Psychological Injury and Law
https://doi.org/10.1007/s12207-022-09444-2

Assessing Negative Response Bias Using Self-Report Measures: New Articles, New Issues

Luciano Giromini[1] · **Gerald Young**[2] · **Martin Sellbom**[3]

Received: 20 December 2021 / Accepted: 2 February 2022

Abstract
In psychological injury and related forensic evaluations, two types of tests are commonly used to assess Negative Response Bias (NRB): Symptom Validity Tests (SVTs) and Performance Validity Tests (PVTs). SVTs assess the credibility of self-reported symptoms, whereas PVTs assess the credibility of observed performance on cognitive tasks. Compared to the large and ever-growing number of published PVTs, there are still relatively few validated self-report SVTs available to professionals for assessing symptom validity. In addition, while several studies have examined how to combine and integrate the results of multiple independent PVTs, there are few studies to date that have addressed the combination and integration of information obtained from multiple self-report SVTs. The Special Issue of *Psychological Injury and Law* introduced in this article aims to help fill these gaps in the literature by providing readers with detailed information about the convergent and incremental validity, strengths and weaknesses, and applicability of a number of selected measures of NRB under different conditions and in different assessment contexts. Each of the articles in this Special Issue focuses on a particular self-report SVT or set of SVTs and summarizes their conditions of use, strengths, weaknesses, and possible cut scores and relative hit rates. Here, we review the psychometric properties of the 19 selected SVTs and discuss their advantages and disadvantages. In addition, we make tentative proposals for the field to consider regarding the number of SVTs to be used in an assessment, the number of SVT failures required to invalidate test results, and the issue of redundancy when selecting multiple SVTs for an assessment.

Keywords Malingering · Negative response bias · SVTs · Symptom validity · Self-report

The expression "negative response bias" (NRB) refers to a tendency to generate less implied healthy or more pathological test results than would be expected based on the overall level of adjustment of the test-taker (Dyonisus et al., 2011; Franzen & Iverson, 2000; Rogers & Bender, 2018). In contrast to terms that have a motivational component, such as malingering or deception, NRB constitutes a response style on instruments, without any inference of motivation. Another term often used to achieve a similar non-inferential connotation is "overreporting" (Ben-Porath, 2013). However, we argue that NRB likely has an even broader meaning, as it encompasses both the overreporting (or overstatement) of experienced problems and the underreporting (or understatement) of psychological resources and strengths. In addition, we would also like to point out that NRB is distinct from both "positive response bias," in which the examinee tries to look overly healthy or good, and "content-unrelated distortion," in which the content of the item(s) or task(s) is not the reason the examinee does not respond in an appropriate or truthful manner to it. Definitional issues aside, it is now agreed that the credibility of presented psychological problems cannot be taken for granted in forensic and related assessment contexts, and that the possibility of NRB should always be considered when conducting forensic assessments (Sherman et al., 2020; Sweet et al., 2021).

To determine whether NRB is present, professionals should use multiple sources of information (APA, 2013; Sweet et al., 2021). With regard to psychological tests, Larrabee (2012) described two types of instruments that address NRB. Tests evaluating the credibility of self-reported symptoms are considered "Symptom Validity

✉ Luciano Giromini
luciano.giromini@unito.it

[1] Department of Psychology, University of Turin, Turin, Italy

[2] Glendon College, York University, Toronto, Canada

[3] Department of Psychology, University of Otago, Dunedin, New Zealand

Published online: 23 February 2022

Tests" (SVTs), whereas those evaluating the credibility of observed performance on cognitive tasks are labeled "Performance Validity Tests" (PVTs). Both SVTs and PVTs aim to detect a similar underlying phenomenon, i.e., evaluees who present themselves in an overly maladjusted/pathological manner. However, SVTs and PVTs accomplish the task in different ways. SVTs assess *overreporting* of experienced psychological problems, whereas PVTs assess *underperforming* on cognitive tasks. Both SVTs and PVTs require standard administration, comparison of test-taker results to relevant norms, acceptable reliability and validity of generated scores, and statistical properties related to their accuracy and precision (e.g., sensitivity, specificity). SVTs generally consist of items that are impossible, rare, or improbable in combination (Rogers & Bender, 2018). PVTs generally present as ostensibly difficult tasks that are in reality quite easy. Additionally, these tests differ in either being separate tests or ones embedded within larger tests. That is, both SVTs and PVTs may be further characterized as (a) "stand-alone" or "free-standing" tests, when the tests have items that are not part of other tests and their main purpose is specifically to assess the credibility of presented symptoms/performance, or (b) "embedded" measures, when they are sub-scales or sub-trials contained within longer, broader, and perhaps more complex instruments, for example, in a neuropsychological battery or multi-scale inventory. Lastly, SVTs may be further differentiated based on whether they use a self-report or an interview-based format.

During the past few decades, many embedded and free-standing PVTs have been developed, validated, researched, and adapted for use in the forensic context (Boone, 2013; Rogers & Bender, 2018). Well-known examples of PVTs are the Test of Memory Malingering (TOMM; Tombaugh, 1996) and Word Memory Test (WMT; Green et al., 1996), among free-standing measures, and the Reliable Digit Span (RDS) of the Wechsler Adult Intelligence Scale (e.g., Axelrod et al., 2006; Babikian et al., 2006; Erdodi & Abeare, 2020; Greiffenstein et al., 1994; Reese et al., 2012) and Forced Choice Recognition Trial of the California Verbal Learning Test (Erdodi et al., 2018; Greve et al., 2009; Slick et al., 2000; Wolfe et al., 2010), among embedded measures.

Some interview-based SVTs also have received much attention in the literature. In particular, the Structured Interview of Reported Symptoms (SIRS; Rogers et al., 1992), along with its revised version (SIRS-2; Rogers et al., 2010), are often referred to (by some) as the preferred methodology for assessing symptom validity (for a debate on the validity of the SIRS-2, see: Rogers et al., 2020; Tylicki et al., 2021). Another widely used and researched interview-based SVT is the Miller Forensic Assessment of Symptoms Test (M-FAST; Miller, 2001), which has recently been the subject of an extensive meta-analysis (Detullio et al., 2019).

Some self-report SVTs have been examined too during the last couple of decades. In particular, the Structured Inventory of Malingered Symptoms (SIMS; Smith & Burger, 1997) and the validity scales of the Minnesota Multiphasic Personality Inventory (MMPI–2: Butcher et al., 2001; MMPI–2–RF: Ben-Porath & Tellegen, 2008) and the Personality Assessment Inventory (PAI; Morey, 1991, 2007) have been the focus of meta-analytic research (Hawes & Boccaccini, 2009; Rogers et al., 2003; Sharf et al., 2017; van Impelen et al., 2014). However, compared to the large and continuously growing number of published PVTs, there still are relatively fewer validated self-report SVTs available to professionals performing symptom validity assessment (Dandachi-FitzGerald et al., 2013; Sherman et al., 2020). In fact, an official consensus statement by the American Academy of Clinical Neuropsychology (Sweet et al., 2021) recently stated that "there remains a need for development and validation of new SVTs, including free-standing measures and embedded measures within current self-report symptom measures" (p. 1084).

Although several studies have investigated how one should combine and integrate the results from several independent PVTs (e.g., Erdodi, 2019, 2021; Larrabee, 2008; Soble et al., 2020) and multivariate models of PVTs are also being adopted by test publishers (Pearson, 2009), the research on how to combine and integrate the information derived from multiple self-report SVTs is currently extremely scarce. In particular, in their recent revision of the Malingered Neuropsychological Dysfunction criteria (MND; Slick et al., 1999), Sherman et al. (2020) noted that, "compared to PVTs, there is very little research on the optimal number of SVTs that should be administered to properly detect malingering" (p. 751).

Another noteworthy aspect is that available SVTs vary in many ways in terms of their content and clinical criteria. Some focus on mixed symptoms and general psychopathology, others psychiatric or somatic disorders, still others cognitive problems. The MMPI-2-RF (Ben-Porath & Tellegen, 2008) and MMPI-3 (Ben-Porath & Tellegen, 2020a, 2020b) provide good examples of the range of SVTs and the domains they tap. They include the Infrequent Responses (F-r on MMPI-2-RF and F on MMPI-3) and the Infrequent Psychopathology Responses (Fp-r on MMPI-2-RF and Fp on MMPI-3) to assess over-reporting of psychopathology; the Infrequent Somatic Responses (Fs) to assess the over-reporting of somatic symptoms; the Response Bias Scale (RBS) to detect over-reporting of memory complaints; and the Symptom Validity Scale (FBS-r on MMPI-2-RF and FBS on MMPI-3) to evaluate the extent to which a presentation is characterized by the endorsement of unusual combination(s) of noncredible cognitive and somatic symptoms.

Different SVTs thus address NRB in different ways, using different detection strategies and focusing on different

domains. However, there likely is some overlap between SVT measurement approaches. Indeed, the position of the authors of the MND (Sherman et al., 2020) is that all available SVTs are potentially "affected by lack of construct unity and clarity," in that they all tap multiple symptom types (p. 750). To give an example, Sherman et al. (2020) argued that even the MMPI-2-RF RBS, which aims to specifically assess over-reporting of memory-related symptoms, in fact, contains "a mix of mostly somatic and psychiatric items along with some cognitive items" (p. 750). As such, the authors concluded that "the field would benefit from derivation of more clearly differentiated scales for the determination of separate dimensions of cognitive, somatic, and psychiatric over-reporting" (p. 751).

The Special Issue of *Psychological Injury and Law* introduced in this article aims at contributing to address these gaps and disputes in the literature, by providing the reader with detailed information on the convergent and incremental validity, strengths and weaknesses, and applicability to different conditions and evaluation contexts of a series of selected self-report measures of NRB.[1] In what follows, we summarize some of the issues that the Guest Editor of this Special Issue (first author) identified with the Editor-in-Chief of this journal (second author) at the outset of this editorial project and describe what attempts were made to address each of these issues during the process. The senior author contributed additional expertise to the overall conceptual framing of this particular article. Moreover, we provide a rationale for the selection of the measures included in this Special Issue and summarize the key information included in each of the articles that were ultimately accepted for publication.

"How Many" and "Which" SVTs Should the Assessor Use?

As noted above, Sherman et al. (2020) recently noted that "compared to PVTs, there is very little research on the optimal number of SVTs that should be administered to properly detect malingering" (p. 751). A major issue associated with using SVTs when performing symptom validity assessments, indeed, pertains to the optimal number of tools to be administered. On the one hand, using more SVTs will allow for gathering more evidence converging toward a supportable, sound forensic conclusion. Accordingly, administering more than one SVT would seem to be a reasonable choice in most symptom validity assessments, and scholars indeed tend to agree that no conclusion regarding symptom validity should be ever made based on one SVT alone (e.g., Sherman et al., 2020; Sweet et al., 2021). However, how much evidence is "enough evidence" to conclude that a given presentation is noncredible or credible, when testing symptom validity? And at what point does administering just one more SVT in the contemplated battery not contribute to incremental validity? Surplus testing along these lines would tap excessively precious resources both in the evaluator and evaluee, and might become a waste of time, an excessive redundancy to what one has already established from the results previously obtained from other SVTs. To date, there are no universally accepted standards in this regard.

A related and perhaps even more challenging issue associated with the use of multiple SVTs in symptom validity assessment concerns the criteria one should use to choose which SVTs to administer, for a given evaluation. Of course, the individual classification accuracies, psychometric properties, addressed conditions, etc., of the available tools for the case at hand will play a key role in making this decision. Aside from these more overt features, however, that the expert forensic evaluator will surely consider, a subtler question concerns whether it would be preferable to administer validity checks that are somehow similar to each other, or would it be better to look for complementary and nonredundant ones? Below, we report some considerations concerning the pros and cons of each possible solution.

On the one hand, a clinician might want to select SVTs that are highly intercorrelated with each other, to gain more confidence on the reliability (in the legal sense) of the outcomes generated by each of the administered measures. The assessor would feel more confident in the determination made on the validity/credibility of presented complaints if the multiple administered SVTs agree with each other that the presentation at hand is noncredible (or that it is credible), compared to if they generate conflicting results. In addition, psychometricians have repeatedly suggested that a given measure should correlate with other similar measures assessing the same construct, so as to demonstrate convergent validity (Campbell & Fiske, 1959; Howell, 2013). From this perspective, it would thus seem reasonable to think that SVTs that are highly intercorrelated with each other may be preferable over those that generate weak correlations and perhaps conflicting results when tested against other similar measures.

[1] In this article, we use the expressions "self-report measures of negative response bias" and "self-report symptom validity tests" interchangeably. However, we argue that the former expression is a broader and more inclusive term than the latter. Indeed, all self-report symptom validity tests are, by definition, self-report measures of negative response bias; however, not all self-report measures of negative response bias necessarily assess the validity of presented *symptoms*. For instance, the Inventory of Problems–29 (IOP-29; Viglione & Giromini, 2020) is a self-report measure of negative response bias that does not uniquely focus on specific psychological *symptoms*, but rather addresses the credibility of various psychological *problems* and *attitudes*.

However, two similarly effective SVTs that strongly correlate with each other may yield excessively redundant information, so that there might be little potential for incremental validity. Indeed, as empirically demonstrated by Tsujimoto et al. (1990), the greater the redundancy between the predictors of a criterion variable, the smaller the gain in predictive validity. Moreover, from a practical perspective, assessing symptom and performance validity is notably different from assessing other constructs in medicine and neuropsychology (Chafetz, 2020). For example, NRB levels are likely to vary across different measures of NRB, because examinees often deliberately choose to restrict their NRB to a very limited number of domains of psychological functioning and do well in other domains (e.g., to appear cognitively impaired, someone might deliberately try to pretend to be unable to perform mathematical calculations but perform well on memory tasks) (Cottingham et al., 2014; Erdodi et al., 2018). From this standpoint, using SVTs that provide information about the same evaluee from different angles, e.g., by relying on different detection strategies or by focusing on multiple domains (such as somatic, cognitive, and psychiatric) might, therefore, be more beneficial than using SVTs that use the same detection strategy or focus on the same one symptom domain. Granted, evaluees will have complaints over multiple domains, but to gain a more complete view of a given evaluee and their NRB, having several SVTs focusing on different aspects of their presentation is probably better than having a limited focus on their symptomatology and its NRB. As in any other clinical evaluation (Hunsley & Meyer, 2003), thus, also in the context of symptom validity assessment, the desire to seek confirmatory evidence should probably be balanced with the recognition that the use of nonoverlapping or nonredundant measures ultimately improves the overall accuracy of evaluation decisions. That said, there are no universally accepted guidelines in this regard, and, based on our experience, while some authors and professionals seem to prefer to include in their assessments SVTs that are highly similar to each other, others seem to prefer to include nonredundant and complementary ones.

To further complicate matters, research demonstrates that the conclusions about the incremental validity of a test are context specific, in that the same test might yield different levels of incremental validity depending on the particular conditions of use, the base rate of the phenomenon under investigation, etc. (Anastasi, 1988; Hunsley & Meyer, 2003; Wiggins, 1973). Consistent with this position, Wygant et al. (2007) noted that civil plaintiffs tend to limit their overreporting to narrow, cognitive domains, whereas criminal litigants exaggerate both psychiatric symptoms and cognitive impairment. As such, the extent to which an SVT yields incremental validity when added to the multi-method assessment of symptom validity likely depends also on its effectiveness in the measurement of the credibility of different symptom presentations. For instance, the SIMS is known to perform well when assessing the credibility of depression- or anxiety-related problems, but sub-optimally when used to assess psychosis-related conditions (Giromini et al., 2018; van Impelen et al., 2014). Conversely, the validity scales of the PAI seem to be more efficient in detecting feigned psychosis rather than feigned mood or anxiety disorders (Hawes & Boccaccini, 2009). Hence, the SIMS possibly adds incremental validity over the PAI validity scales when assessing depression-related presentations, but not when assessing psychosis-related symptoms, though this hypothesis requires empirical testing.

This scenario is further complicated by the fact that real-life situations often present with multiple reasons for why NRB indicators might be present, including genuine problems. Individuals who score in the clinical range for NRB may mix genuine and exaggerated and even malingered symptomatology in their evaluation presentations. Others might present in a particular manner unintentionally (e.g., excessively negative self-evaluation). Therefore, SVTs should never be used in isolation in evaluations of intentional NRB (e.g., malingering) in any forensic assessment context. That said, they should not be dismissed when there are significantly elevated results. These findings would need to be considered carefully for all possible interpretations, with the best one based on the overall pattern of information, data inconsistencies, and data gathered throughout the evaluation (Erdodi et al., 2018; Merten & Merckelbach, 2013; Young, 2019, 2021).

In order to improve the precision of an assessment in these regards, astute forensic evaluators will consider all of the aforementioned factors when deciding *how many* and *which* SVTs to include in their assessment battery. Additionally, these decisions will also depend on many other factors, such as the available resources, the cognitive and attentive abilities of the evaluee at hand, and so forth. Indeed, because there are so many factors involved in the choice of how many and which SVTs to select for a given evaluation, one cannot expect to find a universal, pre-selected battery of SVTs as a valid solution for all circumstances. Nevertheless, just as some steps forward have been made in the development of guidelines for the use of PVTs (e.g., Erdodi, 2019, 2021; Heilbronner et al., 2009; Larrabee, 2008; Soble et al., 2020; Sweet et al., 2021), we believe that the time has come to make initial general recommendations toward this direction with regard to the use of SVTs in forensic civil assessment contexts. This is the main reason why the editorial project introduced in this article was started.

In our opinion, to appreciate how different SVTs could uniquely contribute to the multi-method assessment of symptom validity, each measure would need to be considered not only in terms of its reliability and validity and

classification accuracy, but also by considering its optimal conditions and contexts of use (evaluations focused on psychological injury, criminal culpability, competency to stand trial, etc.), its addressed domains (e.g., psychiatric, somatic, cognitive), diagnostic targets (assessment of cognitive impairment, PTSD, schizophrenia, etc.), and required resources; and its unique strengths and weaknesses. The primary purpose of the Special Issue introduced in this article, thus, is to try to better understand how the eight selected self-report instruments used to assess NRB might perform under different conditions and when used in combination with other, different, and independent validity checks.

The Everlasting Dilemma of Optimal Cut Off Score(s)

An everlasting dilemma in the field of symptom validity assessment concerns the "optimal" cut score(s) to consider when interpreting the result of an SVT. In principle, the credibility of a given presentation could be conceived as dimensional in nature. Certain presentations are just more credible than others, and at the ends of this continuum there are those presentations that are definitely noncredible or invalid versus those that are definitely credible or valid. Nevertheless, in real-life and applied settings, forensic practitioners are often asked to make a dichotomous determination and opine on whether the presentation at hand *is* or *is not* credible. As such, this Special Issue will also consider the issue of which cut scores might be more useful for the different cases at hand, considering their demographic characteristics, the evaluation context, etc.

As is the case with PVTs, SVT cutoffs also are typically calibrated to have a low false-positive rate (≤ 10%, i.e., specificity ≥ .90) (Sherman et al., 2020). This calibration is because making erroneous determinations of NRB based on test results is generally considered to be more harmful to the evaluee than missing the detection of a noncredible presentation using the tests. However, with instruments designed to be used for screening purposes, sensitivity should be considered more important than specificity because only positive classifications would be further evaluated with more specific testing. Accordingly, in these cases, more optimal cut scores would be those that generate sensitivity levels of about .90 or more (Giromini et al., 2020). Furthermore, at times the expression "optimal cut scores" is used to refer to those cut scores that maximize the overall classification rates (Rogers et al., 2003). Thus, an initial potential source of confusion is that the term *optimal* cut scores means different things in different contexts, and what is an optimal cut score in a high-stake forensic evaluation may be non-optimal in a screening or in a research context.

Additionally, and more importantly, for the great majority of available SVTs, what may be conceived of as the "optimal cut score" of an SVT ultimately varies across contexts and research studies. For instance, when the SIMS was originally published, Smith and Burger (1997) recommended considering a total SIMS score ≥ 15 as indicative of a noncredible presentation. However, research later demonstrated that the ≥ 15 cut score would yield sub-optimal specificity values so that higher cut scores—namely, ≥ 17, ≥ 20, and ≥ 25—have later been proposed as "more optimal" cutoffs (van Impelen et al., 2014). Besides, the same SIMS cut score would generate dramatically different sensitivity, specificity, and overall classification rates depending on whether one is assessing individuals tested for possible psychosis or intellectual disability versus depression- or anxiety-related problems (Giromini et al., 2018; van Impelen et al., 2014).

Similar considerations also apply, albeit with some differentiations, to all other popular instruments used to assess symptom validity. For instance, Morey (2003) initially recommended considering a Negative Impression Management (NIM) score ≥ 73 T on the PAI as indicative of some exaggeration, and a score ≥ 84 T as possibly associated with intentional response distortion. After reviewing different cut scores performance statistics from 13 NIM studies, Sellbom and Bagby (2008) instead suggested that a cutoff of ≥ 77 T would be more effective for identifying suspected malingering, with ≥ 110 T being the most effective cutoff for identifying a strong likelihood of malingering. Only 1 year later, however, Hawes and Boccaccini (2009) published a meta-analysis in which they concluded that "for NIM, a cut score of ≥ 81 T yielded the highest overall classification rate while demonstrating relatively strong levels of sensitivity and specificity. This optimal cut score differs from Morey's (2003) recommended cut score of 84 T and Sellbom and Bagby's (2008) recommended cut score ≥ 77 T" (p. 121). Additionally, Hawes and Boccaccini's (2009) meta-analysis also suggested that NIM cut scores recommended by Morey (2003) and Sellbom and Bagby (2008) would yield an excessively high number of false positives, which would be appropriate only when the PAI is being used for screening purposes (see also Boccaccini & Hart, 2018).

This uncertainty as to which cut scores one should consider when making a determination on the credibility of the overall forensic presentation of the case at hand poses a serious challenge to the use of existing SVTs. On the other hand, as Rogers et al. (2012) clearly pointed out when elaborating on the "laser accuracy myth of cut scores" (p. 79), SVT scores are not free from measurement error. As such, one should be very careful when interpreting single-point cut scores.

When envisioning the overall organization of the articles to be included in this Special Issue, it was determined

that the field would benefit from an update on the research informing on the diagnostic efficiency statistics of selected SVTs at various cut scores. Accordingly, in addition to inviting contributing authors to provide the readers with data informing on the conditions of use and convergent and incremental validity of each measure, it was also asked them to offer some general guidelines on which cut scores the forensic evaluator should consider in different contexts and situations. A full description of the Call for Papers prepared before contacting authors potentially interested in contributing to this Special Issue is reported in the next section.

Content and Structure of the Articles Included in the Special Issue

As described above, the primary purpose of this editorial project was to advance the field of symptom validity assessment by summarizing key information concerning the conditions of use, convergent and incremental validity, cut scores and hit rates, strengths and weaknesses, and needed research for a set of selected measures of NRB. Accordingly, the following text was prepared, and subsequently emailed to a number of scholars and SVT experts potentially interested in contributing to this editorial project:

> "This Special Issue will present a series of research review articles sharing the same structure and authored by some leading scholars in the field. Each article focuses on a specific self-report SVT (e.g., the SIMS) or set of SVTs (e.g., the MMPI validity scales) summarizing their conditions of use, strengths, weaknesses, and possible cut scores and relative hit rates. The first section of each paper, *Background and Conditions of Use*, will identify the detection strategies implemented by the target measure, as well as the types of evaluations (e.g., psychological injury, not guilty by reason of insanity, competency to stand trial, etc.) and disorders to which it applies (e.g., traumatic brain injury and other neuropsychological disorders, PTSD and adjustment disorders, schizophrenia). To better appreciate how to use each measure in a multi-method assessment aimed at evaluating negative response bias, the *Convergent and Incremental Validity* section will then describe the external criterion variables (e.g., SIRS based determinations, TOMM scores, SIMS results) each measure tends to correlate the most with, as well as the extent to which the target SVT yields incremental validity when used together with other symptom or performance validity tests. Each paper will then address cut scores and relative hit rates, by summarizing test manual guidelines and empirical research findings reported in the literature (*Cut Scores and Hit Rates*). This will be followed by *Strengths and Weaknesses*, which presents evidence to support explicit conclusions and opinions about the applicability, strengths, and weakness of the target measures, i.e., what we know about the characteristics of each measure and about what makes it unique or different compared to available alternatives. The final section, *Future Perspectives*, will summarize the important characteristics of the measure that have been established by research, as well as what we do not yet know about it that would be helpful to know, i.e., what type of research is most urgently needed for this measure."

Choosing the Instruments to be Included in the Special Issue

To determine which self-report SVTs should be included in this Special Issue, two chief criteria were considered, i.e., the frequency with which professionals use a given tool in their practice, and the number of research articles reporting on available SVTs during the past few years. With regard to the former criterion, Neal and Grisso (2014) recently conducted an international survey in which 434 experts described their two most recent forensic evaluations. Data analyses revealed that three of the ten most frequently used tools included multi-scale personality inventories that embedded scales aimed at measuring negative response bias: the MMPI (any version), PAI, and the Millon Clinical Multiaxial Inventory (MCMI-III: Millon et al., 2009; MCMI-IV: Millon et al., 2015). As such, it was concluded that this Special Issue needed to include those instruments.

With regard to the second of the criteria considered to identify self-report SVTs suitable for inclusion in this Special Issue, a brief literature search was conducted to identify a pool of self-report measures of NRB that had been the subject of research investigations during the past 20 years (i.e., within the first 20 years of the third millennium). This second approach yielded the following additions to the list of suitable SVTs: the Structured Inventory of Malingered Symptoms (SIMS; Smith & Burger, 1997), the Inventory of Problems–29 (IOP-29; Viglione et al., 2017), the Self-Report Symptom Inventory (SRSI; Merten et al., 2016), the Memory Complaints Inventory (MCI; Green, 2019), and the Atypical Response scale (ATR) of the Trauma Symptom Inventory (TSI-2; Briere, 2011).

Lastly, in an attempt to make sure that this Special Issue would not miss any other popular self-report SVTs that would be reasonable to include, the Guest Editor surveyed a few expert practitioners among his colleagues. This final step added to the list a few additional instruments—e.g., the Psychological Screening Inventory (PSI; Lanyon, 2006), the Personal Problems Questionnaire (PPQ; van den Broek

Table 1 Instruments and articles included in the Special Issue

Instrument(s)	Title of the article	Author(s)
MMPI-2-RF & MMPI-3	Assessing Negative Response Bias: A Review of the Noncredible Over-reporting Scales of the MMPI-2-RF and MMPI-3	Burchett, D., & Bagby, M
PAI	Exaggeration or Fabrication? Assessment of Negative Response Distortion and Malingering with the Personality Assessment Inventory	Kurtz, J. E., & McCredie, M. N
MCMI-III & MCMI-IV	Negative Response Bias with the MCMI	Choca, J. P., & Pignolo, C
TSI-2	Detecting Negative Response Bias within the Trauma Symptom Inventory – 2 (TSI-2): A Review of the Literature	Ales, F., & Erdodi, L
SIMS	Structured Inventory of Malingered Symptomatology: A Psychometric Review	Shura, R., Ord, A. S., & Worthen, M. D
IOP-29	Assessing Negative Response Bias with the Inventory of Problems – 29 (IOP-29): A Quantitative Systematic Review	Giromini, L., & Viglione, D. J
SRSI	The Self-Report Symptom Inventory	Merten, T., Dandachi-FitzGerald, B., Boskovic, I., Puente-López, E., & Merckelbach, H
MCI	Memory Complaints Inventory: Review of Psychometric Properties	Armistead-Jehle, P., & Shura, R

et al., 2012), the M Test (Beaber et al., 1985), etc. However, none of the authors contacted to review the available literature on these instruments agreed to contribute to this Special Issue. As such, the final list of instruments addressed in this issue is summarized in Table 1.

Summary of Contents

The Validity Scales of the MMPI-2-RF and MMPI-3

The MMPI-2-RF and the MMPI-3 probably are the most popular self-report personality inventories for assessing adult personality and psychopathology in forensic evaluations (Neal & Grisso, 2014). The MMPI-2-RF was published in 2008 as a shorter (338 items) and psychometrically improved alternative to the MMPI-2. In 2020, a newer iteration of the family of MMPI instruments, comprised of 335 items, was released: the MMPI-3. The main reasons for developing this newer version were (a) the need to update the instrument's normative data and (b) the desire to update the item content of the test to capture important areas of psychopathology not included in the MMPI-2-RF (e.g., eating disorders, compulsivity, impulsivity). In addition, some minor wording changes were also made to simplify existing items. The first article in this Special Issue focuses on the validity scales of the MMPI-2-RF and MMPI-3 (Burchett & Bagby, 2021).

The MMPI instruments have a rich history of incorporating embedded strategies to identify invalidating response styles. The function, operation, and detection strategies of the validity scales embedded in the MMPI-3 are very similar to their corresponding counterparts in the MMPI-2-RF. In fact, the correlation between the MMPI-2-RF and the MMPI-3 versions of the validity scales that address NRB is $r \geq .95$ (Ben-Porath & Tellegen, 2020b). Two scales (F/F-r and Fp/Fp-r) assess overreporting of general psychopathology; one scale (Fs) assesses overreporting of somatic symptoms; one scale (RBS) assesses overreporting of memory complaints; and one scale (FBS/FBS-r) assesses the endorsement of unusual combination(s) of noncredible cognitive and somatic symptoms.

The amount of empirical research supporting the effectiveness of the validity scales of the MMPI-2-RF is impressive, and studies contributing to the validity of their MMPI-3 counterparts are also rapidly accumulating. Based on their review of the relevant literature, Burchett and Bagby (2021) conclude that the Fp-r is by far the most effective MMPI-2-RF validity scale for capturing overreporting of various types of mental health problems, with a cut score ≥ 100 T maximizing the balance between overall hit rates and specificity. Indeed, the Fp-r has shown the most satisfactory classification accuracy not only in detecting feigned emotional disturbance, but also in detecting feigned cognitive impairment. On the other hand, the FBS-r and the RBS also contribute uniquely to the MMPI-based assessment of symptom and performance validity because they presumably provide more specific insight into the possibility of feigned cognitive or memory impairment. A similar conclusion probably holds for the MMPI-3, too. That is, the MMPI-3 F and Fp are presumably best at detecting feigned emotional disorders, whereas the MMPI-3 Fs, FBS, and RBS are presumably best at detecting feigned somatic and/or cognitive complaints.

Regarding convergent validity, Burchett and Bagby (2021) cite a study by Tylicki et al. (2020) that analyzed 550 MMPI-3 protocols from disability claimants. The authors hypothesized that while all overreporting scales of the

MMPI-3 should correlate with available SVTs and PVTs, the F and Fp should correlate more strongly with SVTs while the Fs, FBS, and RBS should correlate more strongly with PVTs. The results only partially supported these hypotheses. Indeed, the correlations of F and Fp ranged from |*r*|= .36 to |*r*|= .75 for SVTs and from |*r*|= .00 to |*r*|= .24 for PVTs; the correlations of Fs, FBS, and RBS ranged from |*r*|= .30 to |*r*|= .60 for SVTs and from |*r*|= .03 to |*r*|= .33 for PVTs. Thus, pending future replications, it appears that the five embedded SVTs of the MMPI-3 are likely to correlate more strongly with other SVTs than with PVTs. It should be noted, however, that in the Tylicki et al. (2020) study, the RBS had the highest effect size of all MMPI-3 validity scales in identifying probable/definite malingering based on Sherman et al. (2020) criteria for malingered neurocognitive dysfunction.

Finally, Burchett and Bagby (2021) provide excellent considerations and comments for deciding which MMPI version to choose for the case at hand. First, given the progress made with the MMPI-2-RF and the contemporary normative sample available with the MMPI-3, they conclude that "it would be difficult to make a strong case for the continued use of the MMPI-2 at this time" (p. 12). Next, the authors highlight the relative advantages of using the MMPI-2-RF versus MMPI-3. With a literature base that spans nearly 15 years, the MMPI-2-RF is obviously more consolidated in the scientific literature at this point. However, as noted earlier, the research base for the use of the MMPI-3 in court is growing rapidly (Ben-Porath et al., in press). Moreover, the overreporting scales of the MMPI-3 are very similar to their corresponding counterparts in the MMPI-2-RF—even identical in the case of FBS and RBS—so that the research base of the MMPI-2-RF for the use of these scales in forensic settings should also apply to the MMPI-3. Furthermore, the normative sample of the MMPI-3 is more updated compared to that of the MMPI-2-RF, which was collected in the 1980s. Clinicians are therefore encouraged to monitor the accumulation of MMPI-3 studies (see, e.g., Morris et al., 2021; Reeves et al., in press; Tylicki et al., 2020; Whitman et al., 2021) and assess the extent to which these studies are conducted in a context similar to their practice. In any case, the considerations to be made in deciding which version of the MMPI to use for a given case can be expected to change significantly in the near future as the research base for the use of the MMPI-3 continues to grow.

The Validity Scales of the PAI

The PAI also is a very popular, broad-band personality inventory aimed at measuring adult personality and psychopathology (Neal & Grisso, 2014). Comprised of 344 items, it includes three standard indicators of NRB: the Negative Impression Management (NIM; Morey, 1991) scale, the Malingering Index (MAL; Morey, 1996), and the Rogers Discriminant Function (RDF; Rogers et al., 1996). In addition, three other supplemental indicators have recently been added to the more updated PAI interpretative report (PAI-plus; Morey, 2020): the Negative Distortion Scale (NDS; Mogge et al., 2010), the Hong Malingering Index (HMI; Hong & Kim, 2001), and the Multiscale Feigning Index (MFI; Gaines et al., 2013). The second of the articles included in this Special Issue, written by Kurtz and McCredie (2021), reviews the available research literature on the effectiveness of each of these six indicators.

The items on the NIM and NDS present exaggerated complaints or symptoms that are uncommon or unexpected in genuine patients. In contrast, the MAL and MFI use a "profile-level approach" to identify NRB, i.e., they focus on unlikely combinations of scores from different full scales or subscales. Finally, the RDF and HMI represent a deeper level of complexity in their approach to assessing NRB, in that they were derived from discriminant function analyses so various PAI scales are weighted differently, based on their unique contribution to identifying NRB.

There is now a large body of research supporting the validity of the NRB indicators embedded in the PAI, particularly with respect to the original NRB indicators. According to Hawes and Boccaccini's (2009) meta-analytic review, the NIM, MAL, and RDF tend to yield larger effect sizes when assessing the credibility of severe mental disorders (e.g., intellectual disability or psychosis) than when assessing the credibility of mood- or anxiety-related presentations. However, all three indicators are strong predictors of both coached and uncoached malingering, and all appear to be effective in detecting overreporting across various types of mental health problems (PTSD, neurocognitive deficits, psychosis, etc.). Furthermore, the three newer supplemental indicators, i.e., the NDS, the HMI, and, although to a lesser extent, the MFI, have also shown promising results in recent research.

Kurtz and McCredie (2021) note that "although the evidence for each indicator is encouraging, Morey and Hopwood (2007) recommend a configural approach that accounts for the unique contributions of NIM, MAL, and RDF to the assessment of negative distortion" (p. 5). Indeed, correlations among the three original indicators of NRB range from $r = .10$ to $r = .62$. The RDF, in particular, correlates weakly with the other two NRB scales in both the community adult normative sample ($r \leq .38$) and the clinical standardization sample ($r \leq .11$) described in the professional manual. Accordingly, it is reasonable to hypothesize that the NIM, MAL, and RDF may capture different aspects of NRB and thus provide incremental information about it. Nevertheless, Kurtz and McCredie (2021) emphasize that the effectiveness of the configural analysis of NRB recommended by Morey and Hopwood (2007) "has not been adequately evaluated in the empirical research literature" (p. 9).

The Modifying Indices of the MCMI-IV

The third of the articles included in this Special Issue (Choca & Pignolo, 2022) focuses on the Modifying Indices of the MCMI-IV, another broad-band clinical and personality assessment inventory. Unlike the validity scales of the MMPI-2-RF, MMPI-3, and PAI, there is currently no research on the effectiveness of the NRB indicators embedded in the MCMI-IV.

The MCMI-IV is a 195-item questionnaire designed to capture DSM-5 (American Psychiatric Association, 2013) personality disorders and clinical syndromes (Choca & Grossman, 2015). It was released in 2015 as an update to the previous version of the test, the MCMI-III. However, the updated version overlaps heavily with the previous one, in that 120 of the 195 items of the MCMI-IV were taken from the MCMI-III. The three embedded indicators of impression management in the MCMI-IV are the Disclosure (X), Desirability (Y) and Debasement (Z) scales. High scores on X and Z and low scores on Y are indicative of possible NRB.

Other than what is included in the test manual, there have been very few published studies using the MCMI-IV, and none of them focuses specifically on NRB. Therefore, Choca and Pignolo (2022) review the literature on the Modifying Indices of the MCMI-III, and assume that the reported results would likely apply to the current version as well. In their article, they summarize the results of four studies aimed at testing the effectiveness of X, Y, and Z. Three used a simulation design, one used a criterion group design. Based on the reported results, they conclude that although Z showed some promise in detecting noncredible cognitive impairment, "taken together, these studies suggest that the Modifier Indices alone showed only modest abilities to detect experimental feigners," (p. 4).

Choca and Pignolo (2022) also note that the MCMI-III Modifier Indices showed a strong relationship with the MMPI-2 validity scales (Morgan et al., 2002; Schoenberg et al., 2004), but were not associated with the TOMM or RDS scores (Ruocco et al., 2008). Thus, it can be concluded that the embedded SVTs of the MCMI-IV are likely to correlate with other SVTs but not with PVTs. However, given the lack of a research base for the efficacy of the MCMI-IV Modifying Indices, caution is warranted in relying on these scales to make decisions about overreporting in forensic psychological assessments.

The Atypical Response (ATR) Scale of the TSI-2

The TSI-2 is a relatively new broad-band self-report inventory designed to assess symptoms of PTSD. It was developed to update the earlier version of the instrument (i.e., the TSI; Briere, 1995), which was unable to detect over-reporting or fabrication of PTSD symptoms (Palermo & Brand, 2019). Two validity scales are embedded within the TSI-2: the ATR and the response level (RL). The former (ATR) assesses NRB whereas the latter (RL) assesses positive impression management (e.g., denial of common problems or understating of psychopathological symptoms). The fourth article in this Special Issue reviews the available literature on the effectiveness of the TSI-2 ATR scale (Ales & Erdodi, 2021).

The ATR scale of the TSI-2 assesses symptom exaggeration and inaccurate representation of PTSD symptomatology. Based on the TSI-2 manual, the items included in the ATR scale seem to indicate PTSD, but would in fact be endorsed rarely by true PTSD patients. However, according to Ales and Erdodi (2021), both the liberal (≥ 8) and conservative (≥ 15) cut scores recommended in the TSI-2 manual result in unacceptably high false positive rates of 49% and 33%, respectively. Indeed, Ales and Erdodi (2021) state that "the limited evidence available suggests that ATR has the potential to serve as measure of symptom validity, although its classification accuracy is generally inferior compared to well-established scales. While the ATR seems sufficiently sensitive to symptom over-reporting, significant concerns about its specificity persist" (p. 1).

Convergent validity is not addressed in detail in Ales and Erdodi (2021). However, in describing the initial research conducted with the first version of the ATR scale (Briere, 1995), the authors report that it correlates with the F scale of the MMPI-2 at $r = .50$ and with the NIM of the PAI at $r = .52$. No information is presented on the extent to which the ATR correlates with PVTs.

The SIMS

The fifth article in this Special Issue, written by Shura et al. (2021), focuses on the SIMS, one of the most commonly used free-standing SVTs. Comprised of 75 items, the SIMS purports to assess overreporting of psychological and cognitive symptoms by presenting the test-taker with 75 rare, atypical, or extreme symptoms that genuine patients presumably tend not to endorse.

In 2014, van Impelen et al. (2014) published a meta-analysis testing the validity of the SIMS for detecting noncredible symptom presentations. In their psychometric review article, Shura et al. (2021) provide an updated diagnostic accuracy table that also includes the SIMS research studies published since the earlier meta-analytic review by van Impelen et al. (2014). Taken together, the results presented by Shura et al. (2021) suggest that (a) the SIMS has been used and researched extensively worldwide over

the past 20 years and (b) the SIMS has excellent sensitivity to overreporting, so it can be very useful in *ruling out* the need for additional symptom validity assessment. However, patients with marked apathy, alexithymia, or schizophrenia, as well as veterans with PTSD and inpatients with extensive trauma history, are likely to generate false positive results. More generally, the cut score of ≥ 15 suggested in the test manual is probably too liberal in any case, so more conservative cut scores such as ≥ 17 may be preferable in any forensic setting. Indeed, Shura et al. (2021) state that "when using common cut off scores, the SIMS does not reliably distinguish feigned psychopathology from severe manifestations of genuine psychiatric illness" (p. 1).

The review by Shura et al. (2021) also shows that the SIMS is highly correlated with other popular SVTs. For example, in a study reported in the test manual, the SIMS total score correlated $r = .84$ with the F scale of the MMPI-2; in a study of 57 men suspected of feigning competence to stand trial, SIMS correlations ranged from $r = .47$ to .50 when considering MMPI-2 validity scales and from .43 to .80 when considering the SIRS; in a study of 115 prison inmates, the SIMS correlated with the SIRS at $r = .81$ and with the NIM, MAL, and RDF of the PAI at $r = .84$, .68, and .45, respectively. In contrast, the relationship between the SIMS and PVT scores is more controversial. The test manual references correlations with the TOMM that range from − .91 to − .89 when considering a small disability sample ($n = 20$). However, an independent study from the Netherlands with a larger sample of mixed psychiatric patients found a much lower correlation of $r = -.22$ with another PVT, the Amsterdam Short Term Memory test (Dandachi-FitzGerald et al., 2011).

Finally, Shura et al. (2021) also note that there is little research on the incremental validity of the SIMS. In a study by Lewis et al. (2002), the SIMS total score showed no incremental validity beyond the MMPI 2 Fb, which was the best predictor of invalid status. Another study by Edens et al. (2007), using the SIMS together with the SIRS, yielded a significantly better prediction of group status compared to using the SIMS alone, but predictive accuracy only increased from 69 to 72% (unfortunately, the authors of that study did not report whether adding the SIMS to the model after the SIRS would increase predictive accuracy). All in all, the concluding comment of Shura et al. (2021) regarding the incremental validity of the SIMS is very reasonable: "In conclusion, SIMS incremental validity is not well established when compared to other SVTs; however, this applies more so in the test battery use than in the screening use, and incremental validity arguably is less important in that situation" (p. 8).

The IOP-29

The sixth article in this Special Issue (Giromini & Viglione, 2021) presents a quantitative literature review examining the psychometric properties of the IOP-29. Of all the free-standing SVTs described in this issue, the IOP-29 is the shortest: comprised of 29 items only, it is designed to identify noncredible presentations of various psychiatric and cognitive disorders, including those related to PTSD, depression, anxiety, schizophrenia, cognitive impairment, and a combination thereof.

The IOP-29 differs from the typical SVT in several ways. First, it not only describes symptoms that are rarely endorsed by genuine patients, but also asks about the strategies and solutions that the test-taker uses to cope with their problems. In fact, the IOP-29 includes a number of detection strategies typically used in PVTs and clinical interviews, but not in SVTs. In addition, its chief feigning scale (i.e., the False Disorder probability Score; FDS) was derived from logistic regression analyses comparing the responses provided of a group of bona fide patients with those of a group of experimental simulators. Thus, to assess the credibility of a given IOP-29, the FDS relies on two (rather than one, as is the case for the typical SVT) sets of reference data, one from valid and one from invalid IOP-29 protocols. Furthermore, the IOP-29 response options do not use the classic true–false dichotomy (as in SIMS or MMPI) or the standard Likert scale (as in PAI). Some of them offer three answer choices, i.e., "true," "false," or "doesn't make sense," while others are open-ended questions about logical or mathematical problems. For all these reasons, the authors of the IOP-29 suggest that the IOP-29 should yield a unique contribution (i.e., incremental validity) when added to the multi-method battery for testing symptom or performance validity (Viglione & Giromini, 2020).

The quantitative literature review described in Giromini and Viglione (2021) shows that although the IOP-29 was introduced relatively recently, in 2017, research supporting its psychometric properties is growing very rapidly. Indeed, the published IOP-29 studies included in the review by Giromini and Viglione (2021) were conducted in ten different countries (i.e., Australia, Brazil, Canada, England, France, Italy, Lithuania, Portugal, Slovenia, and North America), and the results of these studies consistently confirmed the excellent effectiveness of the instrument, with minimal differences across studies. More specifically, "when considering the 3777 IOP-29 protocols included in the statistical analyses comparing credible ($k = 16$) versus noncredible ($k = 17$) presentations, the standard IOP-29 cut score of FDS ≥ .50 yielded a weighted mean sensitivity of .86 (weighted $SD = .07$; range .63–.96) at a weighted mean specificity of 92 (weighted $SD = .06$; range .79–1.00). The weighted mean Cohen's d was 3.02

(weighted $SD = .98$; range 1.48–5.31) and the weighted mean *AUC* was 95 (weighted $SD = .04$; range .83–1.00)" (Giromini & Viglione, 2021; p. 1). The more conservative cut score, FDS ≥ .65, also demonstrated excellent signal detection, with a weighted mean sensitivity of .76 (weighted $SD = .08$) at a weighted mean specificity of .96 (weighted $SD = .03$). Particularly noteworthy about these data are the very low weighted standard deviations, indicating that the results of the individual studies differed only minimally.

In their concluding remarks, Giromini and Viglione (2021) point out as a limitation of their study that most of the studies included in their review used a simulation design, which is known to inflate effect sizes, especially when nonclinical volunteers are used as control groups. Indeed, they highlight that when their analyses were based only on the datasets of the simulation studies that used real patients as control groups (total $n = 962$), the weighted mean Cohen's *d* decreased to 2.01 (weighted $SD = .45$), and the weighted mean *AUC* was .90 (weighted $SD = .03$). Nonetheless, these results are still impressive and they are similar to those reported in a criterion group study by Roma et al. (2020) in an ecologically valid sample of 75 court-ordered psychological injury evaluations, in which Cohen's *d* was 2.98 and *AUC* was .98.

In 14 samples from 11 of the articles included in the review by Giromini and Viglione (2021), the IOP-29 had been administered together with other SVTs and/or PVTs. These datasets were therefore analyzed to test convergent and, for a smaller subset of seven samples, incremental validity. Interestingly, but not surprisingly, the IOP-29 correlated more strongly with other SVTs than with PVTs. In terms of incremental validity, the models improved statistically significantly whenever the IOP-29 was entered into the second step of hierarchical logistic regression models predicting group membership (valid/credible versus invalid/noncredible) after the TOMM, Fifteen Item Test, MMPI-2 F scales, PAI validity scales, or SIMS were entered in the first step. That is, the use of these other SVTs or PVTs along with the IOP-29 significantly improved classification accuracy compared with the use of these other SVTs or PVTs alone. In summary, Giromini and Viglione's (2021) "findings confirm that the IOP-29 could be a useful addition to the toolbox of assessors performing multi-method assessment of symptom or performance validity" (p. 6).

The SRSI

The seventh article in the Special Issue, written by Merten et al. (2021), describes another relatively new SVT: the SRSI. Like other self-report measures of NRB, the SRSI is designed to detect invalid or excessive symptom reports by examining the test-taker's willingness to endorse not only potentially genuine symptoms, but also bizarre, atypical, extreme, or infrequently occurring symptoms (i.e., "pseudosymptoms"). However, unlike the typical free-standing SVT, a unique feature of the SRSI is that it contains not only a set of items describing pseudosymptoms, but also an equal number of items describing relatively common, potentially genuine symptoms. This methodological decision aims to make the actual measurement intent of the instrument less obvious and thus potentially increase its robustness to coaching attempts.

The SRSI includes 107 items: two are warm-up items, five assess consistency, 50 describe potentially genuine symptoms, and 50 describe pseudosymptoms. As Merten et al. (2021) write in their review article, its main purpose is "to detect noncredible symptom endorsement (overreporting) in forensic and clinical patients presenting symptomatology from a spectrum of what may be called "soft" psychopathology (Plomin, 1986), in contrast to the presentation of psychotic, confusional, amnestic, dementia-like symptoms, or intellectual disability" (p. 6). From a conceptual standpoint, "the SRSI can best be seen as a psychometric relative of the SIMS" (p. 6). As such, its psychometric properties have been tested primarily using the SIMS as a "gold standard" or validity criterion. On the other hand, it should be noted that one of the main reasons for developing the SRSI was precisely to overcome some of the well-known limitations of the SIMS (e.g., its exclusive coverage of noncredible symptoms, the resulting potential vulnerability to coaching, etc.).

Originally developed in German, the SRSI is now available in ten languages, namely German, Dutch, French, Norwegian, English, Russian, Portuguese, Italian, Serbian, and Spanish. In their review article, Merten et al. (2021) cite two published cross-validation or equivalence studies in which the German, French, and Dutch versions showed particularly encouraging results. Thus, initial research suggests that the SRSI may be similarly valid in different cultural environments and contexts. However, Merten et al. (2021) emphasize that both researchers and practitioners should adhere to the conditions for using the instrument, in that any deviation (e.g., Internet administration as opposed to the standard and recommended paper-and-pencil format) could affect the outcome both at the level of individual decision making and in terms of the instrument's research database.

Regarding convergent validity, Merten et al. (2021) describe the results of a number of studies conducted during the development of the instrument, and of two additional studies published after the test manual was finalized. Taken together, the results of all these studies indicate that, as expected, the SRSI correlates strongly ($.72 \leq r \leq .82$) with the SIMS and with some of the validity scales of the MMPI instruments (e.g., the Fr of the MMPI-2-RF), whereas the correlations with PVTs fall in the small to medium range. Of

note, in one of the studies cited in the test manual, the SRSI correlated .73, .68, and .55, respectively, when considering MMPI-2-RF scales RBS, Fs, and FBS.

Finally, regarding incremental validity, Merten et al. (2021) state that: "up until now, no study has explicitly focused on incremental validity of the SRSI. Arguably, given the close relationship between the two instruments, no (or, at most, only a subtle) incremental validity is expected between SRSI and SIMS scores. This might not be the case with SVTs that resort to different approaches, such as some of the validity scales of the MMPI family or the Inventory of Problems–29 (IOP-29; Viglione et al., 2017; Viglione & Giromini, 2020)" (p. 5).

The MCI

The eighth and final article in this Special Issue, by Armistead-Jehle and Shura (2021), is a review of the psychometric properties of the MCI. The MCI consists of 58 computerized items and includes six scales that capture plausible memory complaints and three scales that capture implausible memory complaints. The average score of all MCI scales provides information about the overall credibility of self-reported memory problems.

In their review article, Armistead-Jehle and Shura (2021) point out that in addition to the data reported in the test manual, five other published studies provide valuable information about the validity of the MCI. Taken together, the results of all of these studies suggest that the MCI has (a) a weak association with test-taker performance on objective measures of verbal memory, (b) a moderate association with scores on various memory-based PVTs such as the WMT, and (c) a strong association with scores on various SVTs. Armistead-Jehle and Shura (2021) also note that, although the MCI is currently available in six languages (English, Dutch, Spanish, French, Portuguese, and German), there are no published studies in which the translated, non-English versions were used.

When tested against memory-based PVT criteria, the MCI has often shown satisfactory specificity (i.e., $\geq .90$) but suboptimal sensitivity (i.e., $< .40$). However, Armistead-Jehle and Shura (2021) correctly point out that "the MCI is conceptually a symptom validity test" (p. 6). Consistent with this reasoning, a study by Armistead-Jehle et al. (2016) designed specifically to assess the diagnostic accuracy of the MCI compared with PVTs and SVTs showed that the average score of all MCI scales generated *AUC* values of .72 to .75 when using PVT-based criteria and of .77 to .86 when using SVT-based criteria. Remarkably, and somewhat surprisingly, *AUC* values were particularly high when the MMPI-2-RF Fr (.86) or the PAI NIM (.85) was used as the criterion variable, and relatively lower when the MMPI-2-RF RBS (.80) or the MMPI-2-RF FBS-r (.77) was used as the criterion variable. That is, the MCI appears to associate more strongly with SVTs that address overreporting of general psychopathology (e.g., MMPI-2-RF Fr) than with SVTs that address overreporting of memory complaints (e.g., MMPI-2-RF RBS).

Although Armistead-Jehle and Shura (2021) do not present specific statistical data on the extent to which the MCI adds incremental validity over existing PVTs or SVTs, they argue that a unique value of the MCI is that it is one of the few free-standing SVTs that focuses specifically on self-reported memory complaints. Thus, the MCI is one of the few available ways to test the credibility of reported memory problems using an assessment method that is an alternative and, to some extent, a complement to memory-based PVTs.

General Conclusions

This Special Issue contains articles examining the psychometric properties of 19 different indicators of NRB. Five of these indicators (and their five counterparts) are embedded in the two most recent versions of the MMPI instruments (i.e., MMPI-2-RF and MMPI-3), six are embedded in the PAI, three are embedded in the MCMI-IV, one is embedded in the TSI-2, and four are free-standing SVTs, i.e., the SIMS, the IOP-29, the SRSI, and the MCI (Table 2). These 19 indicators differ from each other in many ways, including the clinical domains they focus on, the detection strategies they use, the available research literature, the classification accuracy they have demonstrated in published studies, their convergent and incremental validity, etc. However, some general considerations and common trends can be identified.

SVTs Tend to Correlate More Strongly with SVTs than with PVTs

First, SVTs tend to correlate more strongly with other SVTs than with PVTs. This observation may seem obvious due to shared method variance among SVTs, but the implications might be less obvious. In a meta-analysis of 41 studies (154 effect sizes) on the relationship between self-assessed and psychometrically measured cognitive abilities, Freund and Kasten (2012) reported an average effect size of $r = .33$. The correlation between self-assessments and performance-based outcomes becomes even weaker when memory abilities are considered. For example, in a meta-analysis of 107 studies (673 effect sizes) that looked at the relationship between memory self-efficacy and memory performance in healthy adults, Beaudoin and Desrichard (2011) reported an average effect size of $r = .15$. In older adults, the relationship between subjective and objective memory performance drops further to $r = .06$, according to a meta-analysis of 53 studies (109 effect sizes) published by Crumley et al.

Table 2 Indicators of NRB reviewed by the articles in the Special Issue

Instrument	Indicator of NRB (Acronym)
MMPI-2-RF	Infrequent Responses (F-r)
	Infrequent Psychopathology Responses (Fp-r)
	Infrequent Somatic Responses (Fs)
	Response Bias Scale (RBS)
	Symptom Validity Scale (FBS-r)
MMPI-3	Infrequent Responses (F)
	Infrequent Psychopathology Responses (Fp)
	Infrequent Somatic Responses (Fs)
	Response Bias Scale (RBS)
	Symptom Validity Scale (FBS)
PAI	Negative Impression Management (NIM)
	Malingering Index (MAL)
	Rogers Discriminant Function (RDF)
	Negative Response Distortion Scale (NDS)
	Hong Malingering Index (HMI)
	Multiscale Feigning Index (MFI)
MCMI-IV	Disclosure (X)
	Desirability (Y)
	Debasement (Z)
TSI-2	Atypical Response (ATR)
SIMS	Total score
IOP-29	False Disorder probability Score (FDS)
SRSI	Total pseudosymptoms
MCI	Average score of all MCI scales

(2014). In terms of assessing symptom validity, this means that bona fide individuals who are tested for possible cognitive impairment (particularly if it is associated with memory problems) are likely to have some inconsistency in their results when taking performance-based tests versus self-report tests. Relatedly, if someone engaged in malingering is going to demonstrate the exact same level of impairment on a performance-based test and on a self-report test that assesses the same abilities, it is reasonable to expect that the results of these two tests will differ, at least to some degree. Consistent with these considerations, SVT and PVT test scores typically load on separate factors in factor analytic studies (e.g., Van Dyke et al., 2013), and experts tend to agree that performance validity and symptom validity should be assessed separately (Sweet et al., 2021).

From a practical perspective, this conceptual (and statistical) distinction means that the forensic evaluator should not expect the SVTs and PVTs administered to necessarily match in terms of the overall credibility of the complaints presented. In fact, it is rather unusual for an evaluee to fail both types of validity tests (Sabelli et al., 2021). For example, Shura et al. (2021) recently reported that of 417 veterans who completed the WMT and PAI after deployment, 20.4% produced invalid scores on the WMT (independent of PAI scores), 13.8% produced an invalid PAI (independent of WMT scores), and only 4.6% were invalid on both tests. SVTs and PVTs simply provide a different type of information about the mental and psychological state of the person being assessed. Thus, in the event of a discrepancy, the forensic evaluator should not conclude that one type of validity check is probably correct and the other is probably not, but should try to understand these discrepancies from a clinical perspective: In the context of all available information, why does this particular person in this particular evaluation show credible results on one type of validity test but noncredible results on the other?

In terms of research implications, researchers who want to study the psychometric properties and effectiveness of SVTs should avoid using PVT scores as the only criterion variables to form their credible and noncredible groups, otherwise estimates of classification accuracy would likely be biased (most likely, underestimated). The optimal criterion variables in SVT research are SVTs, or maybe SVTs combined with PVTs, but not PVTs alone. Not even when the target construct is related in some way to cognitive performance, as is the case with measures such as the MCI or the RBS of the MMPI-2-RF/MMPI-3. Consistent with this position, Armistead-Jehle and Shura (2021), in reviewing the psychometric properties of the MCI, found that the MCI, although developed to assess the credibility of presented memory problems, actually correlates more strongly with SVTs that measure overreporting of general psychopathology (e.g., the F-r scale of the MMPI-2-RF) than with PVTs that focus more specifically on memory. Similarly, in an MMPI-3 study by Tylicki et al. (2020), it was found that although the RBS had the highest effect size of all MMPI-3 validity scales in identifying probable/definite malingering based on Sherman et al.'s (2020) criteria, Fs, FBS, and RBS correlated more strongly with other SVTs (from $|r|=.30$ to $|r|=.60$) than with PVTs (from $|r|=.03$ to $|r|=.33$).

One SVT Failure Might Not Be Enough

In a seminal article by Larrabee (2008), it was shown that although PVTs are typically calibrated to achieve a minimum specificity of 90%, when multiple PVTs are used in the same assessment, the probability of randomly failing a PVT despite a valid presentation (i.e., false positive) is too high compared to the standards required in forensic work. Accordingly, it has been repeatedly suggested that—unless the performance on a single PVT is in the significantly below-chance range—examiners

should consider ≥ 2 failures (or even ≥ 3; Larrebee et al., 2019) failures as the standard for concluding that a given presentation is invalid (Boone, 2013; Davis & Millis, 2014a, 2014b; Larrabee, 2008, 2012, 2014; Sherman et al., 2020; Victor et al., 2009). In particular, the revised Multidimensional Malingering Criteria for Neuropsychological Assessment (Sherman et al., 2020) include the following general recommendation, "If only a relatively small number of PVTs are administered, use a criterion of at least two or more PVT failures as indicative of invalid performance" (p. 747).

However, in discussing SVT results, Sherman et al. (2020) stated, in that same article, that "on theoretical grounds, one SVT failure could be deemed sufficient to determine the presence of invalid self-reported symptoms because it is based on a sufficiently large sample of self-reported behaviors" (p. 752). Later on the same page, the authors then continue by saying, "In the absence of specific guidance from the research literature on the optimal number of SVTs to administer or on the optimal number of SVT failures required for the detection of malingering, although we recommend administering more than one SVT which could be accomplished by administering one psychological scale with more than one embedded SVT score, the new model requires only one SVT failure for invalid responding" (p. 752).

In our opinion, the articles included in this Special Issue challenge the validity of the Sherman et al. (2020) position in this regard. First, because not all SVTs actually achieve the minimum target specificity of 90%—for example, when using the traditional cutoffs of ≥ 15 or ≥ 17, the SIMS has repeatedly shown suboptimal specificity in empirical research studies (van Impelen et al., 2014). Second, but relatedly, because we find that there is substantial variation across SVTs. Thus, a high score on an SVT with questionable specificity (e.g., the SIMS or the ATR of the TSI-2) or with a poor research base (e.g., the Modifying Indices of the MCMI-IV) should not be equated with a high score on an SVT with more optimal psychometric properties or research foundation. In addition, the extent to which the evaluee's score deviates from the proposed cut score should also be considered: SVT failures with scores closer to the recommended cut scores should be considered less problematic than those with more extreme departures. For all these reasons, we believe that further research is needed to confirm the position of Sherman et al. (2020) that only one SVT failure is required for invalid responding. Indeed, generally, we recommend against this position when possible, e.g., by giving more SVTs than less, as discussed next.

The Issue of Redundancy in Symptom Validity Assessment

In their revised Multidimensional Malingering Criteria for Neuropsychological Assessment, Sherman et al. (2020) also noted that evaluators must be aware of the issue of redundancy when counting the number of PVT failures. As a general heuristic, Sherman et al. (2020) stated, "Derived scores from the same PVT will necessarily include shared variance if these are based on the same items, such as the use of consistency scores, ratio scores, and immediate/delayed trials. If these are treated as independent PVT scores, this introduces redundancy. For example, the three main Effort scores from the Word Memory Test (Green, 2003) would not constitute independent PVTs because of high shared variance including both shared response format and administration format. Similarly, immediate and delayed trials of the same PVT, such as the TOMM (Tombaugh, 1996), would also not be considered independent PVTs for demonstrating failure on more than one PVT" (p. 747). Later in the same article, the authors then concluded, "In the revised criteria, it is specified that PVTs not be redundant. The field does not yet have clear guidelines on what the degree of maximal shared variance between two PVTs should be, but we would propose that PVTs that tap the same item pool or consist of derived scores from the same items would not be considered independent." (p. 748).

In this regard, the articles in this Special Issue suggest that SVTs from different tests may actually be even more redundant with each other than multiple SVTs embedded in the same instrument. For example, the review article by Shura et al. (2021) cites one study in which the SIMS was strongly correlated with scores on the PAI NIM at $r = .84$ and another in which it was strongly correlated with the MMPI-2 F at $r = .84$. In contrast, the review article by Kurtz and McCredie (2021), which focused on the PAI, showed that the RDF correlated with the other two standard SVTs embedded in the PAI (i.e., NIM and MAL) with $r \leq .38$ and $r \leq .11$, respectively, when considering the community adult normative sample and the clinical standardization sample described in the professional manual. Taken together, then, these results suggest that the fact that two SVTs are embedded in the same test is not prima facie evidence that they are excessively redundant with each other, just as the fact that two SVTs are from different instruments is not per se evidence that they are nonredundant.

In probability theory, the likelihood that a number of events with two outcomes (i.e., Bernoulli trials) that are correlated with each other occur together can be calculated using a correlated binomial distribution. Thus, to provide

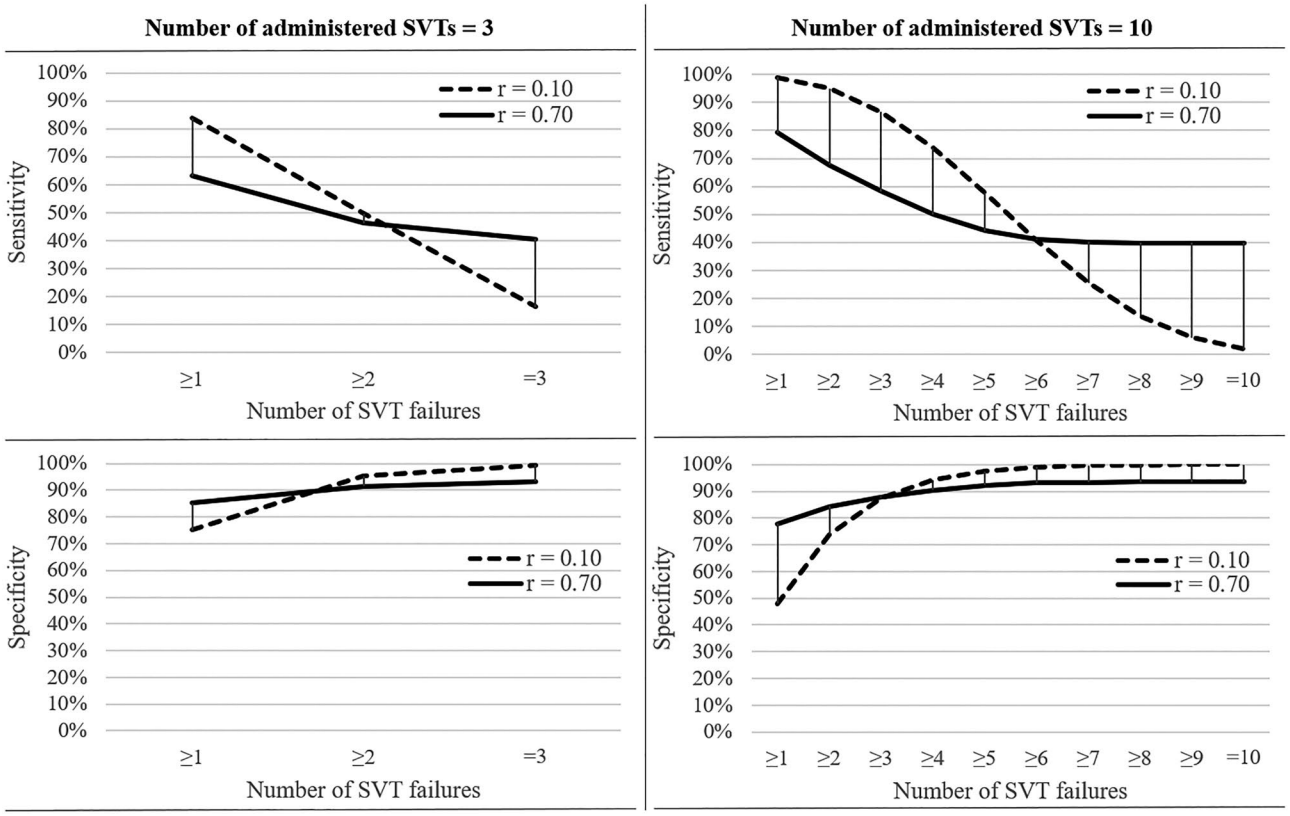

Fig. 1 Examples of battery-wise sensitivity and specificity values for various symptom invalidity criteria Solid lines represent battery-wise sensitivity and specificity values calculated assuming that administered SVTs correlate with each other at $r = .70$; dashed lines represent battery-wise sensitivity and specificity values calculated assuming that administered SVTs correlate with each other at $r = .10$. Battery-wise sensitivity and specificity statistics were calculated considering a Moody's correlated binomial default distribution (Witt, 2004), with individual SVT sensitivities set at 50% and individual SVT specificities set at 90%

the reader with a better theoretical framework for the issue of redundancy in symptom validity assessment, we calculated the battery-wise sensitivity and specificity values that would be obtained if the SVTs included in a battery were correlated with each other at either $r = .10$ or $r = .70$. We considered two possible scenarios: In one case, three SVTs were administered; in the other case, ten SVTs were included. In both cases, we assumed that each SVT had an individual sensitivity of .50 and an individual specificity of .90 (Larrabee limit; Crisan et al., 2021; Erdodi et al., 2014). Therefore, in our model, for each SVT, the probability of a false negative by pure chance would be .50, and the probability of a false positive by pure chance would be .10. The results of these analyses, performed by referring to Moody's correlated binomial default distribution (Witt, 2004), are shown in Fig. 1.

Based on these projections, when considering the "≥ 1 SVT failures" and "≥ 2 SVT failures" criteria, the higher the redundancy in the included SVTs, the lower the battery-wise sensitivity and the higher the specificity. Noteworthy, the loss in sensitivity is generally higher than the gain in specificity. Particularly when considering the scenario in which ten SVTs are included, the battery including less redundant SVTs (i.e., when $r = .10$) tends to generate a substantially higher sensitivity. At least on a theoretical ground, thus, in assessing the overall presentation of the case at hand, the forensic expert should consider not only the number of SVT failures, but also the amount of collinearity.

Although these theoretical models based on probability theory may not apply to the field of symptom validity assessment (for a thorough discussion on this topic, please see Chafetz, 2020), they overall suggest that it might be advantageous to administer SVTs that are not excessively redundant. The models also predict that as the number of SVTs administered increases, so does the battery-wise sensitivity. In addition, when a high number of SVTs are administered (e.g., ten), the one SVT failure criterion might be too liberal and generate an unacceptably low (i.e., < 90) battery-wise specificity. Conversely, if only a few SVTs (e.g., < 4) are administered, the ≥ 2 SVT failures criterion might be too conservative at the expense of sensitivity.

In any case, these are only very general and rather abstract considerations. The correlation values—and, by extension, the strength of associations between multiple SVTs—are indeed dependent on numerous factors that cannot be fully addressed here. To name just a few, these statistical values vary depending on the nature of the populations and diagnostic targets considered, the evaluation context (e.g., criminal versus civil), the research setting (e.g., simulation versus criterion group designs), etc. Thus, while it is important to understand that failing by pure chance (i.e., despite a valid presentation) two highly correlated SVTs out of two administered SVTs is mathematically more likely than failing by pure chance two relatively independent SVTs out of two administered ones, further research is needed to provide more detailed recommendations on this topic.

Is the Number of Failures Really that Important?

In the previous sections, we suggested that (1) one SVT failure may or may not be sufficient to infer that a symptom presentation is invalid, depending on the SVT(s) under consideration; (2) when using multiple SVTs, failure of two SVTs that are less redundant with each other is more unlikely and problematic than failure of two SVTs that are somehow connected. Nevertheless, the number of failed SVTs is unlikely to be one of the most important indicators to consider in assessing NRB. Some SVTs have been studied more intensively than others, some SVTs have been shown to be more optimal in a particular assessment context (e.g., psychological injury evaluations, competency to stand trial, etc.) than in others and/or in a particular diagnostic target (e.g., mTBI, psychosis, depression, etc.) than in others, some combinations of SVTs may be more effective than others, etc. Thus, when counting the number of SVT failures, the forensic evaluator should not implicitly assume that all SVTs are the same and that one failure is equivalent to any other in the battery used. There are some important differences among SVTs that cannot be ignored, and the articles included in this Special Issue are a useful resource to help the professional recognize them.

In addition, although different authors have different opinions on this issue (see, in particular, the results reported by Davis & Millis, 2014a, 2014b), it may not be an optimal solution to count the number of failed SVTs without considering the number of SVTs administered. For example, if an examinee fails two SVTs out of two SVTs administered, does that provide the same level of evidence as when the examinee fails the same two SVTs out of 10 SVTs administered? In this regard, in performance validity assessment, many embedded PVTs are dispersed throughout the examination (Boone, 2009, 2013), along with free-standing ones. Approaches such as Larrabee's (2012), that statistically determine that "failing" two PVTs in a battery is sufficient to determine that the entire profile is invalid no matter the number of PVTs administered, might make statistical sense for the formula used, but would have to confront the issue of face validity. For example, in court the assessor's decision may be challenged on rational grounds: "How could someone who passed 10 of 12 PVTs be considered to express test invalidity"? The same face validity question applies to SVTs too, so that the number of tests administered should perhaps be considered, too. This question constitutes a ripe area of research for the field, considering the multiple questions, pitfalls, and uncertainties on the matter. It would be premature to offer firm practice recommendations for all contexts given the present state of research in the field.

Similarly, the very concept of "SVT failure" can be questioned. As mentioned in the introductory section of this article, there is a great deal of variability in what the "optimal cut score" might be for a given SVT. For example, the professional manual of the SIMS recommends that ≥ 15 be considered the standard cut score to assess the credibility of the symptoms presented (Smith & Burger, 1997). However, subsequent meta-analytic research has shown that a more optimal cut score would likely be ≥ 17 van Impelen et al., 2014). Thus, which of these two cut scores should be used to determine SIMS failure? What if a person scores 16? The answers to these and many other similar questions are not easily addressed in this Special Issue. More research is needed to guide practitioners and researchers in this regard.

Recommendations for Practice

Although this article and the Special Issue it introduces do not provide conclusive solutions to the complex questions the professional faces when performing symptom validity assessments, we would like to share some general recommendations for practice based on what we have learnt from this topic review. First, in an evaluation, the assessor needs to list all embedded and free-standing SVTs and PVTs used and which ones were failed at levels either indicated in the test manuals or, if it applies, in the most recent reliable and valid research. Next, they should use an approach like Sherman et al.'s (2020) or Erdodi's (2019) to indicate different criteria for test failure and where the evaluee falls with respect to them. Moreover, they should review the literature that applies to the specific psycho-legal question, propose multiple hypotheses, and settle on the alternative for which the evidence most strongly supports for the index evaluee. More specifically, they should consider all sources of information and arrive at the opinion that is best supported when considering all the reliable data gathered in the assessment, including with respect to the PVTs and

SVTs. Indeed, the results of SVTs are only one of many important data that the assessor must consider when determining the credibility of a particular clinical presentation.

The field needs general guidelines to deal with the number PVT and SVT failures; the systems developed by Sherman et al. (2020) and Erdodi (2019, 2021) are good starting points. Others have been developed to include both PVTs and SVTs (Bianchini et al., 2005; Young, 2014, 2015). None of these systems, including the widely used prior version of the MND (Slick et al., 1999), have been tested in toto to determine their reliability and validity and should be used with special care and justification. The Sherman et al. (2020) article indicates that the MND has been widely researched and put into practice. However, the changes the system has undergone from one version to the next and the open questions it still poses, as discussed here, indicates that further work is required. Moreover, the other systems mentioned here should be considered for their relative advantages. For example, the Malingered Pain-Related Disability (MPRD; Bianchini et al., 2005) is aimed at a specific condition, that of chronic pain; Young's system includes 60 points on its implementation, and the Erdodi system is quantitative and can apply in most circumstances. Future research needs to evaluate the relative merits of these various approaches more directly before this issue can be settled.

Next, we make a tentative proposal for the field to consider. The field considers validity test failure according to empirically established cut-offs. But these are not etched in stone and, as has been shown herein, can vary according to the most recent research, the context of the evaluation, etc. Considering test invalidity over multiple symptom validity tests should also be considered as variable and contextual, for example, depending on the tests administered, the referral question or context, etc. In PVT research, suggestions have been made to consider more than the dichotomous test validity/invalidity decision. According to the American Academy of Clinical Neuropsychology position paper (Guilmette et al., 2020), the field should consider an "indeterminate" category placed between tests being valid or invalid, which is consistent with prior recommendations (Bigler, 2015; Erdodi, 2019; Young, 2014). For the SVT field, we are recommending multiple SVT administration in forensic and related disability assessments, and would like to see research on the value of considering a similar "indeterminate" range as the third outcome in addition to "valid" and "invalid." This suggestion would also afford the field time to work out the contentious issues of how many SVTs should be administered in a case at hand, what number of failures constitutes invalidity, etc.

To conclude the article, we summarize by stating that we have accomplished several goals, and they point the way to research and practice directions. First, we have reviewed articles on the most commonly used SVTs in psychological injury and related forensic evaluations. We have pointed out their relevant psychometric properties, and their advantages and disadvantages, thereby orienting future research and their application in practice. Second, we have reviewed the most contentious issues in the field, including on the number of SVTs to use in an assessment, choosing the most appropriate multivariate cutoff, collinearity, relative independence/dependence, and the logic and foundational assumptions underlying their use. Thus, the article represents both a plateau of where we stand on SVTs and a steep slope the field has climb through future research and practice.

Declarations

Conflict of Interest Luciano Giromini owns a share in the corporate (LLC) that possesses the rights to Inventory of Problems. Gerald Young have published books on malingering, symptom, and performance validity and received royalties for this. Martin Sellbom is a paid consultant to the University of Minnesota Press, publisher of the MMPI-3.

References

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

American Psychological Association. (2013). Specialty guidelines for forensic psychology. *The American Psychologist, 68*(1), 7–19. https://doi.org/10.1037/a0029889

Anastasi, A. (1988). *Psychological testing* (6th ed.). Macmillan.

Armistead-Jehle, P., & Shura, R. D. (2021). Memory complaints inventory: Review of psychometric properties. *Psychological Injury and Law, Advance Online Publication.* https://doi.org/10.1007/s12207-021-09430-0

Armistead-Jehle, P., Grills, C. E., Bieu, R., & Kulas, J. (2016). Clinical utility of the Memory Complaints Inventory to detect invalid test performance. *The Clinical Neuropsychologist, 30*(4), 610–628.

Axelrod, B. N., Fichtenberg, N. L., Millis, S. R., & Wertheimer, J. C. (2006). Detecting incomplete effort with Digit Span from the Wechsler Adult Intelligence Scale – Third Edition. *The Clinical Neuropsychologist, 20(3)*, 513–523.

Babikian, T., Boone, K. B., Lu, P., & Arnold, G. (2006). Sensitivity and specificity of various Digit Span scores in the detection of suspect effort. *The Clinical Neuropsychologist, 20*(1), 145–159.

Beaber, R., Marston, A., Michelli, I., & Mills, M. (1985). A brief test for measuring malingering in schizophrenic individuals. *The American Journal of Psychiatry, 142*, 1478–1481.

Beaudoin, M., & Desrichard, O. (2011). Are memory self-efficacy and memory performance related? *A Meta-Analysis. Psychological Bulletin, 137*(2), 211.

Ben-Porath, Y. S. (2013). Forensic applications of the Minnesota Multiphasic Personality Inventory-2-Restructured Form. In R. P. Archer & E. M. A. Wheeler (Eds.), *Forensic uses of clinical assessment instruments* (pp. 63–107). Routledge/Taylor & Francis Group.

Ben-Porath, Y. S., & Tellegen, A. (2008). *Minnesota Multiphasic Personality Inventory-2-Restructured Form: Manual for administration, scoring and interpretation*. University of Minnesota Press.

Ben-Porath, Y. S., & Tellegen, A. (2020a). *MMPI-3 Manual for administration, scoring, and interpretation*. University of Minnesota Press.

Ben-Porath, Y. S., & Tellegen, A. (2020b). *MMPI-3 Technical manual*. University of Minnesota Press.

Ben-Porath, Y.S., Heilbrun, K., & Rizzo, M. (in press). Using the MMPI-3 in legal settings. Journal of Personality Assessment.

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

Bigler, E. D. (2015). Neuroimaging as a biomarker in symptom validity and performance validity testing. *Brain Imaging and Behavior, 9*(3), 421–444.

Boccaccini, M. T., & Hart, J. R. (2018). Response style on the Personality Assessment Inventory and other multiscale inventories. *Clinical Assessment of Malingering and Deception, 4*, 280–300.

Boone, K. B. (2009). The need for continuous and comprehensive sampling of effort/response bias during neuropsychological examination. *The Clinical Neuropsychologist, 23*(4), 729–741. https://doi.org/10.1080/13854040802427803

Boone, K. B. (2013). Clinical Practice of Forensic Neuropsychology—An evidence-based approach. New York, NY: Guilford.

Briere, J. (1995). Trauma Symptom Inventory (TSI) Professional Manual. Odessa, FL: Psychological Assessment Resources.

Briere, J. (2011). Trauma Symptom Inventory-2nd edition (TSI-2) professional manual. Odessa, FL: Psychological Assessment Resources.

Burchett, D., & Bagby, R. M. (2021). Assessing negative response bias: A review of the noncredible overreporting scales of the MMPI-2-RF and MMPI-3. *Psychological Injury and Law, Advance Online Publication.* https://doi.org/10.1007/s12207-021-09435-9

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., Dahlstrom, W. G., & Kaemmer, B. (2001). Minnesota Multiphasic Personality Inventory—2: Manual for administration, scoring and interpretation (rev ed.). Minneapolis, MN: University of Minnesota.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

Chafetz, M. D. (2020). Deception is different: Negative validity test findings do not provide "evidence" for "good effort. *The Clinical Neuropsychologist*, Epub Ahead of Print. https://doi.org/10.1080/13854046.2020.1840633

Choca, J. P., & Grossman, S. D. (2015). Evolution of the Millon Clinical Multiaxial Inventory. *Journal of Personality Assessment, 97*(6), 541–549. https://doi.org/10.1080/00223891.2015.1055753

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, Advance online publication. https://doi.org/10.1007/s12207-022-09442-4

Cottingham, M. E., Victor, T. L., Boone, K. B., Ziegler, E. A., & Zeller, M. (2014). Apparent effect of type of compensation seeking (disability vs. litigation) on performance validity test scores may be due to other factors. *The Clinical Neuropsychologist, 28(6)*, 1030–1047. https://doi.org/10.1080/13854046.2014.951397.

Crişan, I., Sava, F. A., Maricuţoiu, L. P., Ciumăgeanu, M. D., Axinia, O., Gîrniceanu, L., & Ciotlăuş, L. (2021). *Evaluation of various detection strategies in the assessment of noncredible memory performance: Results of two experimental studies* (p. 10731911211040104). Advance online publication.

Crumley, J. J., Stetler, C. A., & Horhota, M. (2014). Examining the relationship between subjective and objective memory performance in older adults: A meta-analysis. *Psychology and Aging, 29*(2), 250.

Dandachi-FitzGerald, B., Ponds, R. W. H. M., Peters, M. J. V., & Merckelbach, H. (2011). Cognitive underperformance and symptom over-reporting in a mixed psychiatric sample. *The Clinical Neuropsychologist, 25*(5), 812–828. https://doi.org/10.1080/13854046.2011.583280

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

Davis, J. J., & Millis, S. R. (2014a). Examination of performance validity test failure in relation to number of tests administered. *The Clinical Neuropsychologist, 28*(2), 199–214. https://doi.org/10.1080/13854046.2014.884633

Davis, J. J., & Millis, S. R. (2014b). Reply to commentary by Bilder, Sugar, and Helleman (2014 this issue) on minimizing false positive error with multiple performance validity tests. *The Clinical Neuropsychologist, 28*(8), 1224–1229. https://doi.org/10.1080/13854046.2014.987167

Detullio, D., Messer, S. C., Kennedy, T. D., & Millen, D. H. (2019). A meta-analysis of the Miller Forensic Assessment of Symptoms Test (M-FAST). *Psychological Assessment, 31*, 1319–1328.

Dionysus, K. E., Denney, R. L., & Halfaker, D. A. (2011). Detecting negative response bias with the Fake Bad Scale, Response Bias Scale, and Henry-Heilbronner Index of the Minnesota Multiphasic Personality Inventory-2. *Archives of Clinical Neuropsychology, 26*(2), 81–88. https://doi.org/10.1093/arclin/acq096

Edens, J. F., Poythress, N. G., & Watkins-Clay, M. M. (2007). Detection of malingering in psychiatric unit and general population prison inmates: A comparison of the PAI, SIMS, and SIRS. Journal of Personality Assessment, 88(1), 33–42.

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

Erdodi, L. A. (2021). Five shades of gray: Conceptual and methodological issues around multivariate models of performance validity. *NeuroRehabilitation, 49*(2), 179–213. https://doi.org/10.3233/NRE-218020

Erdodi, L. A., & Abeare, C. A. (2020). Stronger together: The Wechsler Adult Intelligence Scale-Fourth Edition as a multivariate performance validity test in patients with traumatic brain injury. *Archives of Clinical Neuropsychology, 35*(2), 188–204. https://doi.org/10.1093/arclin/acz032/5613200

Erdodi, L. A., Abeare, C. A., Medoff, B., Seke, K. R., Sagar, S., & Kirsch, N. L. (2018). A single error is one too many: The Forced Choice Recognition trial on the CVLT-II as a measure of performance validity in adults with TBI. *Archives of Clinical Neuropsychology, 33*(7), 845–860.

Erdodi, L. A., Kirsch, N. L., Lajiness-O'Neill, R., Vingilis, E., & Medoff, B. (2014). Comparing the recognition memory test and the word choice test in a mixed clinical sample: Are they equivalent? *Psychological Injury and Law, 7*(3), 255–263. https://doi.org/10.1007/s12207-014-9197-8

Franzen, M. D., & Iverson, G. L. (2000). Detecting negative response bias and diagnosing malingering: the dissimulation exam. In J.

Snyder, & P. J. Nussbaum (Eds.), Clinical neuropsychology: A pocket handbook for assessment. Washington, DC: American Psychological Association.
Freund, P. A., & Kasten, N. (2012). How smart do you think you are? A meta-analysis on the validity of self-estimates of cognitive ability. *Psychological Bulletin, 138*(2), 296.
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.
Giromini, L., & Viglione, D. J. (2021). *Assessing negative response bias with the inventory of problems – 29 (IOP-29): A quantitative literature review*. Advance online publication.
Giromini, L., Pignolo, C., Zennaro, A., & Viglione, D. J. (2018). A clinical comparison, simulation study testing the validity of SIMS and IOP-29 with an Italian sample. *Psychological Injury and Law, 11*(4), 340–350.
Giromini, L., Viglione, D. J., Pignolo, C., & Zennaro, A. (2020). An inventory of problems-29 sensitivity study investigating feigning of four different symptom presentations via malingering experimental paradigm. *Journal of Personality Assessment, 102*, 563–572.
Green, P. (2003). *Green's Word Memory Test for Microsoft Windows*. Green's Publishing Inc.
Green, P. (2019). *Users' manual for the Memory Complaints Inventory (MCI)*. Green's Publishing.
Green, P., Allen, L. M., & Astner, K. (1996). The word memory test: A user's guide to the oral and computer-administered forms, US version 1.1. Durham, NC: CogniSyst.
Greiffenstein, M. F., Baker, W. J., & Gola, T. (1994). Validation of malingered amnesia measures with a large clinical sample. *Psychological Assessment, 6(3)*, 218–224. https://doi.org/10.1037/1040-3590.6.3.218
Greve, K. W., Curtis, K. L., Bianchini, K. J., & Ord, J. S. (2009). Are the original and second edition of the California Verbal Learning Test equally accurate in detecting malingering? *Assessment, 16(3)*, 237–248.
Guilmette, T. J., Sweet, J. J., Hebben, N., Koltai, D., Mahone, E. M., Spiegler, B. J., & Participants, C. (2020). American Academy of Clinical Neuropsychology consensus conference statement on uniform labeling of performance test scores. *The Clinical Neuropsychologist, 34*(3), 437–453.
Hawes, S. W., & Boccaccini, M. T. (2009). Detection of overreporting of psychopathology on the Personality Assessment Inventory: A meta-analytic review. *Psychological Assessment, 21*(1), 112–124. https://doi.org/10.1037/a0015036
Heilbronner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., Millis, S. R., & Participants, C. (2009). American Academy of Clinical Neuropsychology Consensus Conference Statement on the neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist, 23*(7), 1093–1129. https://doi.org/10.1080/13854040903155063
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.
Howell, D. C. (2013). *Statistical methods for psychology* (8th ed.). Cengage.
Hunsley, J., & Meyer, G. J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. *Psychological Assessment, 15*(4), 446–455. https://doi.org/10.1037/1040-3590.15.4.446
Kurtz, J. E., & McCredie, M. N. (2021). Exaggeration or fabrication? Assessment of negative response distortion and malingering with the personality assessment inventory. *Psychological Injury and Law*, Advance online publication. https://doi.org/10.1007/s12207-021-09433-x
Lanyon, R. I. (2006). Mental health screening: Utility of the Psychological Screening Inventory. *Psychological Services, 3*, 170–180.
Larrabee, G. J. (2008). Aggregation across multiple indicators improves the detection of malingering: Relationship to likelihood ratios. *The Clinical Neuropsychologist, 22*(4), 666–679. https://doi.org/10.1080/13854040701494987
Larrabee, G. J. (2012). Performance validity and symptom validity in neuropsychological assessment. *Journal of International Neuropsychological Society, 18*(4), 625–630.
Larrabee, G. J. (2014). False-positive rates associated with the use of multiple performance and symptom validity tests. *Archives of Clinical Neuropsychology, 29*(4), 364–373. https://doi.org/10.1093/arclin/acu019
Larrabee, G. J., Rohling, M. L., & Meyers, J. E. (2019). Use of multiple performance and symptom validity measures: Determining the optimal per test cutoff for determination of invalidity, analysis of skew, and inter-test correlations in valid and invalid performance groups. *The Clinical Neuropsychologist, 33*(8), 1354–1372.
Lewis, J. L., Simcox, A. M., & Berry, D. T. R. (2002). Screening for feigned psychiatric symptoms in a forensic sample by using the MMPI-2 and the Structured Inventory of Malingered Symptomatology. Psychological Assessment, 14, 170–176.
Merten, T., & Merckelbach, H. (2013). Symptom validity testing in somatoform and dissociative disorders: A critical review. *Psychological Injury and Law, 6*, 122–137. https://doi.org/10.1007/s12207-013-9155-x
Merten, T., Merckelbach, H., Giger, P., & Stevens, A. (2016). The Self-Report Symptom Inventory (SRSI): A new instrument for the assessment of distorted symptom endorsement. *Psychological Injury and Law, 9*(2), 102–111. https://doi.org/10.1007/s12207-016-9257-3
Merten, T., Dandachi-FitzGerald, B., Boskovic, I., Puente-López, E., & Merckelbach, H. (2021). The self-report symptom inventory. Psychological Injury and Law. Advance online publication. https://doi.org/10.1007/s12207-021-09434-w
Miller, H.A. (2001). M-FAST: Miller Forensic Assessment of Symptoms Test professional manual. Odessa, FL: Psychological Assessment Resources
Millon, T., Davis, R., Millon, C., & Grossman, S. (2009). Millon Clinical Multiaxial Inventory-III, 4th ed. (MCMI-III). Minneapolis, MN: Pearson Assessments.
Millon, T., Grossman, S., & Millon, C. (2015). Millon Clinical Multiaxial Inventory IV (MCMI-IV). Minneapolis, MN: Pearson Assessments.
Mogge, N. L., LePage, J. S., Bell, T., & Ragatz, L. (2010). The negative distortion scale: A new PAI validity scale. *Journal of Forensic Psychiatry and Psychology, 21*(1), 77–90.
Morey, L. C. (1991). Personality assessment inventory (PAI). Professional manual. Odessa, FL: Psychological Assessment Resources.
Morey, L. C. (1996). An interpretive guide to the Personality Assessment Inventory (PAI). Odessa, FL: Psychological Assessment Resources.
Morey, L. C. (2003). *Essentials of PAI assessment*. Wiley.
Morey, L. C. (2007). *Personality Assessment Inventory (PAI). Professional manual* (2nd ed.). Psychological Assessment Resources.
Morey, L. C. (2020). PAI Plus: Professional manual supplement. Odessa, FL: Psychological Assessment Resources.
Morey, L. C., & Hopwood, C. J. (2007). Casebook for the personality assessment inventory: A structural summary approach. Odessa, FL: Psychological Assessment Resources.
Morgan, C. D., Schoenber, M. R., Dorr, D., & Burke, M. J. (2002). Overreport on the MCMI-III: Concurrent validation with the MMPI-2 using a psychiatric inpatient sample. *Journal of Personality Assessment, 78*(2), 288–300. https://doi.org/10.1207/S15327752JPA7802_05

Morris, N. M., Mattera, J., Golden, B., Moses, S., Ingram, P. B. (2021). Evaluating the performance of the MMPI-3 over-reporting scales: Sophisticated simulators and the effects of comorbid conditions. *The Clinical Neuropsychologist*, 1–9. https://doi.org/10.1080/13854046.2021.1968037

Neal, T. M. S., & Grisso, T. (2014). Assessment practices and expert judgment methods in forensic psychology and psychiatry: An international snapshot. *Criminal Justice and Behavior, 41*(12), 1406–1421. https://doi.org/10.1177/0093854814548449

Palermo, C. A., & Brand, B. L. (2019). Can the trauma symptom inventory-2 distinguish coached simulators from dissociative disorder patients? *Psychological Trauma: Theory, Research, Practice, and Policy, 11*(5), 477–485. https://doi.org/10.1037/tra0000382

Pearson (2009). Advanced Clinical Solutions for the WAIS-IV and WMS-IV – Technical Manual. San Antonio, TX: Author.

Plomin, R. (1986). *Development, genetics, and psychology*. Lawrence Erlbaum.

Reese, C. S., Suhr, J. A., & Riddle, T. L. (2012). Exploration of malingering indices in the Wechsler Adult Intelligence Scale-Fourth Edition Digit Span subtest. *Archives of Clinical Neuropsychology, 27*, 176–181.

Reeves, C.K., Brown, T.A., & Sellbom, M. (in press). An examination of the MMPI-3 validity scales in detecting overreporting of psychological problems. *Psychological Assessment*.

Rogers, R., & Bender, D. (2018). Clinical assessment of malingering and deception. New York, NY: Guilford.

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

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*, 77–88.

Rogers, R., Sewell, K. W., & Gillard, N. D. (2010). *Structured interview of reported symptoms, second edition: Professional test manual* (2nd ed.). Psychological Assessment Resources.

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. https://doi.org/10.1177/1073191103010002007

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.

Rogers, R., Velsor, S. F., & Williams, M. M. (2020). A brief commentary on SIRS versus SIRS-2 critiques. *Psychological Injury and Law, 13*(3), 275–283. https://doi.org/10.1007/s12207-020-09379-6

Roma, P., Giromini, L., Burla, F., Ferracuti, S., Viglione, D. J., & Mazza, C. (2020). Ecological validity of the Inventory of Problems-29 (IOP-29): An Italian study of court-ordered, psychological injury evaluations using the Structured Inventory of Malingered Symptomatology (SIMS) as criterion variable. *Psychological Injury and Law, 13*(1), 57–65.

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

Sabelli, A. G., Messa, I., Giromini, L., Lichtenstein, J. D., May, N., & Erdodi, L. A. (2021). Symptom versus performance validity in patients with mild TBI: Independent sources of non-credible responding. *Psychological Injury and Law, 14*(1), 17–36.

Schoenberg, M., Dorr, D., & Morgan, D. (2004). A comparison of the MCMI-III personality disorder and modifier indices with the MMPI-2 clinical and validity scales. *Journal of Personality Assessment, 82*(3), 273–280. https://doi.org/10.1207/s15327752jpa8203_03

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). Guilford Press.

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*(3), 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. https://doi.org/10.1093/arclin/acaa019

Shura, R. D., Ord, A. S., & Worthen, M. D. (2021). Structured inventory of malingered symptomatology: A psychometric review. *Psychological Injury and Law, Advance Online Publication.* https://doi.org/10.1007/s12207-021-09432-y

Shura, R. D., Yoash-Gantz, R. E., Pickett, T. C., McDonald, S. D., & Tupler, L. A. (2021). Relations among performance and symptom validity, mild traumatic brain injury, and posttraumatic stress disorder symptom burden in postdeployment veterans. *Psychological Injury and Law, Advance Online Publication.* https://doi.org/10.1007/s12207-021-09415-z

Slick, D. J., Iverson, G. L., & Green, P. (2000). California verbal learning test indicators of suboptimal performance in a sample of head-injury litigants. *Journal of Clinical and Experimental Neuropsychology, 22*(4), 569–579.

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. https://doi.org/10.1076/1385-4046(199911)13:04;1-Y;FT545

Smith, G. P., & Burger, G. K. (1997). Detection of malingering: Validation of the structured inventory of malingered symptomatology (SIMS). *Journal of the American Academy on Psychiatry and Law, 25*, 180–183.

Soble, J. R., Alverson, W. A., Phillips, J. I., Critchfield, E. A., Fullen, C., O'Rourke, J. J. F., & Marceaux, J. C. (2020). Strength in numbers or quality over quantity? Examining the importance of criterion measure selection to define validity groups in performance validity test (PVT) research. *Psychological Injury and Law, 13*, 44–56. https://doi.org/10.1007/s12207-019-09370-w

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

Tombaugh, T. N. (1996). Test of memory malingering (TOMM). New York, NY: Multi Health Systems.

Tsujimoto, R. N., Hamilton, M., & Berger, D. E. (1990). Averaging multiple judges to improve validity: Aid to planning cost-effective clinical research. *Psychological Assessment, 2*, 432–437.

Tylicki, J. L., Gervais, R. O., & Ben-Porath, Y. S. (2020). Examination of the MMPI-3 over-reporting scales in a forensic disability sample. *The Clinical Neuropsychologist, Advance Online Publication.* https://doi.org/10.1080/13854046.2020.1856414

Tylicki, J., Glassmire, D., Tarescavage, A., Wygant, D., & Sellbom, M. (2021). A response to Rogers and Colleagues' (2020) analysis of a "Trio" of SIRS vs. SIRS-2 comparison studies. *Psychological Injury and Law,* Manuscript accepted for publication.

van den Broek, M. D., Monaci, L., & Smith, J. G. (2012). Clinical utility of the Personal Problems Questionnaire (PPQ) in the

assessment of non-credible complaints. *Journal of Experimental Psychopathology, 3*(5), 825–834. https://doi.org/10.5127/jep.024311

Van Dyke, S. A., Millis, S. R., Axelrod, B. N., & Hanks, R. A. (2013). Assessing effort: Differentiating performance and symptom validity. *The Clinical Neuropsychologist, 27*(8), 1234–1246. https://doi.org/10.1080/13854046.2013.835447

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

Victor, T. L., Boone, K. B., Serpa, J. G., Buehler, J., & Ziegler, E. A. (2009). Interpreting the meaning of multiple symptom validity test failure. *The Clinical Neuropsychologist, 23*(2), 297–313. https://doi.org/10.1080/13854040802232682

Viglione, D. J., & Giromini, L. (2020). Inventory of problems–29*:* Professional manual. IOP-Test, LLC.

Viglione, D. J., Giromini, L., & Landis, P. (2017). The development of the Inventory of Problems–29: A brief self-administered measure for discriminating bona fide from feigned psychiatric and cognitive complaints. *Journal of Personality Assessment, 99*(5), 534–544. https://doi.org/10.1080/00223891.2016.1233882

Whitman, M. R., Tylicki, J. L., & Ben-Porath, Y. S. (2021). Utility of the MMPI-3 validity scales for detecting overreporting and underreporting and their effects on substantive scale validity: A simulation study. *Psychological Assessment, 33*(5), 411–426. https://doi.org/10.1037/pas0000988

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment*. Addison Wesley.

Witt, G. (2004). Moody's correlated binomial default distribution. Moody's Investor Service, Special Report, August.

Wolfe, P. L., Millis, S. R., Hanks, R., Fichtenberg, N., Larrabee, G. J., & Sweet, J. J. (2010). Effort indicators within the California Verbal Learning Test-II (CVLT-II). *The Clinical Neuropsychologist, 24*(1), 153–168.

Wygant, D. B., Sellbom, M., Ben-Porath, Y. S., Stafford, K. P., Freeman, D. B., & Heilbronner, R. L. (2007). The relation between symptom validity testing and MMPI-2 scores as a function of forensic evaluation context. *Archives of Clinical Neuropsychology, 22*, 489–499.

Young, G. (2014). Malingering, feigning, and response bias in psychiatric/psychological injury. *International Library of Ethics, Law, and the New Medicine, 56*, 817–856.

Young, G. (2015). Detection system for malingered PTSD and related response biases. *Psychological Injury and Law, 8*(2), 169–183.

Young, G. (2019). The cry for help in psychological injury and law: Concepts and review. *Psychological Injury and Law, 12*(3–4), 225–237. https://doi.org/10.1007/s12207-019-09360-y

Young, G. (2021). The call for aid (cry for help) in psychological injury and law: Reinterpretation, mechanisms, and a call for research. *Psychological Injury and Law, 14*(3), 185–200. https://doi.org/10.1007/s12207-021-09414-0