampling methods, demographics, etc.) Fixed effects are preferred when studies are assumed to have similar characteristics and therefore provide an accurate effect size estimate for that context. Random effects analyses are preferred as variability between samples and groups is assumed

(Borenstein et al., 2009). Broadly, meta-analytic techniques provide an empirical means for researchers and clinicians to assess the literature and its context.

This meta-analysis analyzed studies that reported mean differences on PAI over-reporting indices between credible and noncredible self-reporting groups (e.g., feigners) using a weighted mean effect size approach. The weighted mean effect size approach assumes that larger sample sizes will closer approximate population effects and, therefore, ensures that data sources with larger samples produce greater influence than those with smaller samples. Random effect analyses were conducted at the broadest level to assess general effects and variability; fixed effect analyses were subsequently calculated at levels of analyses empirically known to moderate performance in noncredible responding behavior research (e.g., simulation vs criterion; Rogers & Bender, 2018). Coached and uncoached designs were examined via random effects analysis given different coaching instructions. All analyses were conducted using version 4.6 of the Metafor R package (Viechtbauer, 2010). Precision in meta-analyses is impacted when analysis is conducted on a small number of studies. Thus, we made an a priori *decision to* require at least five studies for primary analysis (Borenstein et al., 2009), with at least three required for exploratory analysis of newer scales.

Lastly, consistent with Hawes and Boccaccini's (2009) methods and recommendations, we collect sensitivity and specificity values for common cut scores across analyzed over-reporting scales. As with other meta-analyses which have done this (Hawes & Boccaccini, 2009; Sharf et al., 2017), the presented sensitivity and specificity values *are not* meta-analytically weighted and, as such, it should not be assumed that performance generalizes (as one might conclude from a simulation or known-group study). Rather, average sensitivity and specificity values reflect trends in classification overall, but do not guarantee field agreement of standards (e.g., 0.90 specificity; see Sherman et al., 2020) will apply to a novel population.

Statistical Analyses

Sequentially, standardized mean differences between credible and noncredible groups across PAI over-reporting indices were calculated. After, these differences were transformed into effect sizes and then converted into weighted effect sizes. These effect sizes (based upon Hedge's *g*; effect size corrected for unequal effect sizes, see Hedges & Olkin, 1985) were interpreted using Rogers and Bender's (2018) effect size estimate guidelines (i.e., Moderate = 0.75, Large = 1.25, Very Large = 1.75). A half a standard deviation difference (e.g., *d* or *g* at or above 0.50) serves as a common threshold for clinical significance for mean differences.

Search Criteria and Procedure

Using Hawes and Boccaccini's (2009) search criteria in the PsycINFO and Proquest databases (*PAI or Personality Assessment Inventory with validity scales, feigning, malingering, faking, simulation, dissimulation, and fake-bad*), an initial screening was conducted by the first author to gather studies for examination. A reverse search was also employed for development and validation articles of newer measures (e.g., MFI, CBS, etc.), and a follow-up search was conducted in 2025 to collect any articles published between the initial screening (October 2023) and manuscript submission (June 2025). The first author contacted corresponding authors on PAI over-reporting manuscripts published over the last 15 years (i.e., 2010 through May 2025) to request unpublished data not currently under review. One investigator responded with an article previously identified in the initial screening. This low response rate (6%) for meta-analytic gray data requests is consistent with response rates previously identified in the literature (Hussey, 2023). Studies were eligible for analysis if they were available online, administered the English version of the PAI, reported means and standard deviations for honest and feigning (either simulation or criterion) groups, and used unique data (i.e., effect estimates were not previously used in another study). Data sources were not screened for inclusion of non-content-based invalid responding (i.e., elevations of the INC and INF scales).

Of the initial 63 studies identified, 20 were removed for not being publicly available, not providing data required to calculate Hedge's *g*, or containing duplicate data. When studies containing duplicate data emerged, the earlier study was selected for analysis. Two independent coders coded the remaining studies ($k = 43$) based on methodology (i.e., simulation vs criterion, coached vs naive, criterion measure used). Coaching was determined based on if a study 1) explicitly stated that participants were coached or 2) provided participants with knowledge regarding the PAI over-reporting detection methods and knowledge regarding symptoms relevant to the feigned condition (symptoms of PTSD, depression, etc.). When coding discrepancies emerged, which was uncommon, the first author reviewed the article to make the final determination with the senior author. Articles were divided into groups based on different feigning groups (e.g., feigned PTSD vs feigned depression) or honest conditions (e.g., student control vs clinical control). This decision was made to parse apart differences arising from contextual moderators (e.g., participant coaching, symptoms feigned, clinical control group). The final analyses included 78 studies across 43 articles with 6,451 participants (see supplemental materials).

Table 1 General random effects analyses for PAI over-reporting scales and supplemental indicators

Scale	*k*	*g*	*SE*	*p*	*95% CI*		I^2	H^2	*Q*	τ^2
NIM	71	1.50	0.11	< .001	1.29	1.71	94.21	17.28	801.13	0.75
MAL	56	1.07	0.10	< .001	0.88	1.26	92.74	13.77	540.08	0.47
RDF	52	0.99	0.12	< .001	0.76	1.22	95.14	20.58	717.82	0.68
NDS	16	1.81	0.29	< .001	1.25	2.37	97.15	35.04	637.88	1.25
MFI	8	1.72	0.28	< .001	1.17	2.27	95.57	22.55	187.70	0.60
HMI	7	1.68	0.45	< .001	0.81	2.56	97.77	44.93	285.02	1.36
CBS	8	0.86	0.07	< .001	0.72	1.00	44.51	1.80	12.55	0.02
CB-SOS-1	6	0.87	0.14	< .001	0.59	1.15	79.51	4.88	23.84	0.10
CB-SOS-2	5	0.84	0.13	< .001	0.58	1.10	72.01	3.57	11.08	0.06
CB-SOS-3	6	0.94	0.14	< .001	0.67	1.22	78.75	4.71	24.60	0.09
M	-	1.23	0.18	-	0.87	1.59	84.74	16.91	324.17	0.54
SD	-	0.38	0.11	-	0.26	0.57	15.96	13.67	303.76	0.46

NIM, Negative Impression Management; *MAL*, Malingering Index; *RDF*, Rogers Discriminate Function; *NDS*, Negative Distortion Scale; *MFI*, Multiscale Feigning Index; *CBS*, Cognitive Bias Scale; *CB-SOS*, Cognitive Bias Scale of Scales

Results

In general, results support PAI's over-reporting scales effective in detection of potential noncredible responding (Table 1). These effects were mostly large to very large in magnitude and generally produced large portions of explanatory variance (I^2). Effects ranged from *0.8*4 (CB-SOS2; 95% CI: 0.58—1.10) to 1.81 (NDS; 95% CI: 1.25 – 2.37), depending upon study design (e.g., simulation vs criterion) and criterion grouping (e.g., known groups vs performance validity failure). Below, we summarize general trends observed across all scales and then present individual results on a scale-by-scale basis.

Simulation studies mostly consisted of coached conditions (55%) using healthy controls (65%). Criterion studies mostly examined feigned neurocognitive dysfunction (75%) across Military and Veteran (39%; $k=7$, $n=824$), forensic (39%; $k=7$, $n=1{,}050$), neuropsychological outpatient (17%; $k=3$, $n=638$), psycho-educational assessment (17%; $k=3$, $n=153$), and inpatient (6%; $k=7$, $n=95$) contexts.[2] Large I^2 values were particularly common in simulation studies. Effect sizes and predicted variance amongst criterion studies were substantially lower relative to simulation-designs ($I^2=0.70.8$ [71%] vs. 91.39 [91%] s on average, respectively (Table 2). Simulation studies outperformed criterion studies ($M_{simulation}$ $g=1.67$, $M_{Criterion}$ $g=0.62$) but with greater variation in effect ($SD_{Simulation}$ $g=0.49$ vs $SD_{Criterion}$ $g=0.25$). In simulation studies, coached studies produced lower effects than uncoached ($M_{Coached}$ $g=1.42$ vs. $M_{Uncoached}$ $g=1.71$), with less variability ($SD_{Uncoached}$ $g=0.45$ vs $SD_{Coached}$ $g=0.24$; Table 3).

[2] Some studies contained multiple contexts (e.g., Active-Duty Military Personnel undergoing Medical Board Evaluations) and were coded as both.

Published studies tended to differ in notable ways from unpublished studies (Tables 4 and 5). First, published studies had slightly smaller effects on average ($g_{published}=1.16$ vs $g_{unpublished}$ 1.50). Second, unpublished studies were also more likely to include coached feigning conditions (70% vs 52%). Third, published studies had higher portions of explained variance ($I^2=93.33$ [93%] vs. 83.23 [83%] on average, respectively). Conversely, unpublished studies had a higher proportion of simulation studies (91%) to criterion (9%) compared to published studies (75% simulation to 25% criterion). Differences in bias between published and unpublished studies emerged when reviewing funnel and forest plots in the supplemental materials. These plots also demonstrated a difference between simulation and criterion designs, with criterion design exhibiting less bias than simulation. In these cases, published and simulation studies were biased toward larger effect sizes. Below, individual analyses for each validity scale are summarized separately.

Standard PAI Over-reporting Indices

Negative Impression Management Random effect analyses across 71 studies demonstrated a large effect (Fig. 1; $g=1.50$, $I^2=94.21$, $\tau^2=.75$). Larger effects were observed in simulation studies ($g=1.51$, $I^2=90.01$) compared to criterion designs ($g=0.74$, $I^2=65.28$; $g_{difference}=0.77$). Very little difference was evident between coached and naive designs.

Malingering Index MAL had a moderate effect (Fig. 2; $g=1.07$, $I^2=92.74$, $\tau^2=0.47$) across 56 studies. MAL performed better in simulation ($g=1.16$, $I^2=89.37$, $k=43$) than criterion designs ($g=0.59$, $I^2=67.06$, $k=13$;

Table 2 Fixed effects analyses for simulation and criterion designs

Scale	*k*	*g*	*SE*	*p*	*95% CI*		I^2	H^2
Simulation (77% of Studies)								
NIM	58	1.51	0.03	< .001	1.45	1.57	90.01	10.01
MAL	43	1.16	0.03	< .001	1.10	1.22	89.37	9.40
RDF	39	1.19	0.03	< .001	1.13	1.25	88.28	8.53
NDS	12	2.27	0.06	< .001	2.16	2.39	96.92	32.42
MFI	5	2.22	0.08	< .001	2.06	2.37	92.41	13.17
M	-	1.67	0.05	-	1.58	1.76	91.39	14.71
SD	-	0.49	0.02	-	0.45	0.52	3.07	8.99
Criterion (23% of Studies)								
NIM	13	0.74	0.05	< .001	0.66	0.83	65.28	2.88
MAL	13	0.59	0.05	< .001	0.50	0.68	67.06	3.04
RDF	13	0.17	0.04	< .001	0.08	0.25	71.11	3.46
NDS	4	0.68	0.08	< .001	0.53	0.82	70.31	3.37
MFI	3	0.92	0.09	< .001	0.75	1.09	80.24	5.06
M	-	0.62	0.06	-	0.50	0.73	70.80	3.56
SD	-	0.25	0.02	-	0.23	0.27	5.18	0.78

NIM, Negative Impression Management; *MAL*, Malingering Index; *RDF*, Rogers Discriminate Function; *NDS*, Negative Distortion Scale; *MFI*, Multiscale Feigning Index; *CBS*, Cognitive Bias Scale; *CB-SOS*, Cognitive Bias Scale of Scales

Table 3 Fixed effects analyses for uncoached and coached designs

Scale	*k*	*g*	*SE*	*p*	*95% CI*		I^2	H^2	*Q*	τ^2
Uncoached (45% of Studies)										
NIM	25	1.87	0.21	< .001	1.46	2.27	94.78	19.15	306.17	0.97
MAL	20	1.33	0.16	< .001	1.01	1.66	92.01	12.52	155.90	0.47
RDF	19	1.26	0.18	< .001	0.91	1.61	93.15	14.60	138.83	0.53
NDS	8	2.36	0.44	< .001	1.50	3.23	97.42	38.77	301.31	1.52
M	-	1.71	0.25	-	1.22	2.19	94.34	21.26	225.55	0.87
SD	-	0.45	0.11	-	0.26	0.66	2.03	10.39	78.44	0.42
Coached (55% of Studies)										
NIM	33	1.53	0.14	< .001	1.26	1.81	90.36	10.38	242.66	0.55
MAL	23	1.11	0.16	< .001	0.80	1.43	92.38	13.11	221.48	0.52
RDF	20	1.27	0.19	< .001	0.89	1.65	94.15	17.09	184.99	0.68
NDS	4	1.75	0.32	< .001	1.11	2.38	84.80	6.58	10.45	0.34
M	-	1.42	0.20	-	1.02	1.82	90.42	11.79	164.90	0.52
SD	-	0.24	0.07	-	0.18	0.35	3.51	3.84	91.52	0.12

NIM, Negative Impression Management; *MAL*, Malingering Index; *RDF*, Rogers Discriminate Function; *NDS*, Negative Distortion Scale

$g_{difference} = 0.57$) designs, with such small a difference suggesting generally equitable effects. Similarly, small differences exist between coached ($g = 1.11$, $I^2 = 92.38$; $\tau^2 = 0.52$, $k = 23$) and naive ($g = 1.33$, $I^2 = 92.01$, $\tau^2 = 0.47$, $k = 23$; $g_{difference} = 0.22$) designs.

Rogers Discriminant Function A moderate effect was observed across all 52 studies examining RDF (Fig. 3; $g = 0.99$, $I^2 = 95.14$, $\tau^2 = 0.68$). Differences in simulation ($g = 1.19$, $I^2 = 88.28$; $k = 39$) and criterion designs ($g = 0.17$, $I^2 = 71.11$; $k = 13$) differed substantially ($g_{difference} = 1.02$). There were negligible differences between coached and naive simulation designs ($g_{difference} = -0.01$), with both achieving detection rates within the expected standard range. Criterion design effect sides, however, fell in the negligible range, suggesting an inability to meaningfully differentiate clinical criterion groups. Slightly more than two percent of simulation ($k = 1$) studies and a notable 30.1% of criterion ($k = 4$) studies produced negative effect sizes (see Fig. 3),

Table 4 Fixed effects analyses for published and unpublished studies

Scale	*k*	*g*	*SE*	*p*	*95% CI*		I^2	H^2
Published (86% of Studies; 75% Simulation, 25% Criterion)								
NIM	60	1.22	0.03	< .001	1.17	1.28	91.47	11.72
MAL	49	0.96	0.03	< .001	0.91	1.01	90.54	10.57
RDF	47	0.82	0.03	< .001	0.77	0.88	93.23	14.78
NDS	13	1.63	0.05	< .001	1.53	1.73	98.07	51.93
M	-	1.16	0.03	-	1.10	1.22	93.33	22.25
SD	-	0.28	0.01	-	0.26	0.29	2.60	15.39
Unpublished (14% of Studies; 91% Simulation, 9% Criterion)								
NIM	11	1.81	0.09	< .001	1.64	1.98	85.41	6.85
MAL	7	1.17	0.09	< .001	1.00	1.34	77.85	4.51
RDF	5	1.11	0.10	< .001	0.92	1.31	86.79	7.57
NDS	3	1.91	0.15	< .001	1.61	2.20	82.88	5.84
M	-	1.50	0.11	-	1.29	1.71	83.23	6.20
SD	-	0.32	0.02	-	0.30	0.35	3.05	1.03

NIM, Negative Impression Management; *MAL*, Malingering Index; *RDF*, Rogers Discriminate Function; *NDS*, Negative Distortion Scale

Table 5 Differences in Hedge's g values between conditions

	Simulation vs Criterion	Naïve vs Coached	Published vs Unpublished
NIM	.77	.33	-.59
MAL	.57	.22	-.21
RDF	1.02	-.01	-.29
NDS	1.60	.62	-.28
MFI	1.30	–	–

Directionality noted in column tile; *NIM*, Negative Impression Management; *MAL*, Malingering Index; *RDF*, Rogers Discriminate Function; *NDS*, Negative Distortion Scale; *MFI*, Multiscale Feigning Index

meaning that honest responders scored *higher* on RDF than their respective feigning conditions.

Supplemental Over-reporting Indicators

Negative Distortion Scale Across 16 studies, NDS demonstrated an overall large effect (Fig. 4; $g = 1.81$, $I^2 = 97.15$, $\tau^2 = 1.25$). Simulation studies ($g = 2.27$, $I^2 = 96.92$, $k = 12$) greatly outperformed criterion designs ($g = 0.68$, $I^2 = 70.31$, $k = 4$; $g_{difference} = 1.6$). NDS demonstrated large effects across coached ($g = 1.75$, $k = 4$) and naive designs ($g = 2.36$; $k = 8$).

Multiscale Feigning Index MFI demonstrated generally large effects (Fig. 5; $g = 1.72$, $I^2 = 95.57$, $\tau^2 = 0.6$, $k = 8$). A large difference was observed between simulation and criterion designs ($g_{difference} = 1.3$) with simulation studies evidencing a very large effect ($g = 2.22$, $I^2 = 92.41$) and criterion designs evidencing a large effect ($g = 0.92$, $I^2 = 80.24$).

Hong Malingering Index A generally large effect was observed in HMI (Fig. 6; $g = 1.68$, $I^2 = 97.77$, $\tau^2 = 1.36$, $k = 7$). All but one study featuring the HMI used a criterion design. The six simulation studies demonstrated a large effect ($g = 1.89$, $I^2 = 97.54$, $\tau^2 = 1.28$).

Cognitive Bias Scale The CBS scale demonstrated a moderate effect in the random effect analysis (Fig. 7; $g = 0.86$, $I^2 = 44.51$, $\tau^2 = 0.02$, $k = 8$). Only one uncoached simulation study using CBS was identified, and this study produced a moderate effect ($g = 0.85$). A fixed effect analysis on the seven criterion studies demonstrated a moderate effect ($g = 0.85$, $I^2 = 52.19$).

Cognitive Bias Scale of Scales-1 The CB-SOS-1 scale demonstrated broadly moderate effects (Fig. 8; $g = 0.87$, $I^2 = 79.51$, $\tau^2 = 0.1$, $k = 6$). Notably, all but one study used a criterion design. The one uncoached simulation study showed a large effect ($g = 1.35$). The other five criterion designs evidenced a small effect ($g = 0.73$, $I^2 = 64.22$).

Cognitive Bias Scale of Scales-2 Five criterion designs were identified for the CB-SOS-2 scale. Random effect analysis of these studies showed a moderate effect (Fig. 9; $g = 0.84$, $I^2 = 72.01$, $\tau^2 = 0.06$). Notably, no simulation studies examining the CB-SOS-2 scale were found.

Cognitive Bias Scale of Scales-3 Moderate effects were observed on the CB-SOS-3 scale (Fig. 10; $g = 0.94$, $I^2 = 78.75$, $\tau^2 = 0.09$, $k = 6$). Five of those studies used a criterion design and evidenced a small effect ($g = 0.78$, $I^2 = 61.48$). The only simulation study identified used an uncoached simulation design, which showed a large effect ($g = 1.45$).

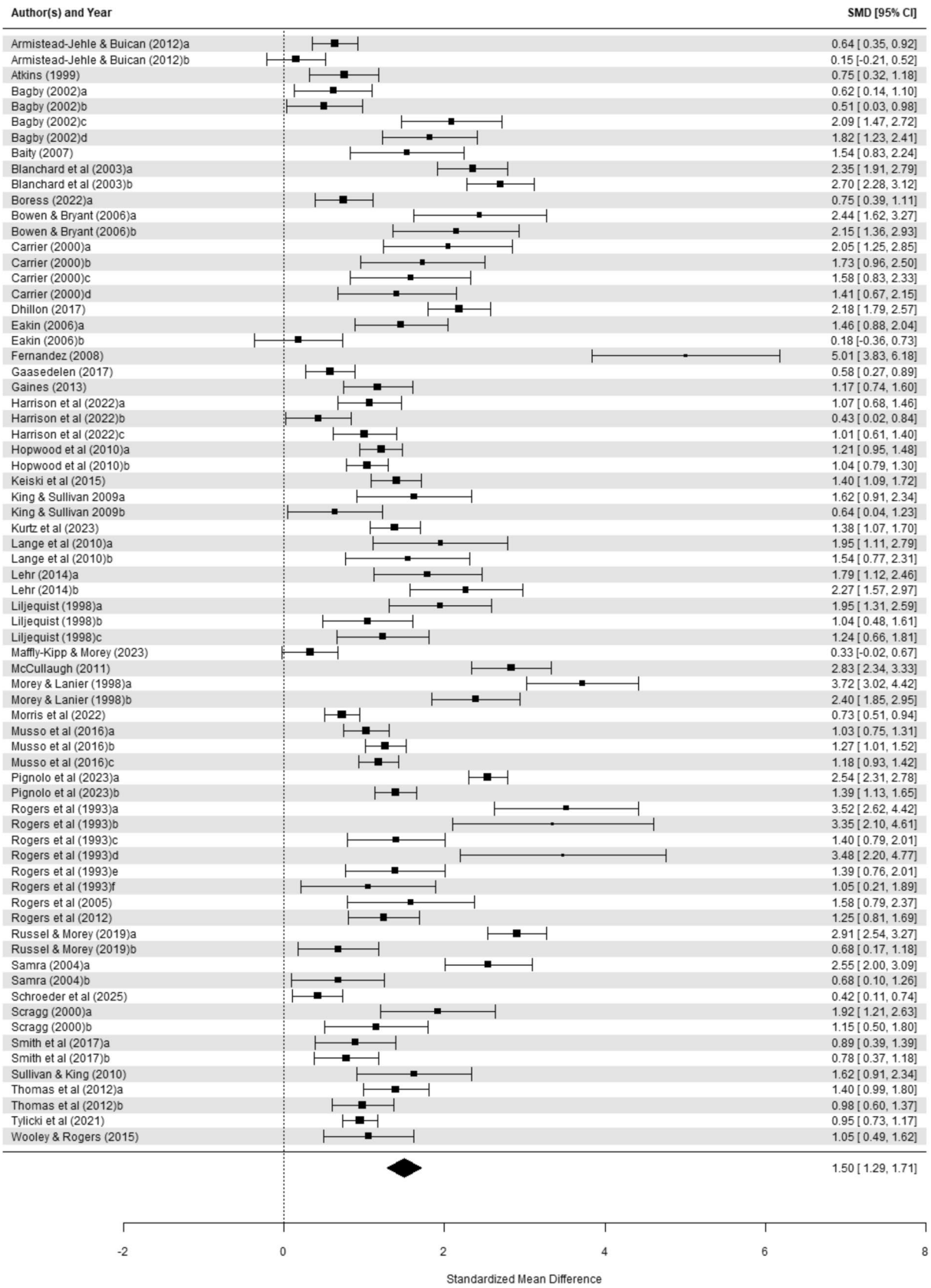

Fig. 1 NIM random effects forest plot

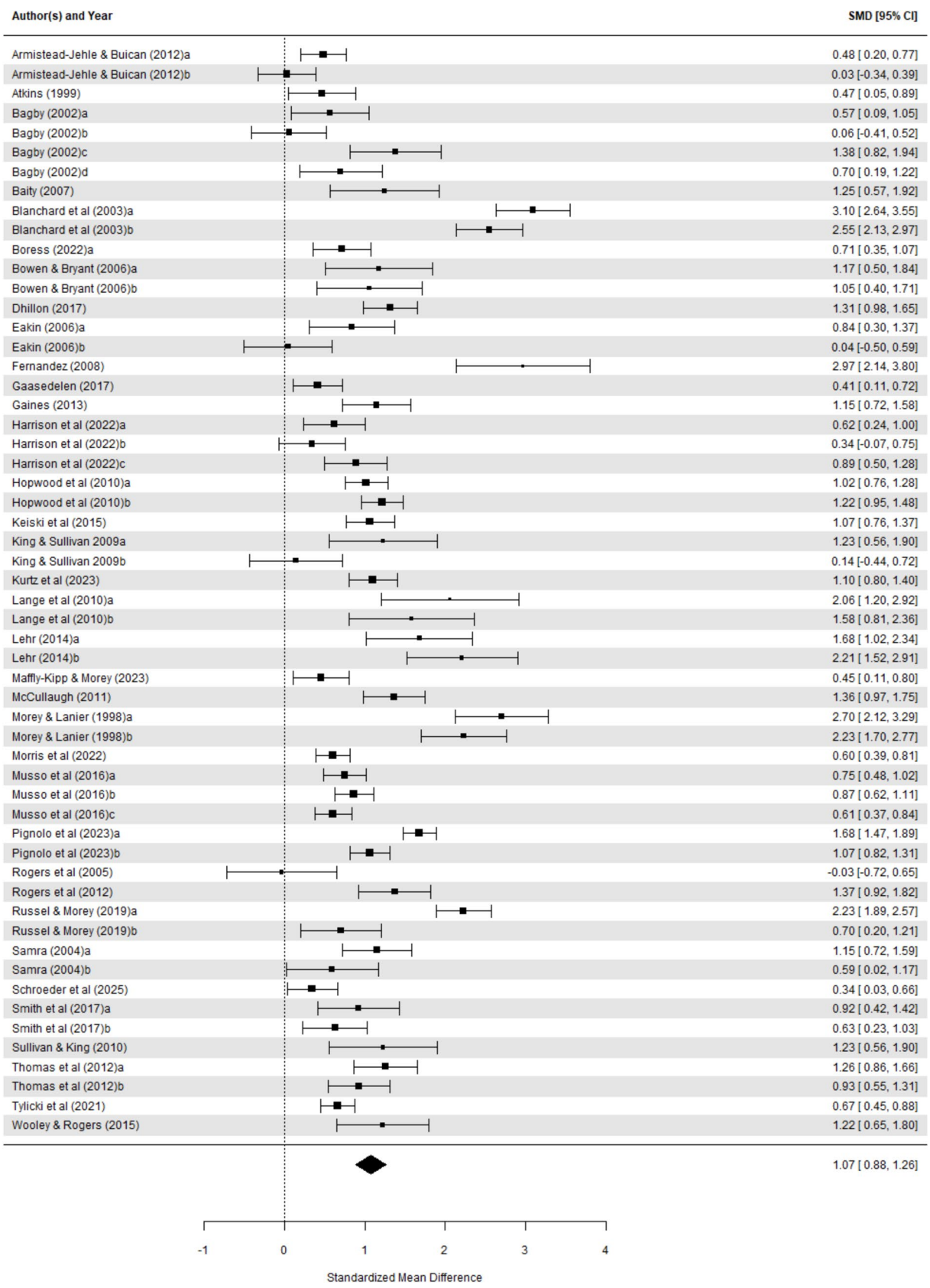

Fig. 2 MAL random effects forest plot

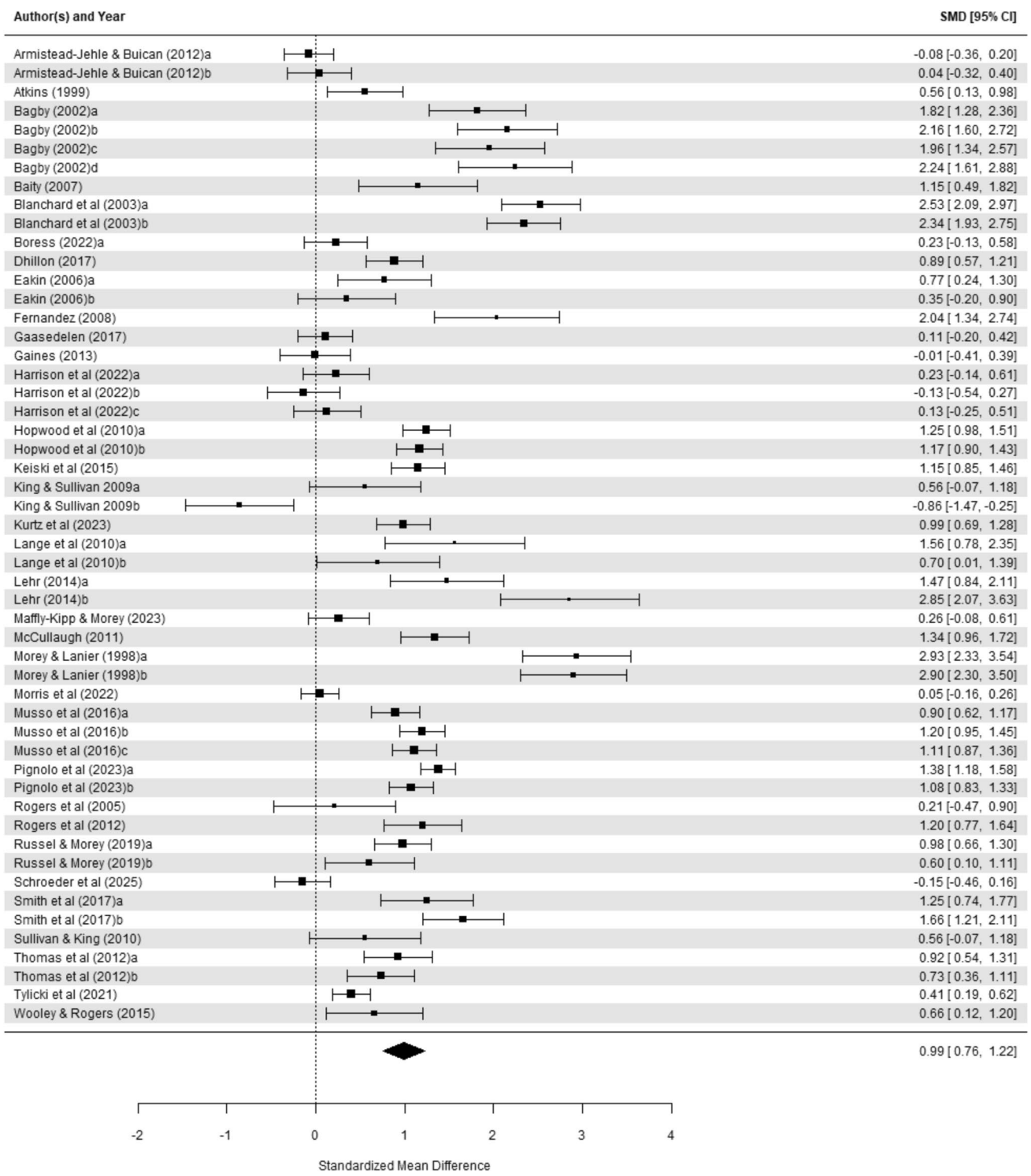

Fig. 3 RDF random effects forest plot

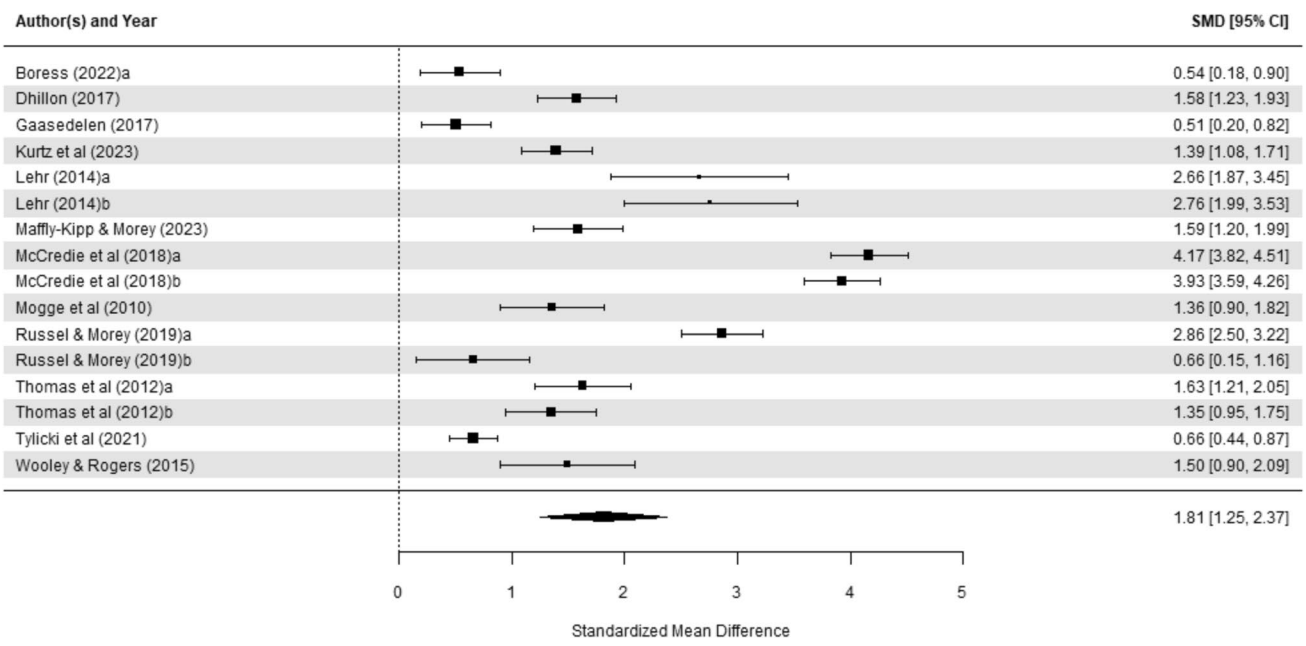

Fig. 4 NDS random effects forest plot

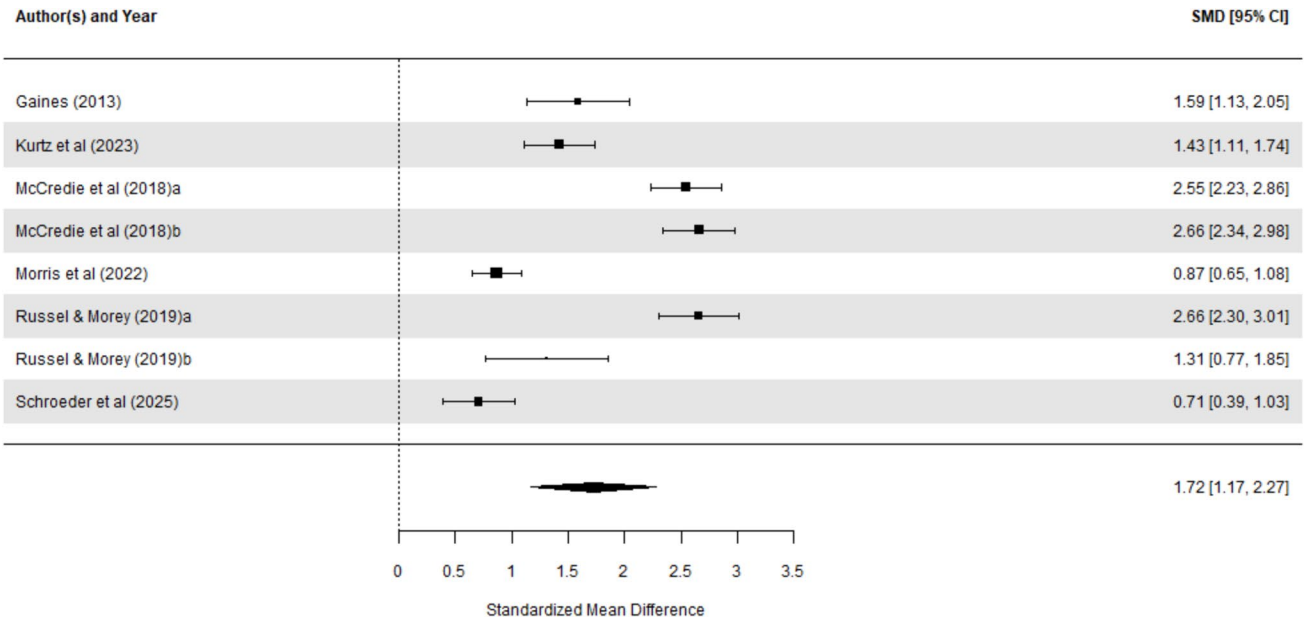

Fig. 5 MFI random effects forest plot

Sensitivity and Specificity

Classification statistics for commonly reported cut-scores (Table 6) ranged, with most scores demonstrating adequate specificity (SP range = 0.80 [RDF ≥ 0] to 0.99 [MAL ≥ 111]) and varying sensitivity (SN range = 0.18 [CBS ≥ 21] to 0.60 [RDF ≥ 0]). However, reported cut-scores varied across studies for many indicators, prohibiting some classification statistics. Only four analyzed studies, for example, included sensitivity and specificity for the NDS, and they had little overlap in reported cut-values in those classification analyses.

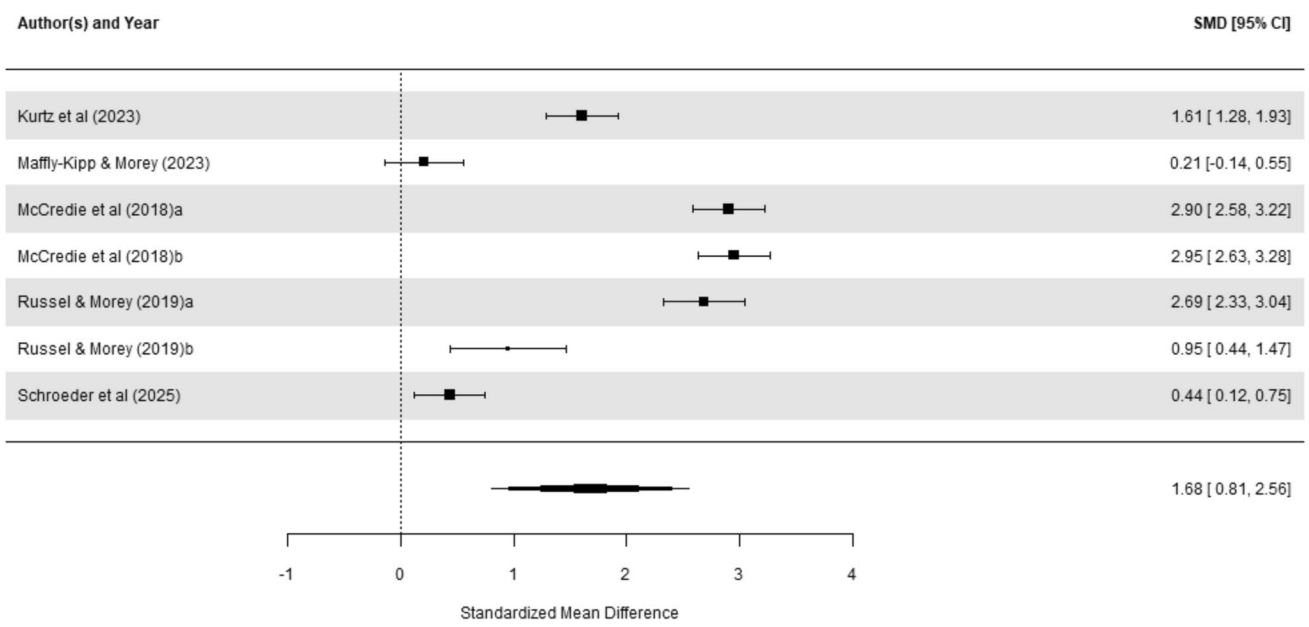

Fig. 6 HMI random effects forest plot

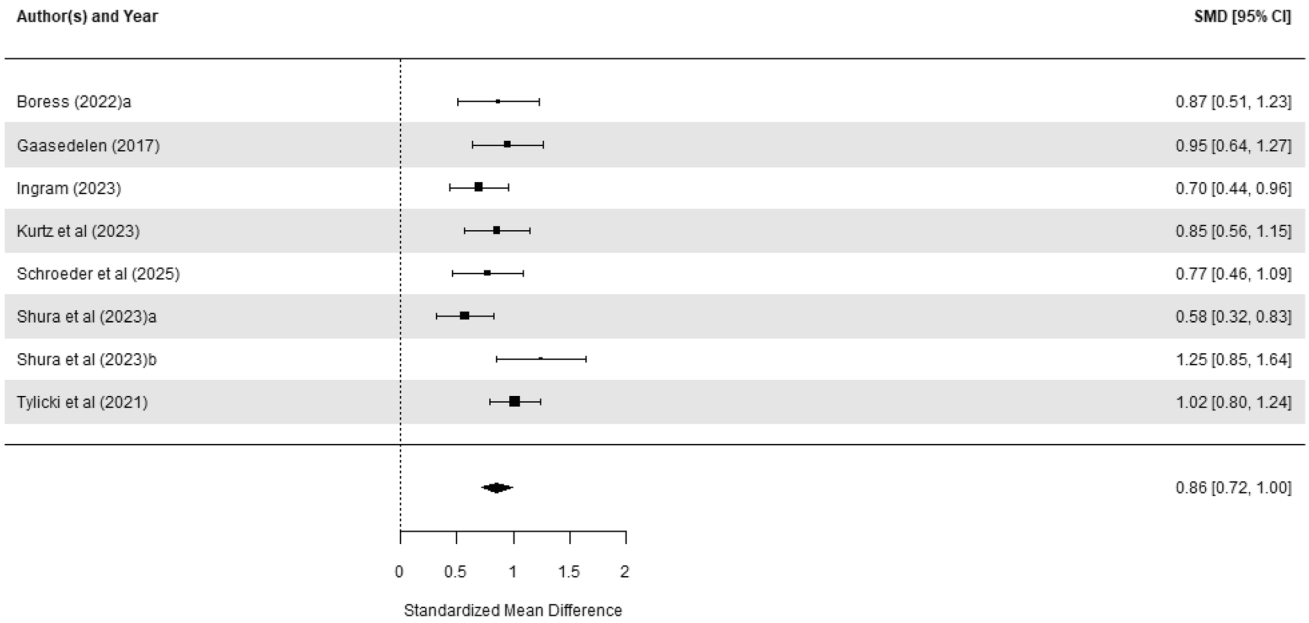

Fig. 7 CBS random effects forest plot

Differences from Hawes and Boccaccini (2009) on NIM, MAL, and RDF

There were several differences between NIM, MAL, and RDF's performance in the current study and their performance in Hawes and Boccaccini's (2009) meta-analysis.[3] Overall differences were mostly negligible (see Table 7). In the simulation condition, effects are smaller in the present analyses than observed previously, although the most pronounced difference was observed in RDF ($\Delta ES = -0.5$). This study also observed smaller effect sizes across criterion designs on NIM ($\Delta ES = -0.32$). MAL ($\Delta ES = -0.35$), and RDF ($\Delta ES = -0.14$). Sensitivity and specificity patterns are generally the same as those observed by Hawes and Boccaccini (2009), with minor decreases in scale sensitivity (NIM > 92 = 0.37 vs 0.59) and corresponding increases in

[3] Cohen's *d* was used by Hawes & Boccaccini (2009). In this study, we use Hedge's *g* for mean estimation rather than Cohen's *d*. While they are similar, Hedge's *g* is preferred in meta-analysis as it adjusts for sample size. This difference in approach may impact reliably of Effect Size (ES) change.

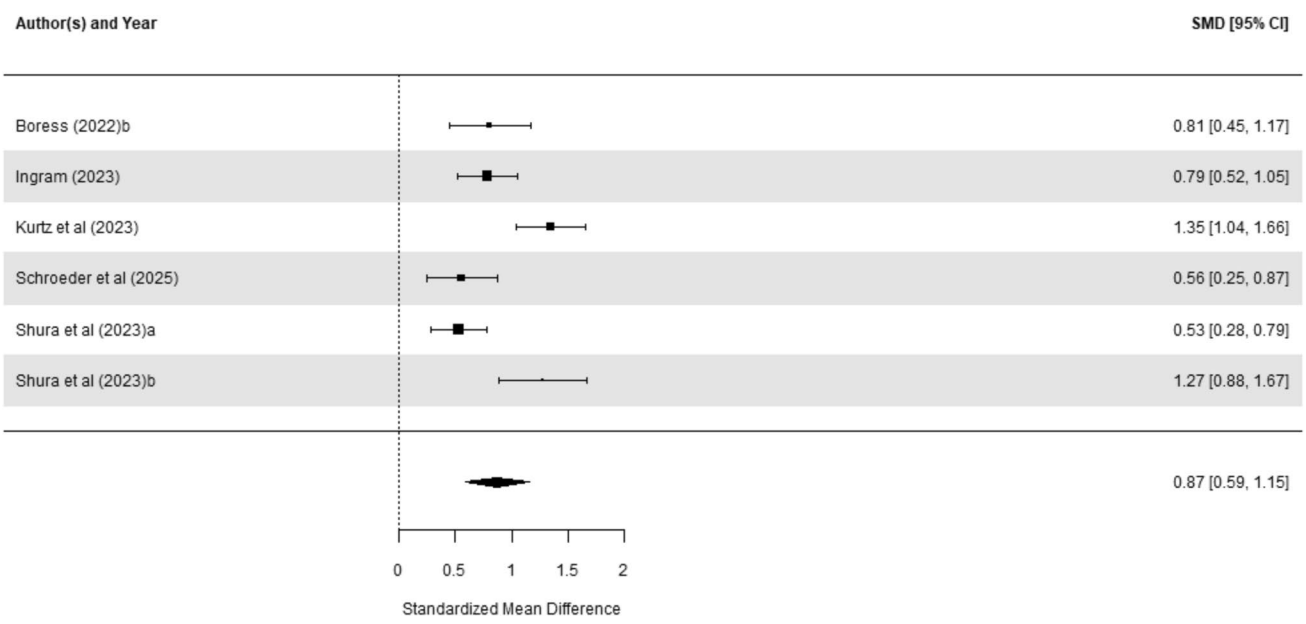

Fig. 8 CB-SOS-1 random effects forest plot

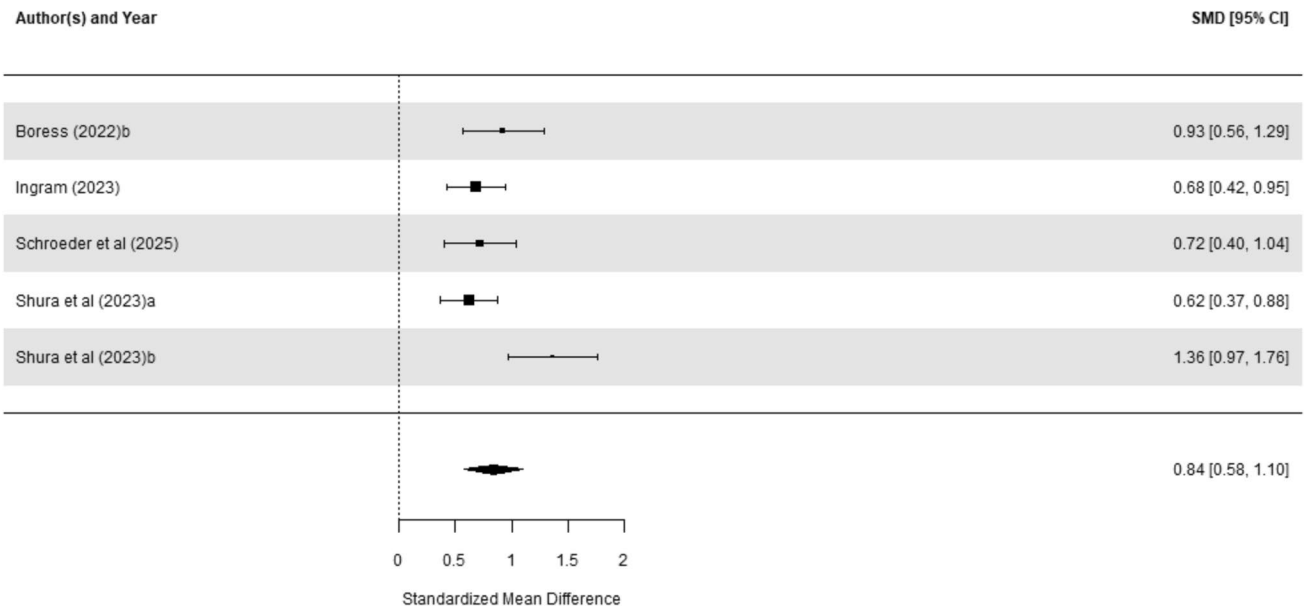

Fig. 9 CB-SOS-2 random effects forest plot

specificity (NIM > 92 = 0.95 vs 0.80) in this study compared to theirs, respectively (Figs. 4, 5, 6, 7, 8, 9, and 10).

Discussion

This study provides an updated and comprehensive meta-analysis and research synthesis of the Personality Assessment Inventory (PAI)'s over-reporting scales. It includes 78 effect sizes across 43 articles and covers both standard (i.e., NIM, MAL, RDF) and supplemental (i.e., NDS, MFI, CBS, SOS, and HMI) scales. Methodologically, we compared scale mean scores using a series of random and fixed effect meta-analyses for each scale and across study-coded criterion groups developed based on historic moderating influences (e.g., published/unpublished, simulation/criterion design, coached and uncoached simulation design). We also compared effects observed to those of Hawes and

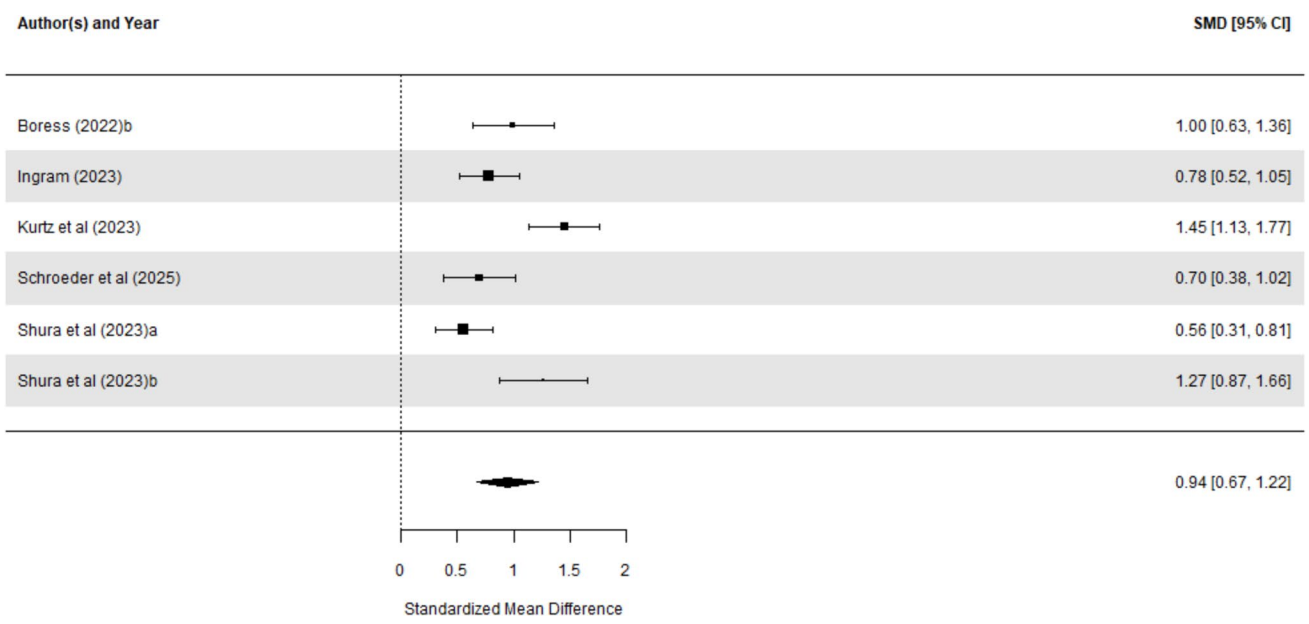

Fig. 10 CB-SOS-3 random effects forest plot

Table 6 Cut-scores for standard PAI over-reporting scales

		Sensitivity		Specificity	
	k	*M*	*SD*	*M*	*SD*
NIM					
73	8	.51	.21	.87	.06
84	10	.39	.21	.93	.10
92	15	.37	.18	.95	.06
110	6	.27	.16	.99	.01
MAL					
84	17	.40	.21	.95	.04
111	9	.19	.15	.99	.00
RDF					
0	13	.60	.24	.80	.06
CBS					
16	5	.39	.13	.85	.08
19	7	.29	.12	.93	.03
21	5	.18	.09	.96	.01

NIM, Negative Impression Management; *MAL*, Malingering Index; *RDF*, Rogers Discriminate Function; *CBS*, Cognitive Bias Scale

Boccaccini (2009) and computed average sensitivities and specificities across scales. In addition to the originally examined Negative Impression Management (NIM; $k = 71$), Malingering Index (MAL; $k = 56$), and Rogers Discriminant Function (RDF: $k = 52$) validity indices, this study analyzes the Negative Distortion Scale (NDS; $k = 16$), Multiscale Feigning Index (MFI; $k = 8$), Hong Malingering Index (HMI; $k = 7$), Cognitive Bias Scale (CBS; $k = 8$), and the Cognitive Bias Scale of Scales (CB-SOS; $k = 5$—6).

Table 7 SMD differences between this study and Hawes and Boccaccini (2009)

	Current Study		Hawes and Boccaccini (2009)		
Type	*k*	*g*	*k*	*d*	*SMD Difference*
NIM					
Overall	71	1.50	23	1.48	.02
Simulation	58	1.51	16	1.68	-.17
Criterion	13	0.74	7	1.06	-.32
MAL					
Overall	56	1.07	19	1.15	-.08
Simulation	43	1.16	12	1.27	-.11
Criterion	13	0.59	7	0.94	-.35
RDF					
Overall	52	0.99	15	1.13	-.14
Simulation	39	1.19	9	1.69	-.50
Criterion	13	0.17	6	0.31	-.14

SMD, standardized mean difference; SMD differences were calculated by subtracting Hawes and Boccaccini (2009) Cohen's *d* values from this study's Hedge's *g* values; *SMD*, Standardized Mean Difference; *NIM*, Negative Impression Management; *MAL*, Malingering Index; *RDF*, Rogers Discriminate Function

We hypothesized (1) each of the PAI validity scales' group difference magnitudes would fall within the standard effect range (0.75 to 1.75), (2) simulation studies would outperform criterion studies by between half and full standard deviation (0.50 to 1.00), and (3) coached simulation designs will more closely approximate criterion groups by a moderate magnitude of effect (approximately 0.30 effect difference).

Our results support these hypotheses as well as the over-reporting scales of the PAI at their detection over-reporting. Specifically, we found that (1) scales generally differentiate between over-reporting and honest responders, (2) these effects were larger in simulation than criterion designs, (3) coached simulation designs more approximated effects seen in criterion designs than uncoached designs, (4) published studies demonstrated smaller effects than unpublished, and (5) these effects are slightly different in magnitude than those observed in Hawes and Boccaccini's (2009) meta-analysis, demonstrating smaller effects overall. These findings are discussed for the PAI over-reporting scales, as well as broader self-report symptom validity detection theory.

Personality Assessment Inventory

Overall, the PAI over-reporting scales and supplemental indicators (i.e., NDS, MFI, CBS, CB-SOS, HMI) fell within Rogers and Bender's (2018) mean-analyzed (e.g., *Cohen's d* or *Hedge's g*, for instance) range of moderate to large. Published studies' effect sizes were also mildly smaller than unpublished studies on all scales, $g_{difference} = 0.21$—0.59. However, published studies paradoxically demonstrated more bias toward larger effects than unpublished studies (see supplemental figures). These minor differences may result from study characteristics (i.e., higher proportion of simulation designs in unpublished studies) as well as fewer unpublished studies analyzed ($k_{unpublished} = 11$ vs $k_{published} = 67$).

Criterion and Simulation Designs Simulation and criterion studies provide data on the efficacy (controlled design) and effectiveness (clinical application), respectively, of the PAI over-reporting scales. As expected, studies differed based on their design characteristics. Simulation studies demonstrated larger effects than criterion studies, with difference ranges from moderate (MAL $g_{difference} = 0.57$) to large (NDS $g_{difference} = 1.6$). This trend is consistent with past research on the PAI (Hawes & Boccaccini, 2009). Criterion-grouped studies were also less accurate in capturing constructs of interest (e.g., explained variance) relative to simulation designs ($I^2 = 71\%$ vs 91% in simulation designs). The methodological distinction between PVT and SVT criterion could play a role (Larrabee, 2012; Ord et al., 2021), with the former serving as a staple of criterion group classification.

Criterion studies examined mostly individuals with suspected neurocognitive dysfunction or were part of standard neuropsychological evaluations. Such over-sampling of a single evaluative context may also influence scale effectiveness due to base rates of symptoms. For instance, one might expect that assessment of over-reported cognitive impairment would be well assessed, whereas severe psychopathology may not (see Sweet et al., 2021). Prior studies have often found negligible to small effect sizes differentiating types of impairment on broadband measures (Ingram et al., 2024, Morris et al., 2022). Small differences are unsurprising given that PAI over-reporting scales generally share 50% (or more) of their variance, regardless of construction method or domain assessed in those scales (Shura et al., In press).

Criterion studies have an advantage over simulation studies—they assess genuine versus feigned symptomology, which more accurately mirrors genuine malingering (i.e., increased external validity; Rogers & Bender, 2018, p. 594). Indeed, only 35% of simulation studies used a clinical control group. This distinction can be found in funnel plots (found in the supplemental materials), wherein simulation studies demonstrate bias for larger effects not evident in criterion designs. This trend suggests that, while the PAI's over-reporting indices are greatly efficacious, their effectiveness in clinical settings is more tempered, particularly for RDF which failed to produce evidence of utility outside of simulation designs.

The large differences between simulation and criterion designs affect several scales in the PAI, and the ratio of simulation studies to criterion studies may explain overall differences in indicator performance (i.e., the higher the simulation to criterion ratio, the larger the overall effect size). For example, RDF has a moderate effect ($g = 0.99$) overall, but it does not effectively differentiate between groups in criterion studies ($g = 0.17$). The proportion of simulation studies (79%) comprising this meta-analytic estimate also likely inflates its performance. Conversely, most studies examining the CBS and CB-SOS indicators are criterion designs with scales resting consistently within a narrow, moderate sized effect range, $g = 0.84$ (CB-SOS-2) to $g = 0.94$ (CB-SOS-3). After excluding the single simulation designs, CBS and the CB-SOS scales (except CB-SOS1) evidenced larger differences in criterion designs over the standard PAI over-reporting indicators ($g = 0.78$ [CB-SOS-3] to $g = 0.85$ [CBS] vs $g = 0.17$ [RDF] to $g = 0.74$ [NIM]). This strong criterion relationship is not surprising given that the CB-SOS scales were developed using a bootstrapped method to predict PVT failure (Burchett & Bagby, 2022), often what was also predicted in those studies (Boress et al., 2022a, 2022b; Gaasedelen et al., 2019). In our analysis, MFI ($g = 0.92$; Table 2) performed the best across criterion designs for all PAI over-reporting scales.

Coached and Uncoached Designs Effect sizes were generally larger in uncoached designs ($g_{difference} = -0.01$ [RDF] to 0.62 [NDS]), a pattern consistent with past moderated random effect models in other broadband measures (e.g., Ingram & Ternes, 2016). Interestingly, the variance explained was slightly larger in the uncoached conditions ($I^2 = 94\%$ vs 90%); however, sample variance

was larger in this condition compared to coached designs ($\tau^2 = 0.87$ vs 0.52, respectively). Such differences in variation reflect the nature of how coaching shapes response patterns (Veltri & Williams, 2013), such as when participants are informed of the scales and how to feign on them. Uncoached designs may outperform coached in overall effect while demonstrating more variance because simulators use diverse feigning approaches and are theoretically less sophisticated than someone informed about the test. Unfortunately, there is limited research on participant-centered perceptions during feigning. Studies focused on participant described feigning methods (e.g., what strategy did you take, how did you decide which items to endorse, what type of information did you want to convey/hide, etc.) would be a fruitful and timely addition to the literature. Additionally, constructs such as insight, mental health literacy, as well as emotion awareness and regulation may be relevant covariates and may explain endorsement approaches themselves.

Comparison to Prior Meta-Analysis The last meta-analysis on the PAI over-reporting indices was conducted by Hawes and Boccaccini (2009). Due to differences in inclusion criteria (e.g., including data not publicly available) and study availability, the current study only included 13 of the original 26 studies they analyzed. Despite these differences, findings from this study align with trends observed in their meta-analysis of the standard PAI over-reporting scales. Indeed, minimal differences between effect sizes across scales and conditions were observed between this study and Hawes and Boccaccini's (2009) original investigations. This study showed mild decreases in standardized mean difference[4] effects on NIM ($\Delta ES = 32$) and MAL ($\Delta ES = 35$) in criterion studies as well as RDF's performance in simulation designs ($\Delta ES = 50$). These differences reflect the increase in study sample size (k) as well as differing study demographics between meta-analyses. Notably, despite these small differences, effect sizes remain in the *over-reporting standard effect range* across PAI both meta-analyses.

Method trends since Hawes and Boccaccini (2009) have largely remained unchanged, despite recommendations and broader SVT development discussions (e.g., Burchett & Ben-Porath, 2019; Whiteside & Basso, 2024). One issue that occurs across meta-analytic studies, as noted by Hawes and Boccaccini (2009), is that standard cut-off values are not always evaluated. Failing to provide standardized cut scores reduces information available for standard clinical decisions, as well as reduces information on extreme scores—which tend to be of the most interest in scale interpretation. Since their paper, reporting of standard cut scores (as provided in Morey, 1991, 2007) remains uncommon. For example, only 19% of studies ($k = 15$) examining NIM included sensitivity and specificity values at the recommended cut score of 92 (Table 2). Posterior probabilistic statistics (e.g., positive and negative predictive power) are also not regularly presented for contextual information about these scales' utility (see Leonhard, 2023a). Additionally, most studies use stringent comparisons (e.g., pass all versus fail 2 + PVTs) that eliminate ambiguity commonly observed in clinical practice among overreporting while inflating the observed effect sizes (Fa). Mixed clinical samples also complicate interpretation for criterion designs because it creates a false dichotomy (i.e., honest versus uncredible) (Faust, 2023), ignoring the continuum of effort and intention in feigning research (Aita & Hill, 2022).

Trends and Future Directions of Self-Report Symptom Validity Detection

Meta-analytic effects observed in this study generally follow effect size trends seen in other broadband personality instruments (Ingram & Ternes, 2016; Sharf et al., 2017). The PAI scales in our study appear to perform generally equitably to one another (particularly in the case of criterion evaluation), suggesting that the scales capture a broad trait (e.g., response bias) rather than specific domains or approaches (*c.f.,* Sherman et al., 2020). Such a possibility is supported by scale-level analysis (Gaines et al., 2013; Ingram et al., 2024), multivariate analysis (Aita et al., 2022; Aita et al., in press), and population level correlation patterns (Shura et al., In press). This similarity in performance highlights several substantial methodological and statistical issues in validity testing (see Leonhard, 2023a, 2023b, 2023c), as well as providing direction for growth. While not revisiting those criticisms, or responses to them (e.g., Bush, 2023; Young & Erdodi, 2024), we discuss three specific recommendations for symptom validity research.

First, researchers should focus on growing mechanistic research within experimental designs (Halligan et al., 2003). This approach might help explain the notable differences between simulation and criterion and target underlying motivational and test engagement processes. While there are decades of meta-analyses on personality assessment over-reporting, there is limited experimental and methodological advancement. For instance, only a single study of the PAI or MMPI-family of measures has used a between-within design in the last 20 years to our knowledge (Baity et al., 2007). These designs may help explain

[4] Hawes & Boccaccini (2009) used Cohen's d whereas this study used Hedge's g, which is preferred in meta-analysis as it adjusts for sample size. This difference in approach may impact reliably of Effect Size (ES) change.

discrepancies of performance observed with RDF in this study or help contextualize longstanding test bias issues such as an over-identification of minoritized individuals (e.g., Geisinger et al., 1998). Additionally, few criterion studies were identified outside neuropsychological settings, which heavily weighs performance-based criterion measures in examining scale effectiveness (73% of criterion studies in the current study examined cognitive dysfunction and used PVT failure as the criterion). These limitations in known group designs examining psychiatric, somatic, or other impairment limits knowledge of a scale's clinical effectiveness (Rogers & Bender, 2018, p. 594). Likewise, there is a lack of research examining what, how, and when (if at all) feigners incorporate coaching to achieve their intended goal. For example, the mainstreaming of Artificial Intelligence can impact the test information that is available to clients, and limitations in test security (e.g., accessibility of item content, scale construction and scoring, etc.) and research-synthesized coaching may impact validity. These are, in our view, several of the more critical ways in which current SVT research approaches fail to capitalize on the strength of experimental design.

Second, it would be wise to consider broader patterns that emerge across the substantive scales, as well as between validity scales. The growth of scale-based infrequency approaches offers a pivotal approach to improving SVT practice, regardless of if conducted at the univariate or multivariate level. Emerging multivariate base rate research has similarly shown its effectiveness on the PAI (Aita et al., in press). Interestingly, sensitivities using the multivariate base rate approach were higher than the standard range seen for individual scales (e.g., 0.10 to 0.50). Within this expanded infrequency framework, scale level indicators could include not only the probability of individual item endorsement but also frequencies of floor and ceiling effects (Aita et al., in press), elevation rates, or mean scores (Ingram et al., 2024). The MFI follows the assumption that "*Individuals tend to report numerous symptoms across a wide variety of diagnostic categories rather than focusing on a particular diagnostic category*" (Gaines et al., 2013; p. 439). Thus, elevated validity scales are likely to co-occur on the PAI, likely as a function of the substantial portion (~ 50%) of shared variance (Schroeder et al., 2025; Shura et al., In press) instead of as a function of domain-specific presentation. Future studies should test meaningful differences between over-reporting scales, particularly when those scales claim to measure distinct and theorized domains of functioning (e.g., cognitive-focused symptoms, such as targeted by CBS and CB-SOS). While there are no agreed upon metrics about the size of effect required to effectively differentiate these domains, such distinctions are important to investigate in SVT research. For instance, while medium effect differences are a standard threshold for clinical significance (e.g., Ferguson, 2009), small effects may also be meaningful in certain types of decisions (Bakker et al., 2019). Determining the meaningful size of such comparisons is beyond the scope of this paper, but careful determination of these needs is critical to help refine symptom validity theory.

Lastly, adherence and standardization of reporting metrics will also improve SVT research. Hawes and Boccaccini (2009) recommended a standard inclusion of cut scores for classification accuracy and this call remains as critical now as it was then. Without these specific cut-points to analyze, meta-analytic findings rely on a mean-centered approach (e.g., on average, is group 1 higher on a scale than group 2), but such approaches produce substantial individual variability (e.g., standard deviation) and fail to provide explicit information about other scale distributional qualities. Given that means become less reliable as the underlying distribution loses its normality, meta-analytic methods should be applied to classification accuracies as current averaging approaches do not provide a population-weighted metric. The PAI appears well positioned to continue contributing to these questions and advancing methodological approaches and statistical considerations for self-report symptom validity detection.

Conclusions

Overall, all PAI standard over-reporting scales and supplemental indices except RDF are effective in discriminating between credible and non-credible responders across contexts and study designs. Notable differences emerged between simulation and criterion designs as well as published and unpublished studies. These trends follow patterns observed in the last meta-analysis on the standard over-reporting scales conducted 15 years ago (Hawes & Boccaccini, 2009). Moreover, mostly equitable effect sizes across PAI scales with one another and other broadband over-reporting measures (i.e., the MMPI-2-RF) support a broad symptom endorsement strategy of over-reporting–consistent with research on multivariate base rates (Aita et al., 2022; Ingram et al., 2024). These findings underscore the need to continue discussions around statistical and methodological issues in validity assessment (Leonhard, 2023a, 2023b, 2023c), standardize reporting of common cut-scores and classification statistics (Hawes & Boccaccini, 2009 for their recommendations), and further examine these indices in real-world contexts.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10862-025-10233-9.

Funding There was no funding for this project.

Data Availability Data are available upon reasonable request.

Declarations

Conflict of interest Dr. Paul Ingram has received research support from various test publishers or distributors, including the University of Minnesota Press and Pearson Clinical Assessment. He also served as a paid advisory board member for PAR Inc on the Personality Assessment Inventory (PAI). No support of any type was received for this project. The authors report there are no other competing interests to declare.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Aikman, G. G., & Souheaver, G. T. (2010). Use of the Personality Assessment Inventory (PAI) in neuropsychological testing of psychiatric outpatients. *Applied Neuropsychology, 15*(3), 176–183. https://doi.org/10.1080/09084280802324283

Aita, S.L., Hill, B.D. (2022). Effort is more than suboptimal: Positive aspects of motivation and engagement in neuropsychological assessment. In: Randolph, J.J. (eds) *Positive Neuropsychology*. Springer. https://doi.org/10.1007/978-3-031-11389-5_6

Aita, S. L., Montgomery, E., Caron, J. E., Pagano, L. A., Broggi, M., Ingram, P. B., Erickson, S. C., Borgogna, N. C., Moncrief, G. G., Roth, R. M., Calamia, M., Armistead-Jehle, P., & Hill, B. D. (in press). Multivariate base rates of standard- and skyline-cutoff elevations on the Personality Assessment Inventory: Do they distinguish simulated from genuine PTSD? Manuscript accepted in Journal of Personality Assessment.

Aita, S. L., Holding, E. Z., Greene, J., Carrillo, A., Moncrief, G. G., Isquith, P. K., & Roth, R. M. (2022). Multivariate base rates of score elevations on the BRIEF2 in children with ADHD, autism spectrum disorder, or specific learning disorder with impairment in reading. *Child Neuropsychology, 28*(7), 979–996. https://doi.org/10.1080/09297049.2022.2060201

Anestis, J. C., Rodriguez, T. R., Preston, O. C., Harrop, T. M., Arnau, R. C., & Finn, J. A. (2021). Personality assessment and psychotherapy preferences: Congruence between client personality and therapist personality preferences. *Journal of Personality Assessment, 103*(3), 416–426.

Arbisi, P. A., & Beck, J. G. (2016). Introduction to the special series "empirically supported assessment." *Clinical Psychology: Science and Practice, 23*(4), 323.

*Armistead-Jehle, P., & Buican, B. (2012). Evaluation context and Symptom Validity Test performances in a US military sample. *Archives of Clinical Neuropsychology, 27*(8), 828–839. https://doi.org/10.1093/arclin/acs086

Armistead-Jehle, P., Ingram, P. B., & Morris, C. S. (2020). Personality assessment inventory cognitive bias scale: Validation in a military sample. *Archives of Clinical Neuropsychology, 35*(7), 1154–1161.

*Atkins, D. G. (1999). *Validity of the Personality Assessment Inventory for detecting malingering of psychosis in a prison population*. (Doctoral Dissertation), The Fielding Institute.

*Bagby, R. M., Nicholson, R. A., Bacchiochi, J. R., Ryder, A. G., & Bury, A. S. (2002). The predictive capacity of the MMPI-2 and PAI validity scales and indexes to detect coached and uncoached feigning. *Journal of Personality Assessment, 78*(1), 69–86. https://doi.org/10.1207/S15327752JPA7801_05

Baity, M. R., Siefert, C. J., Chambers, A., & Blais, M. A. (2007). Deceptiveness on the PAI: A study of naive faking with psychiatric inpatients. *Journal of Personality Assessment, 88*(1), 16–24. https://doi.org/10.1207/s15327752jpa8801_03

Bakker, A., Cai, J., English, L., Kaiser, G., Mesa, V., & Van Dooren, W. (2019). Beyond small, medium, or large: Points of consideration when interpreting effect sizes. *Educational Studies in Mathematics, 102*, 1–8.

Bergquist, B. K., Keen, M. A., Ingram, P. B., Morris, N. M., & Schmidt, A. T. (2023). Doctoral students' intention to use assessments in their career: The incremental role of self-reported competence. *Journal of Clinical Psychology, 79*, 374–390. https://doi.org/10.1002/jclp.23415

*Blanchard, D. D., McGrath, R. E., Pogge, D. L., & Khadivi, A. (2003). A comparison of the PAI and MMPI-2 as predictors of faking bad in college students. *Journal of Personality Assessment, 80*(2), 197–205. https://doi.org/10.1207/S15327752JPA8002_08

Boccaccini, M. T., & Hart, J. R. (2018). Response style on the personality assessment inventory and other multiscale inventories. In R. Rogers & S. D. Bender (Eds.), *Clinical assessment of malingering and deception* (4th ed., pp. 280–300). The Guilford Press.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons. https://doi.org/10.1002/9780470743386

Boress, K., Gaasedelen, O. J., Croghan, A., Johnson, M. K., Caraher, K., Basso, M. R., & Whiteside, D. M. (2022a). Replication and cross-validation of the personality assessment inventory (PAI) cognitive bias scale (CBS) in a mixed clinical sample. *The Clinical Neuropsychologist, 36*(7), 1860–1877. https://doi.org/10.1080/13854046.2021.1889681

Boress, K., Gaasedelen, O. J., Croghan, A., Johnson, M. K., Caraher, K., Basso, M. R., & Whiteside, D. M. (2022b). 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

Bow, J. N., Gould, J. W., Flens, J. R., & Greenhut, D. (2006). Testing in child custody evaluations–Selection, usage, and Daubert admissibility: A survey of psychologists. *Journal of Forensic Psychology Practice, 6*(2), 17–38.

*Bowen, C., & Bryant, R. A. (2006). Malingering posttraumatic stress on the Personality Assessment Inventory. *International Journal of Forensic Psychology, 1*(3), 22–28.

Burchett, D., & Bagby, R. M. (2014). Multimethod assessment of response distortion: Integrating data from interviews, collateral records, and standardized assessment tools. In C. Hopwood & R. Bornstein (Eds.), Multimethod clinical assessment (pp. 345–378). Guilford.

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

Burchett, D., & Ben-Porath, Y. S. (2019). Methodological considerations for developing and evaluating response bias indicators.

Psychological Assessment, 31(12), 1497–1511. https://doi.org/10.1037/pas0000680

Bush, S. S. (2023). Questioning what we thought we knew: Commentary on Leonhard's performance validity assessment articles. *Neuropsychology Review, 33*(3), 624–627.

*Carrier, S. (2000). *Les effets de deux types d'informations sur la simulation d'un e´tat de stress post-traumatique measures par le Personality Assessment Inventory* [The effects of two types of information on the simulation of posttraumatic stress disorder measured by the Personality Assessment Inventory]. Unpublished master's thesis, Universite´ De Moncton.

Cersosimo, B. H., Hilsenroth, M. J., Bornstein, R. F., & Gold, J. R. (2022). Personality assessment inventory clinical scales in relation to patient and therapist rated alliance early in treatment. *Assessment, 29*(4), 806–816.

Chafetz, M., & Underhill, J. (2013). Estimated costs of malingered disability. *Archives of Clinical Neuropsychology, 28*(7), 633–639. https://doi.org/10.1093/arclin/act038

Charles, N. E., Cowell, W., & Gulledge, L. M. (2022). Using the personality assessment inventory-adolescent in legal settings. *Journal of Personality Assessment, 104*(2), 192–202.

Denning, J. H., & Shura, R. D. (2017). Cost of malingering mild traumatic brain injury-related cognitive deficits during compensation and pension evaluations in the veterans benefits administration. *Applied Neuropsychology: Adult, 26*(1), 1–16. https://doi.org/10.1080/23279095.2017.1350684

DeViva, J. C., & Bloem, W. D. (2003). Symptom exaggeration and compensation seeking among combat veterans with posttraumatic stress disorder. *Journal of Traumatic Stress, 16*(5), 503–507. https://doi.org/10.1023/A:1025766713188

*Dhillon, S. (2017). *The assessment and detection of feigned symptoms that may persist after a mild traumatic brain injury: An analogue investigation* (Master's thesis). University of Toronto, Graduate Department of Psychological Clinical Science.

*Eakin, D. E., Weathers, F. W., Benson, T. B., Anderson, C. F., & Funderburk, B. (2006). Detection of feigned posttraumatic stress disorder: A comparison of the MMPI-2 and PAI. *Journal of Psychopathology and Behavioral Assessment*, *28*(3), 145–155. https://doi.org/10.1007/s10862-005-9006-5

Faust, D. (2023). Invited commentary: Advancing but not yet advanced: Assessment of effort/malingering in forensic and clinical settings. *Neuropsychology Review, 33*(3), 628–642. https://doi.org/10.1007/s11065-023-09605-3

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

*Fernandez, K., Boccaccini, M. T., & Noland, R. M. (2008). Detecting over- and underreporting of psychopathology with the Spanish-language Personality Assessment Inventory: Findings from a simulation study with bilingual speakers. *Psychological Assessment, 20*(2), 189–194. https://doi.org/10.1037/1040-3590.20.2.189

Fokas, K. F., & Brovko, J. M. (2020). Assessing symptom validity in psychological injury evaluations using the MMPI-2-RF and the PAI: An updated review. *Psychological Injury and Law, 13*(4), 370–382.

*Gaasedelen, O. J., Whiteside, D. M., Altmaier, E., Welch, C., & Basso, M. R. (2019). The construction and the initial validation of the Cognitive Bias Scale for the Personality Assessment Inventory. *The Clinical Neuropsychologist, 33*(8), 1467–1484. https://doi.org/10.1080/13854046.2019.1612947

*Gaasedelen, O. J., Whiteside, D. M., & Basso, M. (2017). Exploring the sensitivity of the Personality Assessment Inventory symptom validity tests in detecting response bias in a mixed neuropsychological outpatient sample. *The Clinical Neuropsychologist, 31*(5), 844–856. https://doi.org/10.1080/13854046.2017.1312700

Gaines, M. V., Giles, C. L., & Morgan, R. D. (2013). The detection of feigning using multiple PAI scale elevations: A new index. *Assessment, 20*(4), 437–447. https://doi.org/10.1177/1073191112458146

Gardner, B. O., Boccaccini, M. T., Bitting, B. S., & Edens, J. F. (2015). Personality assessment inventory scores as predictors of misconduct, recidivism, and violence: A meta-analytic review. *Psychological Assessment, 27*(2), 534.

Garriga, M. (2007). Malingering in the clinical setting. *Psychiatric times, 24*(3), 15–15.

Geisinger, K. F., Ramos-Grenier, J., & Scheuneman, J. D. (1998). *Test interpretation and diversity: Achieving equity in assessment* (pp. 181–211). J. Sandoval (Ed.). American Psychological Association.

Gouvier, W. D., Lees-Haley, P. R., & Hammer, J. H. (2003). *The neuropsychological examination in the detection of malingering in the Forensic Arena: Costs and benefits*. Psychology Press.

Halligan, P. W., Bass, C., & Oakley, D. A. (2003). Wilful deception as illness behaviour. In *Malingering and illness deception* (pp. 3–28). Oxford University Press. https://doi.org/10.1093/oso/9780198515548.003.0001

*Harrison, A. G., Harrison, K. A., & Armstrong, I. T. (2022). Discriminating malingered attention Deficit Hyperactivity Disorder from genuine symptom reporting using novel personality assessment inventory validity measures. *Applied Neuropsychology: Adult, 29*(1), 10–22. https://doi.org/10.1080/23279095.2019.1702043

Hawes, S. W., & Boccaccini, M. T. (2009). Detection of overreporting of psychopathology on the Personality Assessment Inventory: A meta-analytic review. *Psychological Assessment, 21*, 112–124. https://doi.org/10.1037/a0015036

Hong, S. H., & Kim, Y. H. (2001). Detection of random response and impression management in the PAI: II. Detection indices. *Korean Journal of Clinical Psychology, 20*(4), 751–761.

*Hopwood, C. J., Orlando, M. J., & Clark, T. S. (2010). The detection of malingered pain-related disability with the Personality Assessment Inventory. *Rehabilitation Psychology, 55*(3), 307–310. https://doi.org/10.1037/a0020516

Hoyt, T., & Walter, F. A. (2022). The relationship of presurgical Personality Assessment Inventory scales to BMI following bariatric surgery. *Health Psychology, 41*(3), 184.

Hussey, I. (2023). Data is not available upon request. *Meta-Psychology*. Advance online publication. https://osf.io/8sp7e

Ingram, P. B., Schmidt, A. T., Bergquist, B. K., & Currin, J. M. (2022). Coursework, instrument exposure, and perceived competence in psychological assessment: A national survey of practices and beliefs of health service psychology trainees. *Training and Education in Professional Psychology, 16*(1), 10–19. https://doi.org/10.1037/tep0000348

*Ingram, P. B., Armistead-Jehle, P., Herring, T. T., & Morris, C. S. (2023). Cross validation of the Personality Assessment Inventory (PAI) Cognitive Bias Scale of Scales (CB-SOS) over-reporting indicators in a military sample. *Military Psychology*. https://doi.org/10.1080/08995605.2022.2160151

Ingram, P. B., Keen, M. A., Greene, T. E., Morris, C., & Armistead-Jehle, P. J. (2024). Development and initial validation of the Scale of Scales (SOS) overreporting scores for the MMPI family of instruments. *Journal of Clinical and Experimental Neuropsychology, 46*(2), 95–110. https://doi.org/10.1080/13803395.2024.2320453

Ingram, P. B., Sharpnack, J. D., Mosier, N. J., & Golden, B. L. (2021). Evaluating symptom endorsement typographies of trauma-exposed veterans on the Personality Assessment Inventory

(PAI): A latent profile analysis. *Current Psychology, 40*(11), 5267–5277.

Ingram, P. B., & Ternes, M. S. (2016). The detection of content-based invalid responding: A meta-analysis of the MMPI-2-Restructured Form's (MMPI-2-RF) over-reporting validity scales. *The Clinical Neuropsychologist, 30*(4), 473–496.

Karlin, B. E., Creech, S. K., Grimes, J. S., Clark, T. S., Meagher, M. W., & Morey, L. C. (2005). The Personality Assessment Inventory with chronic pain patients: Psychometric properties and clinical utility. *Journal of Clinical Psychology, 61*, 1571–1585. https://doi.org/10.1002/jclp.20209

*Keiski, M. A., Shore, D. L., Hamilton, J. M., & Malec, J. F. (2015). Simulation of traumatic brain injury symptoms on the Personality Assessment Inventory: An analogue study. *Assessment, 22*(2), 233–247. https://doi.org/10.1177/1073191114539380

Kim, K. W., Lim, J. S., Yang, C. M., Jang, S. H., & Lee, S. Y. (2021). Classification of adolescent psychiatric patients at high risk of suicide using the Personality Assessment Inventory by machine learning. *Psychiatry Investigation, 18*(11), 1137–1143. https://doi.org/10.30773/pi.2021.0191

*King, J., & Sullivan, K. A. (2009). Deterring malingered psychopathology: The effect of warning simulating malingerers. *Behavioral Sciences & the Law, 27*(1), 35–49. https://doi.org/10.1002/bsl.839

Kurtz, J. E., Ghosh, A., & Martin, V. A. (2023). Diagnostic efficiency of the PAI negative distortion indicators for detecting feigned head injury. *Psychology & Neuroscience, 16*(2), 155–166. https://doi.org/10.1037/pne0000308

Kurtz, J. E., & McCredie, M. N. (2022). Exaggeration or Fabrication? Assessment of Negative Response Distortion and Malingering with the Personality Assessment Inventory. *Psychological Injury and Law, 15*, 37–47. https://doi.org/10.1007/s12207-021-09433-x

Kurtz, J. E., & Pintarelli, E. M. (2024). The Daubert standards for admissibility of evidence based on the Personality Assessment Inventory. *Psychological Injury and Law*. https://doi.org/10.1007/s12207-024-09508-5

*Lange, R. T., Sullivan, K. A., & Scott, C. (2010). Comparison of MMPI-2 and PAI validity indicators to detect feigned depression and PTSD symptom reporting. *Psychiatry Research, 176*(2), 229–235. https://doi.org/10.1016/j.psychres.2009.03.004

Larrabee, G. J. (2012). Performance validity and symptom validity in neuropsychological assessment. *Journal of the International Neuropsychological Society, 18*(4), 625–630. https://doi.org/10.1017/S1355617712000240

*Lehr, E. Y. C. (2014). *A comparison of the MMPI-2-RF and PAI as predictors of naïve and informed faking* (Doctoral dissertation). The New School for Social Research.

Leonhard, C. (2023a). Review of statistical and methodological issues in the forensic prediction of malingering from validity tests: Part I: Statistical issues. *Neuropsychology Review*. https://doi.org/10.1007/s11065-023-09601-7

Leonhard, C. (2023b). Review of statistical and methodological issues in the forensic prediction of malingering from validity tests: Part II—Methodological issues. *Neuropsychology Review, 33*(3), 604–623.

Leonhard, C. (2023c). Quo vadis forensic neuropsychological malingering determinations? Reply to Drs. Bush, Faust, and Jewsbury. *Neuropsychology Review, 33*(3), 653–657.

*Liljequist, L., Kinder, B. N., & Schinka, J. A. (1998). An investigation of malingering posttraumatic stress disorder on the Personality Assessment Inventory. *Journal of Personality Assessment, 71*(3), 322–336. https://doi.org/10.1207/s15327752jpa7103_3

LoPiccolo, C. J., Goodkin, K., & Baldewicz, T. T. (1999). Current issues in the diagnosis and management of malingering. *Annals of Medicine, 31*(3), 166–174. https://doi.org/10.3109/07853899909115975

*Maffly-Kipp, J., & Morey, L. C. (2023). Detecting attention deficit hyperactivity disorder and its feigning using the personality assessment inventory. *Applied Neuropsychology: Adult*, 1–10. https://doi.org/10.1080/23279095.2023.2207215

McCredie, M. N., Hopwood, C. J., spsampsps Morey, L. C. (2023). Personality Assessment Inventory (PAI) assessment of bipolar spectrum disorders. In J. H. Kleiger spsampsps I. B. Weiner (Eds.), *Psychological assessment of bipolar spectrum disorders* (pp. 63–78). American Psychological Association. https://doi.org/10.1037/0000356-005

*McCredie, M. N., & Morey, L. C. (2018). Evaluating new supplemental indicators for the Personality Assessment Inventory: Standardization and cross-validation. *Psychological Assessment, 30*(10), 1292–1299. https://doi.org/10.1037/pas0000574

*McCullaugh, J. M. (2011). *The convergent and ecological validity of the Inventory of Problems with a community-supervised, forensic sample* (Doctoral dissertation). California School of Professional Psychology.

McDermott, B. E., & Feldman, M. D. (2007). Malingering in the medical setting. *Psychiatric Clinics of North America, 30*(4), 645–662. https://doi.org/10.1016/j.psc.2007.07.007

Meaux, L., Cox, J., Edens, J. F., DeMatteo, D., Martinez, A., & Bownes, E. (2022). The personality assessment inventory in U.S. case law: A survey and examination of relevance to legal proceedings. *Journal of Personality Assessment, 104*(2), 179–191.

Medoff, D. (2003). The scientific basis of psychological testing. *Family Court Review, 41*(2), 199–213. https://doi.org/10.1111/j.174-1617.2003.tb00884.x

*Mogge, N. L., Lepage, J. S., Bell, T., & Ragatz, L. (2010). The negative distortion scale: A new PAI validity scale. *The Journal of Forensic Psychiatry & Psychology, 21*(1), 77–90. https://doi.org/10.1080/14789940903174253

Morey, L. C. (1991). *Personality Assessment Inventory professional manual*. Psychological Assessment Resources.

Morey, L. C. (2007). *Personality Assessment Inventory professional manual* (2nd ed.). Psychological Assessment Resources.

*Morey, L. C., & Lanier, V. W. (1998). Operating characteristics of six response distortion indicators for the Personality Assessment Inventory. *Assessment, 5*(3), 203–214. https://doi.org/10.1177/107319119800500301

Morris, C. S., Ingram, P. B., & Armistead-Jehle, P. (2022). Relationship of Personality Assessment Inventory (PAI) over-reporting scales to performance validity testing in a military neuropsychological sample. *Military Psychology, 34*(4), 484–493. https://doi.org/10.1080/08995605.2021.2013059

*Musso, M. W., Hill, B. D., Barker, A. A., Pella, R. D., & Gouvier, Wm. D. (2016). Utility of the Personality Assessment Inventory for Detecting Malingered ADHD in College Students. *Journal of Attention Disorders*, *20*(9), 763–774. https://doi.org/10.1177/1087054714548031

Nevid, J. S., Gordon, A. J., & Haggerty, G. (2020). Clinical utility of the Personality Assessment Inventory in predicting symptom change and clinical outcome in an inpatient chemical dependency rehabilitation unit. *Journal of Personality Assessment, 102*(5), 587–593.

Oltmanns, J. R., Rivera Rivera, J., Cole, J., Merchant, A., & Steiner, J. P. (2020). Personality psychopathology: Longitudinal prediction of change in body mass index and weight post-bariatric surgery. *Health Psychology, 39*(3), 245.

Ord, A. S., Shura, R. D., Sansone, A. R., Martindale, S. L., Taber, K. H., & Rowland, J. A. (2021). Performance validity and symptom validity tests: Are they measuring different constructs? *Neuropsychology, 35*(3), 241–251. https://doi.org/10.1037/neu0000722

Pignolo, C., Giromini, L., Ales, F., & Zennaro, A. (2023). Detection of feigning of different symptom presentations with the PAI and

IOP-29. *Assessment, 30*(3), 565–579. https://doi.org/10.1177/10731911211061282

Rogers, R., Bagby, R. M., & Dickens, S. E. (1992). *SIRS: Structured interview of reported symptoms professional manual*. Psychological Assessment Resources.

*Rogers, R., Jackson, R. L., & Kaminski, P. L. (2005). Factitious psychological disorders: The overlooked response style in forensic evaluations. *Journal of Forensic Psychology Practice*, *5*(1), 21–41. https://doi.org/10.1300/J158v05n01_02

*Rogers, R., Gillard, N. D., Wooley, C. N., & Ross, C. A. (2012). The detection of feigned disabilities: The effectiveness of the personality assessment inventory in a traumatized inpatient sample. *Assessment*, *19*(1), 77–88. https://doi.org/10.1177/1073191111422031

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

*Rogers, R., Ornduff, S. R., & Sewell, K. W. (1993). Feigning specific disorders: A study of the Personality Assessment Inventory (PAI). *Journal of Personality Assessment, 60*(3), 554–560. https://doi.org/10.1207/s15327752jpa6003_12

Rogers, R., Sewell, K. W., Martin, M. A., & Vitacco, M. J. (2003). Detection of feigned mental disorders: A meta-analysis of the MMPI-2 and malingering. *Assessment, 10*(2), 160–177.

Rotgers, F., & Barrett, D. (1996). *Daubert v. Merrell* Dow and expert testimony by clinical psychologists: Implications and recommendations for practice. *Professional Psychology: Research and Practice, 27*(5), 467–474. https://doi.org/10.1037/0735-7028.27.5.467

*Russell, D. N., & Morey, L. C. (2019). Use of validity indicators on the Personality Assessment Inventory to detect feigning of posttraumatic stress disorder. *Psychological Injury and Law, 12*(3–4), 204–211. https://doi.org/10.1007/s12207-019-09349-7

*Samra, J. (2004). *The impact of depression on multiple measures of malingering*. National Library of Canada = Bibliothèque nationale du Canada.

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 noncredible 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

*Scragg, P., Bor, R., & Mendham, M.-C. (2000). Feigning post-traumatic stress disorder on the PAI. *Clinical Psychology & Psychotherapy, 7*(2), 155–160. https://doi.org/10.1002/(SICI)1099-0879(200005)7:2%3c155::AID-CPP237%3e3.0.CO;2-Z

Sharf, A. J., Rogers, R., Williams, M. M., & Henry, S. A. (2017). The effectiveness of the MMPI-2-RF in detecting feigned mental disorders and cognitive deficits: A meta-analysis. *Journal of Psychopathology and Behavioral Assessment, 39*, 441–455.

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.

Shura, R., Ingram, P.B., Schroder, R., & Armistead-Jehle., P. (In Press). Interpreting the Personality Assessment Inventory (PAI) Validity Scales in Veteran Affairs (VA): Leveraging a longitudinal national sample for large data analytics..

*Shura, R. D., Ingram, P. B., Miskey, H. M., Martindale, S. L., Rowland, J. A., & Armistead-Jehle, P. (2023). Validation of the personality assessment inventory (PAI) cognitive bias (CBS) and cognitive bias scale of scales (CB-SOS) in a post-deployment veteran sample. *The Clinical Neuropsychologist, 37*(7), 1548–1565. https://doi.org/10.1080/13854046.2022.2131630

Sinclair, S. J., Smith, M., Chung, W. J., Liebman, R., Stein, M. B., Antonius, D., & Blais, M. A. (2014). Extending the validity of the Personality Assessment Inventory's (PAI) level of care index (LOCI) in multiple psychiatric settings. *Journal of Personality Assessment, 97*(2), 145–152. https://doi.org/10.1080/00223891.2014.941441

Sleep, C. E., Petty, J. A., & Wygant, D. B. (2015). Framing the results: Assessment of response bias through select self-report measures in psychological injury evaluations. *Psychological Injury and Law, 8*, 27–39.

Slick, D. J., Sherman, E. M., & Iverson, G. L. (1999). Diagnostic criteria for malingered neurocognitive dysfunction: Proposed standards for clinical practice and research. *The Clinical Neuropsychologist, 13*(4), 545–561.

*Smith, S. T., Cox, J., Mowle, E. N., & Edens, J. F. (2017). Intentional inattention: Detecting feigned attention-deficit/hyperactivity disorder on the Personality Assessment Inventory. *Psychological Assessment, 29*(12), 1447–1457. https://doi.org/10.1037/pas0000435

*Sullivan, K., & King, J. (2010). Detecting faked psychopathology: A comparison of two tests to detect malingered psychopathology using a simulation design. *Psychiatry Research, 176*(1), 75–81. https://doi.org/10.1016/j.psychres.2008.07.013

Sweet, J. J., Heilbronner, R. L., Morgan, J. E., Larrabee, G. J., Rohling, M. L., & Boone, K. B. (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*. https://doi.org/10.1080/13854046.2021.1896036

*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

Tylicki, J. L., Rai, J. K., Arends, P., Gervais, R. O., & Ben-Porath, Y. S. (2021). A comparison of the MMPI-2-RF and PAI overreporting indicators in a civil forensic sample with emphasis on the Response Bias Scale (RBS) and the Cognitive Bias Scale (CBS). *Psychological Assessment, 33*, 71–83. https://doi.org/10.1037/pas0000968

Veltri, C. O. C., & Williams, J. E. (2013). Does the Disorder Matter? Investigating a moderating effect on coached noncredible overreporting using the MMPI-2 and PAI. *Assessment, 20*(2), 199–209. https://doi.org/10.1177/1073191112464619

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Whiteside, D. M., & Basso, M. R. (2024). Innovations in performance and symptom validity testing: Introduction to symptom validity section of the special issue. *Journal of Clinical and Experimental Neuropsychology, 46*(2), 81–85.

Woody, R. H. (2016). Psychological testimony and the Daubert standard. *Psychological Injury and Law, 9*, 91–96. https://doi.org/10.1007/s12207-016-9255-5

*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

Wright, A. J., Pade, H., Gottfried, E. D., Arbisi, P. A., McCord, D. M., & Wygant, D. B. (2022). Evidence-based clinical psychological assessment (EBCPA): Review of current state of the literature and best practices. *Professional Psychology: Research and Practice, 53*(4), 372.

Young, G., & Erdodi, L. (2024). Forensic prediction of malingering from performance validity tests: Review of Leonhard (2023, a, b, c). *Psychological Injury and Law*. https://doi.org/10.1007/s12207-024-09504-9