Psychological Injury and Law
https://doi.org/10.1007/s12207-021-09434-w

The Self-Report Symptom Inventory

Thomas Merten[1] · Brechje Dandachi-FitzGerald[2] · Irena Boskovic[3] · Esteban Puente-López[4] · Harald Merckelbach[2]

Received: 8 August 2021 / Accepted: 2 November 2021

Abstract
The Self-Report Symptom Inventory (SRSI) was developed to expand the toolbox of self-report instruments available to detect symptom overreporting. Such instruments, today known as symptom validity tests, play a crucial role in both forensic evaluations and in a range of clinical referral questions. The SRSI was originally designed in the German language; items were selected from a larger pool on the basis of empirical results. Scores on the Structured Inventory of Malingered Symptomatology served as external criterion for the item selection procedure and empirical cut-score determination (gold standard). The SRSI is composed of five subscales describing potentially genuine symptoms and five pseudosymptoms subscales. Ten different language test versions have been developed so far. The article describes the background of the construction of the scale, the main empirical results with the SRSI, the conditions of use, and the limits of applicability. With research ongoing in several countries and with a variety of language versions, a larger body of empirical evidence can be expected to accumulate in the coming years.

Keywords Self-Report Symptom Inventory · Questionnaire · Symptom validity test · Symptom overreporting · Malingering · Forensic assessment · Psychological assessment

Background, Detection Strategy, and Conditions of Use

Measurement Intention. The aim and the strategy of the instrument can be shortly summarized as follows: it is a self-report questionnaire developed to identify invalid, noncredible excessive symptom report (i.e., symptom overreporting) by investigating the willingness of the patient to endorse not only a high number of potentially genuine symptoms, but also bizarre, extreme, or rarely occurring symptoms. As yet, the genuine symptom subscales can only be analyzed qualitatively; their inclusion was mainly motivated by making the true measurement intention of the instrument (its face validity) less obvious and increase the instrument's robustness against coaching attempts.

Background of Scale Development. Investigating the validity of test profiles and the credibility of reported symptoms is a core issue in forensic evaluations (e.g., Bush et al., 2014; Sweet et al., 2021). Its growing interest is also evident in clinical and rehabilitation contexts (e.g., Carone & Bush, 2018; McWhirter et al., 2020). While performance validity tests (PVTs) aim to detect underperformance, self-report validity tests (today called symptom validity tests, SVTs) evaluate whether the symptomatology claimed is credible or not. Only if this is the case, the symptom report given by an individual patient can be trusted with a sufficient degree of confidence. The following text describes the main aspects of the development of a new SVT and the main research findings obtained with it so far.

The development of a psychometric instrument from scratch to the publication of a professional manual can be lengthy. In the case of the Self-Report Symptom Inventory (SRSI; Merten et al., 2019), it took more than 12 years, countless hours of work, about twenty studies in several European countries including a variety of language versions, with various methodologies and samples, and the commitment of many undergraduate students and professionals,

✉ Thomas Merten
thomas.merten@vivantes.de

[1] Department of Neurology, Vivantes Klinikum im Friedrichshain, Landsberger Allee 49, 10249 Berlin, Germany

[2] Forensic Psychology Section, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands

[3] Department of Psychology, Erasmus University, Education & Child Studies, Rotterdam, The Netherlands

[4] Applied Psychology Service, University of Murcia, Murcia, Spain

Published online: 12 November 2021

mostly neuropsychologists and forensic psychologists, from Germany, The Netherlands, Switzerland, Serbia, Great Britain, Norway, Belgium, Austria, Portugal, and Italy, to name the most important ones.

The history of the scale construction reaches back to 2006 when two psychologists and a psychiatrist analyzed a sample of neuropsychiatric civil forensic cases available from a private practice in southern Germany. For 198 claimants, results on the Structured Inventory of Malingered Symptomatology (SIMS; Widows & Smith, 2005) were available. The number of protocols allowed for an item-wise analysis of the German language version of the SIMS (Cima et al., 2003). For a number of items, the basic quality markers, such as item-total correlations, turned out to be insufficient in that sample (Merten et al., 2007). Moreover, the intercorrelations between the five SIMS subscales appeared to be low. This was a worrisome observation, given that (1) they were summed up to a total score and (2) a separate analysis of the subscales is recommended only for subsequent qualitative analyses (Smith & Burger, 1997). Among the subscale intercorrelations, the lowest one was only 0.17 (between Affective Disorders and Low Intelligence).

One consideration relevant for understanding those results was the nature of the sample (as is always the case with classical test theory analyses). Participants' referral background was neuropsychiatric forensic assessment in the context of civil and social law claims, rather than criminal forensic evaluations. Among the claimants, many presented a history of soft psychopathology (such as dysthymia, chronic fatigue, somatoform pain disorder, or adjustment disorder). Hence, reports of more extreme forms of psychopathology, such as amnesia, intellectual disability, delusions, or other psychotic symptoms, were rare in this population. Thus, the civil forensic context and the sort of self-reported symptoms typically found in such a setting were deemed to be another factor that potentially limited the power of the SIMS. What was missing from an instrument to be used with such a target population was a spectrum of symptoms encompassing domains such as pain, anxiety, fatigue, and milder cognitive impairment (beyond hardcore presentations as amnesia or intellectual disability).

These considerations set the stage for the idea of developing a new instrument targeted at measuring overreporting in both non-criminal forensic and clinical settings, with predominantly soft psychopathological symptom claims by patients. Item selection was planned to be based on a strictly empirical procedure because, through numerous previous test analyses of the first author (e.g., Merten, 2006) and the classical test development literature (e.g., Anstey, 1966; Helmstadter, 1966), it is well known that neither face validity nor expert opinion can reliably predict whether an individual item passes the test of psychometric validity. For item analysis, methods of classical test theory were employed because the authors assumed that distorted presentation of illness in general and symptom overreporting in particular were *not* unidimensional constructs, but rather presented in real-world situations in multiple facets.

Further Ad hoc Considerations for Scale Construction. The two main approaches in questionnaire-based symptom validity assessment are the following: (1) The compilation of a list of multiple symptoms each of which occurs with a non-negligible frequency in patients with psychological problems. If the number of endorsed symptoms exceeds an empirically established limit, this makes the presence of such a heavy and overgeneralized symptom burden unlikely and not credible. An example of a scale following this approach is the Fake Bad Scale of the MMPI family (FBS; Lees-Haley et al., 1991), also called the Symptom Validity Scale. To illustrate this, typical items of this kind might be as follows: *I find it hard to keep concentrated on a given task,* or *I have a great deal of headaches.*[1] (2) The compilation of a list of bizarre, atypical, extreme, or rarely occurring symptoms that seemingly belong to existing symptom domains. Good examples of this approach are provided by many items of the SIMS (with the exception of the Affective Disorder subscale). The construction of the pseudosymptoms subscales of the SRSI also followed this approach. To give an example, an item of the extreme type would be as follows: *On a scale from 0 (no headache) to 10 (maximum headache), it is at "10" almost all the time.* (3) Some tests like the M test (Beaber et al., 1985) combine both approaches.

Simple language to keep educational requirements for respondents low was another important consideration for SRSI item construction. This included easy-to-understand syntax, avoidance of negatively formulated items (to avoid problems with double negations), and avoidance of conditional sentence structures. A further consideration had to do with one of the main problems identified by many users of the SIMS and some other validity measures: Often, these measures are readily identifiable as instruments for the detection of false symptom claims. This was why we opted for disguising the measurement intention by mixing potentially genuine symptoms with pseudosymptoms.

Resulting Scale Structure. Starting from an ad hoc scale structure of the SRSI (Table 1), three forensic experts compiled, on a rational basis, an initial collection of potential items for each genuine symptom and pseudosymptom domain. The details of the subsequent item selection procedure were described by Merten et al. (2016). The empirical item selection procedure was based on the participants' group membership depending on their SIMS scores. The

[1] Both for copyright issues and for test security, the FBS and SRSI items formulated in this paragraph are not real, but made-up items to illustrate their nature.

Table 1 Ad hoc Scale Structure of the Self-Report Symptom Inventory (SRSI)

Domain	Scale	Preliminary version	Final version
Potentially genuine symptoms	Cognitive Symptoms	15 items	10 items
	Depressive Symptoms	15 items	10 items
	Pain Symptoms	15 items	10 items
	Nonspecific Somatic Symptoms	15 items	10 items
	Anxiety Symptoms (including PTSD)	15 items	10 items
Subtotal	*Total Genuine Symptoms*	*75 items*	*50 items*
Pseudosymptoms	Cognitive Pseudosymptoms	15 items	10 items
	Neurological: Motor Pseudosymptoms	15 items	10 items
	Neurological: Sensory Pseudosymptoms	15 items	10 items
	Pain Pseudosymptoms	15 items	10 items
	Mental Pseudosymptoms (Anxiety, Depression, PTSD)	15 items	10 items
Subtotal	*Total Pseudosymptoms*	*75 items*	*50 items*
A priori cooperativeness (warming up)		2 items	2 items
Additional consistency check		5 items	5 items
Embedded index: Ratio Pseudosymptom Score/Genuine Symptom Score		–	–
Total item number		157 items	107 items

Modified from Merten et al. (2016)

PTSD Post Traumatic Stress Disorder

subscales were not intended to constitute independent, distinct or, mathematically speaking, orthogonal dimensions.

Conditions of Use. The questionnaire was constructed and validated as a paper-and-pencil measure, with no evidence so far on what possible effects the computerized or Internet-based administration might have. The paper-and-pencil version is the standard of use for the questionnaire and the application of empirically derived cut scores. Whenever another presentation mode or another answer format is used, this should be made explicit, including potential interpretation problems of results.

The questionnaire should only be given to people who master the given language approximately at the level of native speakers with at least lower secondary formal school education. This usually also applies to bilingual or multilingual people who obtained formal schooling in the language in question. The results obtained from people with lower degrees of language proficiency are potentially contaminated to an unknown degree. Such distorting effects on questionnaire results are often underestimated or neglected; foreign speakers of a language may be able to communicate fluently, but, at the same time, they may not be able to understand colloquial expressions, more subtle meanings, or fine nuances that may be central to the item content and that native speakers usually understand without difficulties. This problem may be of the utmost importance for symptom validity items to work as intended (Lilienfeld et al., 2013; Nijdam-Jones & Rosenfeld, 2017; but see also van der Heide et al., 2017).

The questionnaire should be given in a quiet atmosphere in the presence of the examiner or another qualified person so that potential problems can be solved immediately (as prescribed, in more details, in the test manual, Merten et al., 2019). Although some examiners (as known from Germany) continue to deliver questionnaires by mail to forensic patients to answer at home, this is a harsh violation of both the conditions of use and test security. Potentially highly contaminated results will be obtained. The examiner is responsible for the absence or non-interference of any third party during questionnaire responding. This is also a point of potential contamination if questionnaires are responded through the Internet or other remote-assessment devices.

Total and Subscale Score Interpretation. Up until now, the only scale score for which empirically developed cut scores are available, is the sum total of endorsed pseudosymptoms. The manual (Merten et al., 2019) contains numerous data on the results of potentially genuine symptoms and on the symptoms and pseudosymptoms subscales. They may be used for interpreting the individual test results as compared to reference data. Meanwhile, there are no cut scores for individual pseudosymptoms subscales. Therefore, cutoff-based decision-making is not possible on the level of the single subscales.

The original concept underlying the SRSI construction, the item selection, the empirical cutoff optimization, and empirical results so far were centered on detecting people who overgeneralize symptom claims across different domains. Many patients with noncredible symptom claims

in the area of soft psychopathology engage in this kind of overgeneralization, but certainly not all of them.

As a consequence, individual subscale item sets should not be given in isolation or be implemented into other scales or instruments, with the exception of special research questions where this might be of interest (e.g., Boskovic et al., 2019). This limitation also applies to the pain symptoms and the pain pseudosymptoms subscales. Up until now, it is not clear how the SRSI performs with patients who grossly exaggerate pain, but do not extend unjustified symptom claims to other domains (of soft psychopathology). Germane to this is a study by Boskovic et al. (2020) relying on instructed malingerers, which found a relatively modest detection rate (of only 48%) in students who were instructed to simulate selectively pain complaints ostensibly caused by a motor vehicle accident. At the same time, one may argue that those 48% constituted a *relatively* high hit rate considering that (1) the scenario was very selective inasmuch as participants were asked to pretend suffering only from pain, not from other health complaints, and (2) the SRSI is a measure of overreporting working best when symptom claims are overgeneralized over different symptom domains.

Foreign Language Adaptations. The preliminary 157-item questionnaire version was exclusively available in German. Soon after the empirically based item selection and construction of the final version, the first versions in foreign language were developed. They were based on a multiple-step procedure of translation, back-translation, and fine-tuning of unclear items. For some languages, professional translators were employed. One of the first foreign-language versions was the French version (Geurten et al., 2018). Giger and Merten (2019) performed a study with Swiss bilingual participants investigating the equivalence of the German and French SRSI versions, with very encouraging results. Recently, similar results have been obtained in an equivalence study that compared the German and Dutch SRSI versions (Pienkohs, 2021). The ten languages for which SRSI versions were developed so far are German, Dutch, French, Norwegian, English, Russian, Portuguese, Italian, Serbian, and Spanish.

Convergent and Incremental Validity

The construction strategy of the SRSI followed, in some important points, the one known from the SIMS (except for the Affective Symptoms subscale of the latter), and the empirical selection of items was based on SIMS total scores of a group of 239 respondents who were given the preliminary test version. As a consequence, high correlations with SIMS scores were expected (and finally obtained) both for the preliminary and for the final version of the SRSI.

The preliminary 157-item version was tested with a mixed sample ($N = 239$) of participants from different studies and referral backgrounds (mostly independent medical examinations [IME] and experimental analogue studies). SRSI pseudosymptom endorsement correlated at 0.81 with the SIMS total scores, and at 0.62 with the scores on the MMPI-2 Fake Bad Scale (Lees-Haley et al., 1991).

Giger and Merten (2013) collected data on the preliminary SRSI version as well as on six more SVTs and PVTs in a demographically representative population sample of 100 German-speaking Swiss adults from 18 to 60 years old. They found a very low rate of positives on all validity measures. The average number of endorsed SRSI pseudosymptoms was 2.0 ($SD = 0.7$).

After empirical item selection and construction of the final 107-item test version, sixteen more studies were conducted with the SRSI before the publication of the comprehensive test manual (Merten et al., 2019). They encompassed a variety of designs, sample characteristics, and comparison instruments. Next to healthy participants instructed to answer honestly, IME patients, analogue malingerers (with a variety of different scenarios, symptom information, and warning conditions), clinical patients, and sentenced prison inmates were studied with the SRSI.

At an intermediate stage of test construction, a combined sample of 520 participants from different studies was analyzed. For three studies, SIMS scores were available. SRSI pseudosymptoms endorsement correlated at 0.82 with total SIMS scores ($n = 367$). Depending on the SRSI cut score employed for classification (see below), a good concordance with SIMS classification was observed (with phi scores ranging from 0.53 to 0.68).

For three studies, data gathered with different PVTs were available at that stage, such as the Amsterdam Short-Term Memory Test (ASTM; Schmand & Lindeboom, 2005), the Malingering Scale (MgS; Schretlen et al., 1992), and the Word Memory Test (WMT; Green, 2003). Correlations of these PVTs with SRSI pseudosymptom endorsement were in the small to medium size range. For example, for a group of neuropsychiatric IME patients ($n = 207$), SRSI pseudosymptoms and Word Memory Test performance correlated at −0.45 (the negative sign indicates a positive correlation between underperformance and overreporting).

Symptom and performance validity measures refer to conceptually related, but different constructs (overreporting and underperformance, respectively), with the common denominator of both allowing determinations about the validity of test results (in self-report scales and performance tests, respectively). This has repeatedly been shown by a number of studies (e.g., factor analyses by Egeland et al., 2015; Ord et al., 2021; van Dyke et al., 2013). However, it is well known that failure in one validity domain does not necessarily invalidate data in the other domain. For data from

another study, an experimental simulator study performed by Reece (2017) in Britain, the SRSI manual (Merten et al., 2019) reports correlations between SRSI pseudosymptom endorsement and Test of Memory Malingering scores (TOMM; Tombaugh, 1996) of −0.78 and −0.81 (second and third TOMM trials, respectively), and between SRSI pseudosymptoms and WMT of −0.79 and −0.83 (WMT Immediate and Delayed Recognition, respectively). However, these high correlations were largely due to significant SVT/PVT associations within the subgroup of experimental simulators ($N=30$) while they were missing in the control group ($N=30$). The combination of both groups increased the size of the correlation due to a considerably larger variance in the combined group ($N=60$). This demonstrates how much empirical results on the relationship between SVT and PVT measures appear to depend on the samples on which analyses are based.

Stevens (in Merten et al., 2019) investigated correlations with MMPI-2-RF (Ben-Porath & Tellegen, 2008) validity scales in a sample of 50 IME patients. He found a substantial correlation with the *F-r* score ($r=0.81$), followed by the Response Bias Scale ($r=0.73$), *Fs* ($r=0.68$), and the Fake Bad Scale ($r=0.55$).

Among the studies published after the completion of the test manual, two of them investigated the convergent validity of the SRSI with inpatients of a psychosomatic rehabilitation clinic, mostly patients diagnosed with mental disorders of the soft psychopathology range. This patient population fully meets the criteria of the instrument's primary target group. In the first study (Merten et al., 2020), complete protocols were available for 537 patients. A correlation of 0.73 was found between their total SRSI pseudosymptoms scores and SIMS total scores. Moreover, excessive reports of depressive symptoms, as reflected by Beck Depression Inventory-II (BDI-II; Beck et al., 1996) total scores over 40 were associated with a higher probability of being classified as overreporting on the SRSI and/or the SIMS. In a subsequent sample of 147 patients from the same clinic, Kaminski et al. (2020) found a correlation between SRSI pseudosymptoms and SIMS total scores of 0.72. Moreover, the number of endorsed SRSI pseudosymptoms correlated strongly ($r=0.82$) with the total scores of a newly developed German-language SVT (Beschwerdenvalidierungstest, BEVA; Walter et al., 2016).

However, 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). With regard to the risk of false-positive classifications, results obtained in memory clinic patients (Czornik et al., 2021; Lehrner, in Merten et al., 2019) suggest that the SRSI pseudosymptoms scale is more robust against the presence of genuine cognitive impairment than the SIMS.

Cut Scores and Hit Rates

As mentioned above, an analysis of the final 107-item SRSI was performed with a pooled sample of 520 participants from seven different studies, combining data from honestly responding controls, forensic inpatients, experimental malingerers, and IME patients (Merten et al., 2016, 2019). For 367 participants, SIMS results were available. A receiver operating curve (ROC) analysis was performed on the pseudosymptoms scores, using the SIMS as the gold standard and a SIMS cut score of 16 as recommended for most European versions of that instrument (van Impelen et al., 2014).

Seventy-six (21%) of the 367 SIMS protocols were positive indicating probable overreporting or invalid, noncredible symptom claims. The ROC analysis of the SRSI pseudosymptom scores yielded an area under the curve (AUC) of 0.931 (standard error of measurement: 0.015; 95% confidence interval: 0.901–0.961). According to commonly used standards, this reflected a highly accurate classification.

The manual and Merten et al. (2016) give more detailed information about possible cut scores and the corresponding sensitivity and specificity estimates, as well as about positive and negative predictive values for a variety of base rates of overreporting. For practical use, two different cut score were recommended, one for screening purposes, with a maximum of 10% false positives, and a standard cut score at which less than 5% false positives are to be expected. More information can be retrieved from Table 2. Two additional cut scores were discussed: (1) A liberal cut score where sensitivity is very high (0.90), but which should only be used for special research questions. In one study, Stevens et al. (2018) used a cut score of >5 pseudosymptoms. (2) A very rigorous cut score for which the specificity is set at 0.99 so the risk of false-positive results is expected to be as low as 1%.

These two special cut scores should not be used for routine clinical or forensic decision making because of a high probability of false-positive and false-negative decisions, respectively; these are risk potentials that can rarely be justified in routine assessment contexts. Table 3 presents a tentative interpretation guideline that integrates different cut scores and different degrees of diagnostic certainty.

For the additional index ratio (endorsed pseudosymptoms/endorsed genuine symptoms), a separate ROC analysis was performed (Merten et al., 2019). It yielded an AUC of 0.876 (standard error of measurement: 0.020; 95% confidence interval: 0.837–0.911). Following conventional

Table 2 Diagnostic statistics at two cut scores proposed for routine use (screening and standard) and additional cut scores

Cut score	Set at	Sensitivity	Specificity	Likelihood ratio
Pseudosymptom Endorsement				
Liberal (high sensitivity)	> 4 endorsed pseudosymptoms	0.90	0.83	5.20
Screening	> 6 endorsed pseudosymptoms	0.83	0.91	9.31
Standard	> 9 endorsed pseudosymptoms	0.62	0.96	13.73
Rigorous	> 15 endorsed pseudosymptoms	0.33	0.99	28.90
Ratio (number of endorsed pseudosymptoms/number of endorsed genuine symptoms)				
Screening	> 0.288	.59	.90	6.17

standards, this AUC value would also be considered high. However, this index is mathematically dependent on the primary variable, the number of endorsed pseudosymptoms. Therefore, it was conceived as an auxiliary variable only and a cut score was only proposed at screening level (Table 2).

Strengths and Weaknesses

The SRSI was developed primarily 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. Thus, it may add to the toolbox of forensic and clinical psychologists, particularly in German-speaking countries and some other non-English-speaking countries where there is a notable lack of freestanding SVTs with a well-established database. More recently, both the BEVA (Walter et al., 2016) and the IOP-29 (Viglione & Giromini, 2020; Viglione et al., 2017) also became available in Germany, to narrow the gap. Even in the case where there will be no incremental validity between pairs of these measures, it will be important to have a sufficiently high number of SVTs available, especially to counterbalance the effects of coaching and retesting. Results from a coaching study indicate that the SRSI appears to be as little immune against the effects of more subtle coaching procedures as is the case with the SIMS (Merten et al., 2010, 2019), whereas the MMPI-2 Fake Bad Scale showed no effects of coaching.

Current trends in SVT development seem to indicate that there is a need for freestanding validity scales outside the more time-consuming inventories with embedded validity scales. The latter encompass the Minnesota Multiphasic Personality Inventory (MMPI) family (e.g., Butcher et al., 1989), the Personality Assessment Inventory (Morey, 2007), and the Millon Clinical Multitaxial Inventory (e.g., Millon, 1987). These inventories and their embedded validity scales are in common use in the USA and a number of English-speaking countries, but this research has had little detectable impact on symptom validity research in other languages. For many practitioners and in many referral contexts, a well-developed validity test that usually takes no more than ten minutes to apply appears to be an attractive option.

Conceptually, the SRSI can best be seen as a psychometric relative of the SIMS. Still, the SRSI pseudosymptoms

Table 3 Tentative interpretation guideline for different scores of pseudosymptom endorsement

Number of endorsed pseudosymptoms	Interpretation
4 or less	From the questionnaire results, no evidence can be obtained to assume symptom overreporting
5 or 6 (failing the liberal cut score)	Area of uncertainty. Symptom overreporting is a distinct possibility, but cannot be established with sufficient confidence. The endorsement pattern is compatible with mild forms of overreporting. Consult ratio in such cases: a low ratio score (up to 0.288) would rather speak against overreporting (possibly indicating a false positive), a high ratio (> 0.288) would rather support the hypothesis of symptom overreporting
7 to 9 (failing the screening cut score)	Elevated probability of significant symptom overreporting. Further assessment is required (note that false-positive rate is up to 10% with this cut score). In cases of convergent lines of evidence, a positive SRSI score at screening level (but not at the standard cut point) can be strongly supportive of symptom overreporting
10 to 15 (failing the standard cut score)	Substantially elevated probability of significant symptom overreporting. False-positive results can be expected to occur in less than 5% of the cases
> 15 (failing the rigorous cut score)	Very strong evidence of symptom overreporting. The probability of false-negative classifications is very low (< 1%)

appear to be more robust against the effects of genuine cognitive impairment in respondents (Czornik et al., 2021; see also van Helvoort et al., 2019) than the SIMS (van Impelen et al., 2014). We assume that this robustness has to do with the systematic way in which the SRSI items were phrased and selected (i.e., the empirically based item analysis and attempts to resort to simple item structure while avoiding any ambiguities, double negation, etc.). Nevertheless, it remains to be seen to what extent the presence of true mental disorders facilitates false positive results on the SRSI. Methodologically, addressing this research question is a great challenge if only because of the difficulty to recruit unequivocal bona fide patient samples in typical mental-health settings, that is, groups of patients who refrain from any form of symptom exaggeration or symptom distortion (e.g., Dandachi-FitzGerald et al., 2011; Merten et al., 2020). What can be said with some confidence is that relatively well-functioning rehabilitation patients with soft psychopathology often fail on instruments like the SRSI, the SIMS, or the BEVA (e.g., Göbber et al., 2012; Kaminski et al., 2020; Merten et al., 2020). However, it is mostly difficult (if not impossible) to tell apart what is the percentage of false and of true positives among them. In this respect, a study by van Helvoort et al. (2019) seems to be worth a note. The authors investigated a highly selective sample of 40 forensic mental inpatients, taking special care not to include patients who might present a motive for symptom overreporting. Of the 40 patients, only two scored above the screening cutoff on the SRSI pseudosymptoms scale, and none scored positive when the more rigorous standard cut score was applied. This indicated that it was not mental health problems that produced high scores on SRSI pseudosymptoms report, but other aspects of response behavior (motivation, cooperation, honesty, possible hidden agendas, primary or secondary gain expectations, etc.).

The limitations of the research base on which the SRSI currently rests warrant some comments. One is that the SRSI has been largely validated against the SIMS as a "gold standard." Meanwhile, the primary reason for developing the SRSI was the imperfection and suboptimal performance of the SIMS, particularly in certain settings. Consequently, what is needed are more studies that validate the SRSI using different standards than the SIMS.

The instrument in its current form, and with the current decision rules, will regularly fail to detect noncredible complaints that are limited to one symptom domain, even more when complaints are limited to only one symptom or a narrow range of related symptoms. It will also fail to detect feigned symptoms outside the domains covered by the pseudosymptoms subscales (e.g., feigned flu symptoms; *cf*. Cheng, 2013).

Another obvious limitation of the SRSI is that the genuine symptom scales have limited diagnostic value so far. Research activities have mostly focused on the central pseudosymptoms scales; however, with a sufficient base of reference data, the instrument could also be used for evaluating the extent of genuine symptoms. Also, the additional index ratio warrants a more detailed analysis.

Still another weakness of the SRSI is that it remains to be coachable. With sufficient sophistication on the side of the claimant and/or a high-level, expertise-like coaching, a pseudopatient fabricating symptomatology may, of course, remain undetected by the instrument. On a positive note, encouraging results can be retrieved from a recent study by Boskovic et al. (2021) who asked psychology students (i.e., future experts) to rate the items of the English language SRSI on prevalence and plausibility. Many of the future experts did not detect the bogus character of the bizarre and implausible pseudosymptoms. (Although this very same inability of many professionals to accurately distinguish between genuine and fabricated or grossly exaggerated symptom claims is one of the core issues in validity assessment research and practice.)

Future Perspectives

The original German-language SRSI was not made available to professionals until 2019; however, ten different language versions were developed in the last few years. Also, the instrument was included in a host of studied with various samples (forensic patients, inmates of a youth prison, instructed malingerers, population-based samples) and with different methodologies. The data available so far show that the number of endorsed pseudosymptoms correlates highly with symptom overreporting on other instruments, notably the SIMS and the BEVA. In IME samples, SRSI scores tend to correlate moderately with underperformance. Reliability estimates were found to be satisfactory (i.e., internal consistency scores > 0.90; test–retest correlations > 0.85).

As this is a recently developed instrument, it is natural that many aspects and potential target populations have not been studied or have been studied insufficiently yet. One important aspect that needs further clarification is to what extent the symptomatology endorsed by overreporters takes on an overgeneralized form or may, alternatively, be restricted to a specific symptom domain. First empirical results (Merckelbach et al., 2018) indicated that SRSI pseudosymptoms subscales respond differentially to target symptomatology in experimental analogue malingerers. In this context, the question of separate cut scores for pseudosymptoms subscales will be of future interest.

As is true for all SVTs, the SRSI is *not* a malingering scale. It was developed to measure symptom overreporting. What factors underlie such overreporting is an important diagnostic question which cannot be answered from individual test scores. Malingering is just one context in which invalid data are

produced by claimants or patients who follow a secondary gain agenda (*cf.*, in particular, Sherman et al., 2020). The conceptual framework of validity assessment and the consequences of positive validity test results is certainly a topic that will continue to be of interest in the coming years. There are, however, a number of safe and well-established guidelines when using SVTs such as the SRSI. These include that (1) true positive results on SVTs invalidate the patients' symptom report regardless of the underlying factors or causes (whether malingering, factitious disorder, disengagement, frustration, boredom, attention seeking or whatever may have determined significant response distortions); (2) positive results on an SVT (fails) cannot be treated as *false*-positive results simply because the patient in question has serious complaints or was tested in a clinical setting with no apparent secondary gain motive; and (3) conversely, negative results on an SVT (passes) are no guarantee of a valid symptom report and should not be interpreted as such.

Future research by independent authors will show whether the SRSI can stand up to expectations and be used as a powerful and highly informative SVT. What is important is that both researchers and practitioners stick to the conditions of use for the instrument; violations may compromise the outcome both at the level of individual decision-making and with respect to the database of the instrument.

Of course, this does not mean that modifications of item presentation will automatically lead to invalid data. In the 1980s and 1990s, a wealth of research into computerized questionnaire presentation yielded mixed results (e.g., Merten & Ruch, 1996; Webser & Compeau, 1996). This was also the case for measures of positive response bias (e.g., Lautenschlager & Flaherty, 1990). However, to the authors' knowledge, no studies about paper–pencil vs. computerized equivalence have been performed so far for SVTs. Future research may also envisage the question of developing a short version of the SRSI and the possibility of embedding a defined set of items into other instruments. Furthermore, first steps have been envisaged to extend the SRSI by adding a genuine and a noncredible attention deficit/hyperactivity scale. Future studies should also investigate the accuracy of SRSI classifications when well-defined sets of criteria for malingered symptom presentations are applied (Bianchini et al., 2005; Sherman et al., 2020).

Declarations

Conflict of Interest Harald Merckelbach and Thomas Merten are two of the authors of the Self-Report Symptom Inventory, which is commercially distributed by Hogrefe Publishers, Göttingen, Germany.

References

Anstey, E. (1966). *Psychological tests.* Nelson.

Beaber, R. J., Marston, A. M., Michelli, J., & Mills, M. J. (1985). A brief test for measuring malingering in schizophrenic individuals. *American Journal of Psychiatry, 142*(2), 1478–1481. https://doi.org/10.1176/ajp.142.12.1478

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck Depression Inventory-Second Edition*. The Psychological Corporation.

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

Bianchini, K. J., Greve, K. W., & Glynn, G. (2005). On the diagnosis of malingered pain-related disability: Lessons from cognitive malingering research. *Spine Journal, 5*(4), 404–417. https://doi.org/10.1016/J.SPINEE.2004.11.016

Boskovic, I., Dibbets, P., Bogaard, G., Hope, L., Jelicic, M., & Orthey, R. (2019). Verify the scene, report the symptoms: Testing the verifiability approach and SRSI in the detection of fabricated PTSD claims. *Legal and Criminological Psychology, 24*, 241–257. https://doi.org/10.1111/lcrp.12149

Boskovic, I., Merten, T., & Merckelbach, H. (2021). How plausible is the implausible? Students' plausibility and prevalence ratings of the Self-Report Symptom Inventory. *Psychological Injury and Law, 14*(2), 127–133. https://doi.org/10.1007/s12207-021-09409-x

Boskovic, I., Merckelbach, H., Merten, T., Hope, L., & Jelicic, M. (2020). The Self-Report Symptom Inventory as an instrument for detecting over-reporting: An explorative study with instructed simulators. *European Journal of Psychological Assessment, 36*(5), 730–739. https://doi.org/10.1027/1015-5759/a000547

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

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., & Dahlstrom, W. G. (1989). *MMPI-2: Minnesota Multiphasic Personality Inventory-2*. University of Minnesota Press.

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

Cheng, M. (2013). *To tell the truth and nothing but the truth: The role of high and low stakes in the decision to malinger*. Master's thesis. Maastricht University, Faculty of Psychology and Neuroscience.

Cima, M., Hollnack, S., Kremer, K., Knauer, E., Schellbach-Matties, R., Klein, B., & Merckelbach, H. (2003). "Strukturierter Fragebogen Simulierter Symptome". Die deutsche Version des "Structured Inventory of Malingered Symptomatology: SIMS" [The German version of the Structured Inventory of Malingered Symptomatology]. *Nervenarzt, 74*(11), 977–986. https://doi.org/10.1007/s00115-002-1438-5

Czornik, M., Merten, T., & Lehrner, J. (2021). Symptom and performance validation in patients with subjective cognitive decline and mild cognitive impairment. *Applied Neuropsychology: Adult, 28*(3), 269–281. https://doi.org/10.1080/23279095.2019.1628761

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

Egeland, J., Anderson, S., Sundseth, O. O., & Schanke, A. K. (2015). Two types of malingering? A confirmatory factor analysis of performance and symptom validity tests. *Applied Neuropsychology: Adult, 22*(3), 215–226. https://doi.org/10.1080/23279095.2014.910212

Geurten, M., Meulemans, T., & Seron, X. (2018). Detecting over-reporting of symptoms: The French version of the Self-Report Symptom Inventory. *The Clinical Neuropsychologist, 32*(Suppl. 1), 164–181. https://doi.org/10.1080/13854046.2018.1524027

Giger, P., & Merten, T. (2013). Swiss population-based reference data for six symptom validity tests. *Clínica y Salud, 24*(3), 153–159. https://doi.org/10.1016/S1130-5274(13)70016-1

Giger, P., & Merten, T. (2019). Equivalence of the German and the French versions of the Self-Report Symptom Inventory. *Swiss Journal of Psychology, 78*(1), 5–13. https://doi.org/10.1024/1421-0185/a000218

Göbber, J., Petermann, F., Piegza, M., & Kobelt, A. (2012). Beschwerdevalidierung bei Rehabilitanden mit Migrationshintergrund in der Psychosomatik [Symptom validation in patients with migration background in psychosomatic medicine]. *Rehabilitation, 51*(5), 356–364. https://doi.org/10.1055/s-0032-1323669

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

Helmstadter, G. C. (1966). *Principles of psychological measurement.* Methuen & Co.

Kaminski, A., Merten, T., & Kobelt-Pönicke, A. (2020). Der Vergleich von drei Beschwerdenvalidierungstests in der stationären psychosomatischen Rehabilitation [Comparison of three symptom validity tests in a sample of psychosomatic inpatients]. *Zeitschrift für Psychiatrie, Psychologie und Psychotherapie, 68*(2), 96–105. https://doi.org/10.1024/1661-4747/a000408

Lautenschlager, G. J., & Flaherty, V. L. (1990). Computer administration of questions: More desirable or more social desirability? *Journal of Applied Psychology, 75*(3), 310–314. https://doi.org/10.1037/0021-9010.75.3.310

Lees-Haley, P. R., English, L. T., & Glenn, W. J. (1991). A Fake Bad Scale on the MMPI-2 for personal injury claimants. *Psychological Reports, 68*(1), 208–210. https://doi.org/10.2466/pr0.1991.68.1.203

Lilienfeld, S. O., Thames, A. D., & Watts, A. L. (2013). Symptom validity testing: Unresolved questions, future directions. *Journal of Experimental Psychopathology*, *4*(1), 78–87. https://doi.org/10.5127/jep.028312

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

Merckelbach, H., Merten, T., Dandachi-FitzGerald, B., & Boskovic, I. (2018). De Self-Report Symptom Inventory (SRSI): En instrument voor klachtenoverdrijving [The Self-Report Symptom Inventory (SRSI): An instrument to measure symptom overreporting]. *De Psycholoog, 53*(3), 32–40.

Merten, T. (2006). An analysis of the VOSP Silhouettes test with neurological patients. *Psychology Science, 48*(4), 451–462.

Merten, T., Friedel, E., & Stevens, A. (2007). Die Authentizität der Beschwerdenschilderung in der neurologisch-psychiatrischen Begutachtung: Eine Untersuchung mit dem Strukturierten Fragebogen Simulierter Symptome [Authenticity of symptom report in independent neurological and psychiatric examinations: A study with the Structured Inventory of Malingered Symptomatology]. *Praxis der Rechtspsychologie, 17*(1), 140–154.

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

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

Merten, T., Lorenz, R., & Schlatow, S. (2010). Posttraumatic stress disorder can easily be faked, but faking can be detected in most cases. *German Journal of Psychiatry, 13*(3), 140–149.

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., & Ruch, W. (1996). A comparison of computerized and conventional administration of the German versions of the Eysenck Personality Questionnaire and the Carroll Rating Scale for Depression. *Personality and Individual Differences, 20*(3), 281–291. https://doi.org/10.1016/0191-8869(95)00185-9

Millon, T. (1987). *Manual for the Millon Clinical Multitaxial Inventory–II (MCMI-II)* (2nd ed.). National Computer Systems.

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

Nijdam-Jones, A., & Rosenfeld, B. (2017). Cross-cultural feigning assessment: A systematic review of feigning instruments used with linguistically, ethnically, and culturally diverse samples. *Psychological Assessment, 29*(11), 131–1336. https://doi.org/10.1037/pas0000438

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

Pienkohs, S. (2021). Equivalence of the Dutch and German Self-Report Symptom Inventory. Unpublished master thesis. Maastricht University.

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

Reece, V. J. (2017). Validation of the symptoms of Post-concussion Syndrome Questionnaire as a self-report symptom validity test: A simulation study. Doctoral dissertation, Staffordshire and Keele Universities, UK. https://eprints.staffs.ac.uk/4469/

Schmand, B., & Lindeboom, J. (2005). *Amsterdam Short-Term Memory Test. Amsterdamer Kurzzeitgedächtnistest.* Manual, Handanweisung. PITS.

Schretlen, D., Wilkins, S., van Gorp, W., & Bobholz, J. (1992). Cross-validation of a psychological test battery to detect faked insanity. *Psychological Assessment, 4*(1), 77–83. https://doi.org/10.1037/1040-3590.4.1.77

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

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*(2), 180–183.

Stevens, A., Schmidt, D., & Hautzinger, M. (2018). Major depression – A study on the validity of clinicians‘ diagnoses in medicolegal assessment. *The Journal of Forensic Psychiatry & Psychology, 29*(5), 794–809. https://doi.org/10.1080/14789949.2018.1477974

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

Tombaugh, T. N. (1996). *Test of Memory Malingering (TOMM).* Multi-Health Systems.

van der Heide, D., Boskovic, I., & Merckelbach, H. (2017). Standard symptom inventories for asylum seekers in a psychiatric hospital: Limited utility due to poor symptom validity. *Psychological Injury and Law, 10*(4), 358–367. https://doi.org/10.1007/s12207-017-9302-x

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 Helvoort, D., Merckelbach, H., & Merten, T. (2019). The Self-Report Symptom Inventory (SRSI) is sensitive to instructed feigning, but not to genuine psychopathology in male forensic inpatients: An initial study. *The Clinical Neuropsychologist, 33*(6), 1069–1082. https://doi.org/10.1080/13854046.2018.1559359

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

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

Walter, F., Petermann, F., & Kobelt, A. (2016). Erfassung von negativen Antwortverzerrungen – Entwicklung und Validierung des Beschwerdenvalidierungstests BEVA [Assessment of negative response bias: Development and validation of the BEVA]. *Rehabilitation, 55*(3), 182–190. https://doi.org/10.1055/s-0042-105939

Webster, J., & Compeau, D. (1996). Computer-assisted versus paper-and-pencil administration of questionnaires. *Behavior Research Methods, Instruments, and Computers, 28*(4), 567–576. https://doi.org/10.3758/BF03200544

Widows, M. R., & Smith, G. P. (2005). *SIMS – Structured Inventory of Malingered Symptomatology. Professional Manual.* Psychological Assessment Resources.